Creating Useful Integrated Data Sets to Inform Public Policy

By Nancy Fagenson Potok

Bachelor of Arts, 1978, Sonoma State University

Master of Administrative Science, 1980, University of Alabama

A Dissertation submitted to

The Faculty of
the Columbian College of Arts and Sciences
of the George Washington University
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy

August 31, 2009

Dissertation directed by

Kathryn Newcomer

Professor of Public Policy and Public Administration

www.manaraa.com

UMI Number: 3368742

INFORMATION TO USERS

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleed-through, substandard margins, and improper alignment can adversely affect reproduction.
In the unlikely event that the author did not send a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

UMI®

UMI Microform 3368742
Copyright 2009 by ProQuest LLC
All rights reserved. This microform edition is protected against unauthorized copying under Title 17, United States Code.

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

المنارة للاستشارات

www.manaraa.com

The Columbian College of Arts and Sciences of the George Washington University certifies that Nancy Fagenson Potok has passed the Final Examination for the degree of Doctor of Philosophy as of August 5, 2009. This is the final and approved form of the dissertation.

Creating Useful Integrated Data Sets to Inform Public Policy

By Nancy Fagenson Potok

Dissertation Research Committee:

Kathryn Newcomer, Professor of Public Policy and Public Administration, Dissertation Director

Joseph Cordes, Professor of Economics, Public Policy and Public Admin, and of International Affairs, Committee Member

Julia Lane, Program Director, National Science Foundation, Committee Member

www.manaraa.com

Acknowledgements

First, my gratitude goes to my Dissertation Research Committee, including Professor Kathryn Newcomer, Professor Joseph Cordes, Professor Donna Infeld, and Professor Marsha Regenstein of The George Washington University and Julia Lane of the National Science Foundation. I appreciate and greatly benefited from your support, direction and insightful suggestions for improving the research. I especially want to thank Kathy Newcomer, my research director, for her unwavering, unrelenting and cheerful encouragement and willingness to help me keep moving this research forward.

Additionally my thanks are extended to the study participants who work with administrative records (currently or in the past) for making yourselves available, providing me with background papers, and being supportive of this work during the long research and writing period. I was touched by your willingness to spend so much time with me and taking time out of your workday on behalf of this research, and I deeply appreciate it.

To my fellow students in the Applied Probability Group, thank you for helping me feel connected, keeping me focused and aware of the passage of time, as well as offering much needed advice on the process of moving forward.

Finally, a gigantic thank you goes to my family for being supportive in all ways and for such a long time. I dedicate this research to my late mother, Harriet Fagenson. Although she did not quite live long enough to see me complete all the requirements for the Ph.D., she was supremely confident that I would do it.

www.manaraa.com

Abstract

Creating Useful Integrated Data Sets to Inform Public Policy

The costs of traditional primary data collection have risen dramatically over the past decade. For example, the cost of the decennial census of population and housing, conducted by the U.S. Census Bureau, has risen from $6 billion in 2000 to an estimated $14.5 billion in 2010. Other surveys and censuses conducted by the government have also risen in costs. Yet some of the same data are collected by other federal agencies and contained in administrative records such as Medicare and tax records. Sharing of administrative record data between federal agencies has the potential to increase the information that is available for policy makers while saving money. Significant policy issues related to safeguarding privacy and confidentiality, as well as questions about data quality have resulted in barriers that slow down or stop record sharing. But do the barriers address real or perceived problems?

This research used two exploratory case studies to examine the creation of integrated data sets among three government agencies, the Internal Revenue Service (IRS), the U.S. Census Bureau (Census), and the Centers for Medicare and Medicaid Services (CMS). It identified the policy issues raised by the creation of such data pools and examined how these issues are approached in a decentralized governmental statistical system, such as that found in the United States. The creation of new, combined data sets and the related policy issues were examined through five dimensions, legal, technical, organizational, perceptual, and human.

The case studies addressed the following research questions related to the sharing of administrative records between U.S. Federal agencies:

www.manaraa.com

1) What is the life cycle flow of administrative records data on individuals and businesses between IRS, CMS, and the Census Bureau?
2) What are the significant issues that have arisen as a result of sharing administrative records related to the need to protect privacy and confidentiality?
3) What insights and potential solutions can be learned from the experience of those who have worked within the federal statistical system that would help address the significant data-sharing issues that have been identified?

The study found that each agency involved in sharing administrative records is governed by a different set of statutes and regulations that only partially overlap. This patchwork of laws and regulations greatly slows down the initiation of record sharing projects. Participants at the agencies believe that privacy safeguards are adequate and effective. Participants at the agencies expend significant effort to assure that data are protected as required by law and by interagency agreements. Each agency has its own distinct internal processes for approving and tracking record sharing projects. There are no mature government-wide shared processes or criteria for reviewing or approving projects involving multiple agencies. The current processes are slow and burdensome and discourage initiation of new projects.

www.manaraa.com

Table of Contents

www.manaraa.com

www.manaraa.com

List of Figures

www.manaraa.com

List of Tables

www.manaraa.com

Chapter 1: Problem Statement

Introduction

The costs of traditional primary data collection have risen dramatically over the past decade. For example, the cost of the decennial census of population and housing, conducted by the U.S. Census Bureau, has risen from $6 billion in 2000 to an estimated $14.5 billion in 2010. Other surveys and censuses conducted by the government have also risen in costs. Yet some of the same data are collected by other federal agencies and contained in administrative records such as Medicare and tax records. Sharing of administrative record data between federal agencies has the potential to increase the information that is available for policy makers while saving money. Significant policy issues related to safeguarding privacy and confidentiality, as well as questions about the quality of the data have resulted in barriers that slow down or stop record sharing. But do the barriers address real or perceived problems?

This research employs an exploratory case study approach to examine the creation of integrated data sets among three government agencies, the Internal Revenue Service (IRS), the U.S. Census Bureau (Census), and the Centers for Medicare and Medicaid Services (CMS). It then identifies the policy issues raised by the creation of such data sets. Two case studies are used to highlight the issues, identify solutions that may have been attempted, recommend possible improved approaches and solutions, and identify additional research that may be needed. The creation of the integrated data sets and the related policy issues are examined through five dimensions, legal, technical,

www.manaraa.com

organizational, perceptual, and human. These dimensions will be discussed in more depth in chapter three.

This chapter has five parts. First it provides background on the challenges facing the U.S. federal statistical system that create pressures to increase the use of administrative records rather relying on primary data collection methods such as surveys. Second, the chapter covers the research questions and discusses the need for this study. Third, the chapter provides an overview of the history of U.S. government confidentiality protection. Fourth, the conceptual framework of the paper is discussed. The framework for the study is conceptual rather than theoretical, because the purpose is to describe the data sets created by sharing records between federal agencies; illuminate the policy issues surrounding these combined data sets; and look for successful practices. The research itself is not intended to develop new theory, although it may contribute to the development of theory by other researchers. Finally, the chapter ends with a summary of the research approach.

Background

Within the federal statistical system, most agencies are currently experiencing flat or decreasing budget appropriations, with the exception of the cyclical upswings for the 2010 decennial census of population and housing. At the same time, the costs of collecting information directly from individual households and businesses are increasing. A number of factors have emerged in recent years that have collectively made it more difficult and expensive for Federal agencies to use primary modes of data collection such as door-to-door surveys, telephone interviews, web based surveys, and electronic-based

www.manaraa.com

questionnaires (2002; Groves & Couper, 1998) . Household response rates are dropping, as evidenced by the trend in the decennial census, where mail response rates (the rate of households that mail back a filled out census questionnaire) dropped from 78% to 67% between 1970 and 2000 (NRC, 2004). Factors contributing to dropping household response rates include busier two-career households whose members are unwilling to take the time to respond; greater public hostility toward telemarketers that spills over to government telephone surveys; an abundance of junk mail that increases a household's likelihood of throwing away a mailed questionnaire; increased desire on the part of respondents to be compensated for their time, which the government does sparingly; and distrust of the government (Singer, 2002). Further, telephone interviewing is being affected by an increasing number of households that are not using land line-based telephone service and have instead substituted exclusive use of cell phones. This creates problems of bias in telephone surveys, because cell phone numbers generally are not included in survey samples. According to a 2004 supplement to the Current Population Survey (CPS), about 6 percent of households have only cell phones, ranging from a high of 12.8 percent of renters to 3.1 percent of home owners. (Tucker, Brick, & Meekins, 2007). More recently, early release estimates from the National Health Interview Survey indicated that nearly 18% of households are wireless only, and that the percentage of adults that live in households with only wireless phones has increased from 6.7% to 16.1% (more than 36 million adults) between 2005 and 2008 (CDC, 2008).

Although it is hard to measure the precise effect that the public's distrust of government has on its willingness to respond to surveys, the level of distrust as evidenced by public opinion polls fluctuates, and is often influenced by external events that do not

www.manaraa.com

relate directly to the agency trying to collect data or the purposes for which they are being collected. Distrust of government often overlaps with concerns about privacy and how information collected by the government is being used.

Many surveys and public opinion polls have been conducted to track attitudes towards privacy. These attitudes are complex and vary by topical area such as medical records, internet use, homeland security, law enforcement, telemarketing, or other activities. However, one 2002 survey conducted by the University of Connecticut found that 60% of the respondents thought the government possessed too much personal information about individuals (Paulson, 2002). While this does not translate directly into a lack of cooperation in responding to government sponsored surveys, it does indicate that agencies are operating in a challenging environment.

In addition to privacy concerns, respondents such as businesses may find responding to government surveys time consuming, due to both the frequency of data collection for certain economic indicators such as retail sales and the complexity of some survey instruments, especially for large corporations. Businesses are asked by the government to respond to many monthly, quarterly, and annual surveys, as well as an economic census conducted every five years by the Census Bureau. The data from on-going surveys and periodic censuses of businesses play a critical role in calculating the key national economic indicators such as gross domestic product, and include data on retail sales, housing starts, manufacturing, services, and other sectors of the economy, as well as employment information. However, businesses find responding to these many, often detailed, inquiries burdensome. While many surveys conducted by the Census Bureau are mandatory, surveys conducted by other agencies often are voluntary and

www.manaraa.com

struggle to achieve high response rates. For example, the 2003 Survey of Small Business Finances, a voluntary telephone survey of about 14,000 small businesses conducted by the Federal Reserve Board, achieved a weighted response rate of 32.4% (Potok et al., 2005).

As a result of the increasing difficulties of primary data collection, costs of collection have risen significantly. This has resulted in agencies being forced to cut or redesign some popular surveys, many of which have large constituencies of researchers and other data users. For example, the Census Bureau's Survey of Income and Program Participation (SIPP) sparked this response from one of its stakeholders, the National Low Income Housing Coalition (NLIHC) when the President's Fiscal Year 2007 Budget contained a proposed funding reduction:

> "The decision of the Census Bureau to discontinue the SIPP this year has received considerable attention, with a letter opposing the move signed by 426 academics from across the social sciences, including two Nobel Laureates, going to the Hill. After considerable planning in the 1970s, the SIPP was implemented in the 1980s to track Americans' use of federal programs over time and the implications for income and wealth. Now, the Census Bureau, claiming growing difficulties with the response rate and refusals to participate, is proposing cutting funding from $40 million to less than $10 million in the FY07 budget. If the reduction survives the budget process, most of the remaining funding would go toward what one official recently described as the "planning and development for a new approach to providing wealth, income, health insurance and program participation data for the United States. Critics of this decision, including NLIHC, say that the SIPP does not stand out among Census surveys as having particular problems with its response rate. In the short term, discontinuing the survey without a

www.manaraa.com

clear plan going forward will disrupt ongoing research and the ability to track program participation. Over the long term there is a significant risk, particularly in the current budget environment, that a replacement survey will not be forthcoming. ***Also, the Census has stated it plans to use more administrative records such as HUD's data on tenants in its new approach, which raises a number of methodological concerns***. (emphasis added)" (NLIHC, 2006)**.**

Similar concerns were expressed during the FY 2008 budget cycle, in which the President's Budget proposed replacing the SIPP with a less expensive survey. Hundreds of academics, social service providers and others sent a letter supporting continuation of SIPP until a reliable replacement is in place. Several members of Congress also expressed concern, including Representative Carolyn Maloney (D-NY) who issued a statement that said, among other things, "We'll have the statistical equivalent of a Katrina on our hands if the OMB refuses to request funding for the SIPP,"(T. Washington Post, 2007). In response, the Census Bureau shelved its plans to develop and test a cheaper replacement using administrative records. While the survey was downsized to $24 million in FY 2008, the budget request for FY 2009 increased funding by $21 million to almost $46 million, bringing the survey back to its previous size.

In spite of potential objections from stakeholders, rising costs have prompted Federal agencies to look for even more ways in which secondary sources of data can be used to substitute for primary data collection. Because Congress is increasingly unwilling to divert funding from existing programs to start new data collection efforts, the need is particularly acute when government agencies and constituent researchers are looking to explore new areas not addressed by current on-going surveys. One example of

www.manaraa.com

this is the Longitudinal Employer-Household Dynamics Program (LEHD), now a part of the LED or Local Employment Dynamics program. LEHD was started by the Census Bureau in 1998 by combining existing data from censuses, surveys, and administrative records to create new data and products (http://lehd.dsd.census.gov). Under this program, unemployment insurance wage records, as well as establishment records are supplied to the Census Bureau by states each quarter. The Census Bureau merges the state records with other data from sources such as Census 2000, the American Community Survey, IRS summary and detailed earnings records, Social Security Administration numident and benefit data, and the Census Bureau Business Register, and economic censuses and surveys to produce new data and products. These new products include a longitudinal national frame of jobs and an associated data infrastructure that provides information on where workers live, where people work, and companion reports on age, earnings, and industries by geographic block.

The goal for the program is to create a new data infrastructure that captures the complex interactions among households and businesses at the microeconomic level and characterizes the dynamics of the modern economy while overcoming problems of high cost and lower response rates associated with primary data collection (Abowd, Lane, & Haltiwanger, 2004). LEHD is one of many efforts underway to make use of combined administrative records from multiple sources to make use of combined administrative records from multiple sources to create new integrated data sets.

Although secondary data collection is likely to play an important role in the future of federal statistics, it also raises policy issues on how the data are protected and confidentiality assured. Secondary data collection activities result in the creation of data

www.manaraa.com

pools that combine data from multiple agencies. A complex system of overlapping laws, regulations, and policies govern how these data are collected, merged, handled, analyzed, and shared among the more than 70 federal agencies or organizational units that carry out statistical activities. OMB has attempted to establish a uniform policy for all federal statistical collections by issuing policy guidance on the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) (OMB, 2006). However, OMB currently recognizes only 12 agencies or units as being "statistical" under the law. Thus, there remains a patchwork approach to administrative records sharing among all federal agencies.

This dissertation addressed the following research questions related to the sharing of administrative records between federal agencies, specifically IRS, Census, and CMS.

1) What is the life cycle flow of administrative records data on individuals and businesses between IRS, CMS, and the Census Bureau?
    a. What are the laws, rules and regulations guiding the sharing of these records?
    b. To what uses are the data put, and how does that affect the handling of the records?
    c. What are the business processes that guide the sharing and use of combined data including: 1) agency policies for internal handling; 2) training received by the people who handle the records; 3) compliance measurement; and 4) granting external access to the combined data.
2) What are the significant issues that have arisen as a result of sharing administrative records related to the need to protect privacy and confidentiality?

www.manaraa.com

a. Where do the laws, rules, and regulations overlap or conflict?

b. Who "owns" the combined data?

c. What are the barriers to achieving the intended benefits of data sharing among agencies?

3) What insights and potential solutions can be learned from the case studies that might be applied to help address the significant data-sharing issues that have been identified?

To answer these questions, a qualitative case study approach was used to examine the data through five dimensions: legal, perceptual, organizational, technical, and human. The study researched the legal authorities governing data sharing among these agencies, mapped the flow of data into and out of the shared data pools, and documented the business processes used to share and safeguard the data records. In addition, current and former employees of the federal agencies in the case studies were interviewed about various aspects of their jobs as they relate to sharing administrative record data. At the end of the research, a subset of the employees was re-interviewed to gain insight into whether the issues and potential solutions were feasible and appropriate.

Table 1 provides a list of key definitions of terms used in this study. To the extent possible, these definitions conform to definitions used by federal statistical agencies and OMB.

www.manaraa.com

Table 1 Definitions of key terms used in the study

Term	Definition
Federal statistical system	A decentralized system consisting of more than 11 separate agencies located in 9 different federal government departments; and some 70 other agencies of the government that produce statistical output as a part of their programmatic responsibilities.
Statistical agency	Agencies or units whose activities are predominantly the collection, compilation, processing, or analysis of information for statistical purposes
Administrative records and administrative records data	Administrative records and administrative records data refer to microdata records contained in files collected and maintained by administrative (i.e., program) agencies and commercial entities. Government and commercial entities maintain these files for the purpose of administering programs and providing services. Administrative records are distinct from systems of information collected exclusively for statistical purposes
Confidentiality	Pledges given by agencies that assure the public that information about or provided by individuals or organizations for exclusively statistical purposes will be held in confidence and will not be used against such individuals or organizations in any agency action
Privacy	How government agencies or other entities respect and minimize intrusion on the personal life or business operations of the respondent by the manner of collecting information and the nature of the information sought
Disclosure review	The procedures that statistical agencies apply to all data products that they publicly release in order to protect confidentiality
Respondent	A person (other than a Federal employee responding to inquiries within the scope of his employment, see CFR 1320.3(c)(4)) who is requested to provide information, or is the subject of that information, or who provides that information
Data stewardship	The process of meeting the public need for statistical information as well as the legal and ethical obligation to respect individual privacy and protect confidentiality
Informed consent	The agreement of the respondent to provide personal data for research and/or statistical purposes based on the full exposure to the facts, including any risks involved and available alternatives to providing the data, needed to make an intelligent decision to participate. It applies when respondents have a clear choice to

www.manaraa.com

	participate or not and are not subject to any penalties for failing to provide data
System of records	Under the Privacy Act, "a group of any records under the control of an agency from which information is retrieved by the name of the individual or by some identifying number, symbol, or other identifying particular assigned to the individual"
Business Register	A current and comprehensive database of U.S. business establishments and companies for statistical program use. *Business Register* information is establishment-based and includes business location, organization type (e.g., subsidiary or parent), industry classification, and operating data (e.g., receipts and employment).
Data pool	Collection of data that have been collected, combined, and stored in a retrievable manner from a variety of sources including administrative records and surveys and censuses, with or without the explicit knowledge of the original data provider.
Legal Dimension	Laws, regulations and policies governing the sharing of data between agencies.
Perceptual Dimension	The views and perceptions that shape the behavior of individuals regarding data sharing between agencies
Organizational Dimension	Processes, procedures and organizational structures that direct how agencies engage in sharing data with other agencies and handle the combined data sets
Technical Dimension	Technological advances, trends, practices, and IT security affecting the ability of agencies to share data and work with combined data sets while still protecting privacy and confidentiality
Human Dimension	The behaviors of individuals within their organizations that affect the development and implementation of administrative record sharing and creation of combined data sets for research purposes.

Need for the Study

The need for a study that creates a better understanding of the life cycle characteristics of data pools is driven by the need to continue to have high quality data supplied by the public in order to accurately develop and assess public policy. The federal government is required to protect the privacy of individuals and businesses that provide data to the government. Importantly from the perspective of the government's success in collecting data, the public's perception of whether the government is handling

www.manaraa.com

personal data in a secure manner will have an effect on the public's willingness to provide information in existing primary data collection vehicles. Of increasing interest to the federal government are the uses for the combined data, including better quality information for the distribution of billions of dollars in Federal funds distributed by formula grants, cost savings gained through more efficient data collection, and providing access to data to researchers studying federal programs and policies. However, many issues remain unaddressed regarding the safeguarding, ownership, quality and life cycle of data once they are shared and combined with other data. Many of the problems stem from the patchwork of individual agency rules and legislative authorities that have emerged over the last several decades. Because the rules and laws are not always harmonious, agencies often spend months or years negotiating individual arrangements for data sharing, causing lost productivity and creating opportunity costs of not taking advantage of greater efficiencies and potentially higher quality data.

In order to begin to address some of the issues, it is important to look at what happens to data once they are collected and combined into data pools. More information about data pools could inform how external agents may manipulate the pools, including controlling the quality of the data that gets into the pools. As a result, policy issues surrounding how data are shared and protected could begin to be systematically addressed.

One characteristic of data pools containing one's personal information is that they may be merged without that individual's knowledge. The merging may occur in the private as well as the public sector. For example, a random group of cash register receipts from Wal-Mart may be data mined to determine what groupings of products

www.manaraa.com

shoppers prefer. Safeway may conduct a similar analysis. Some of the shoppers may overlap between the two stores, causing an unidentified partial merging of the two pools.

A small number of companies are in the business of merging and reselling massive amounts of data on individuals that have been collected from a variety of sources to corporations and government agencies. More recently, acquisitions by large internet search engine companies such as Yahoo and Google of targeted on-line ad companies have raised questions among privacy advocates about uses of these newly created data pools (Washington Post 2007). This trend is continuing, and creating opportunities for more collection of data that is specific to individuals.

Sweeny (2001) notes that there are three behavioral trends that have arisen among entities that collect data on individuals: (1) collect more (expanding the number of fields being collected on an individual); (2) collect specifically (replacing an existing aggregate data set with one that identifies individual characteristics); and (3) collect if you can (starting a new data base with information specific to individuals to answer a new question or because its doable). Some data pools identified by Sweeny (2001) that have been recently created and continue to grow include Immunization Registries, the National Directory of New Hires, on-line birth certificates, supermarket and other customer loyalty cards, and health care cost data.

Events in recent years have raised questions about government data retention and the safeguarding of personal data include the loss of laptop computers containing personal information by several government agencies such as the Commerce Department, the Internal Revenue Service (Washington Post, 2007), the Veterans Administration, and the Transportation Security Administration (Hsu, 2007); as well as the display on the

www.manaraa.com

internet of farmers' social security numbers embedded in Department of Agriculture data, which was displayed on a Census Bureau operated web site (Nakashima, 2007)

According to Sweeney, other considerations arise when there are secondary uses of the data. Even when parties consent to give data, they don't always know to what use their data will be put after it is collected. Sometimes, the secondary uses of data are not even identified until after data are collected. This may create tensions, because there could be pressure at that point to use more specific data that identifies individuals, rather than more conglomerate data, or data that have been made anonymous. That is because the more identities are protected, the more data are changed, which affects data quality. The tensions result from the researchers trying to get as much individualized data as possible and the owners of the data pool trying to protect privacy as much as possible while recognizing that the research could provide great benefit to society. In some instances however, the secondary use of data is not research but marketing. Currently there are no general rules related to how much privacy should be protected for secondary uses. Because there are many data holders, the current decision- making processes on how to find the correct balance between protecting privacy and having high quality useful data are crude and often spontaneous.

This dissertation used a case study methodology to describe the lifecycle and characteristics of specific data pools created within the U.S. federal statistical system. The research examines the laws, rules, and policies that govern the handling and use of the data in the pools, and how those are implemented and followed in practice. The case studies focus on data that originate in three federal agencies: IRS, Census and CMS, because they are good illustrations of agencies with very different missions and methods

www.manaraa.com

of primary data collection, different governing authorities, and frequent collaboration in combining data into new integrated data pools.

Overview of U.S. Government Confidentiality Protection History

Agency Authorities

The three U.S. agencies examined in this research, IRS, CMS and the Census Bureau, each have their own statutes governing how they protect the privacy and confidentiality of data they are collecting, and under what conditions data can be shared with other agencies. These protections have evolved over time, reflecting changing public attitudes and the introduction of new technologies for collecting, handling and storing data. The following is a brief overview of the current specific authorities governing these three agencies.

The IRS is governed by Title 26, subtitle F of the U.S. Code. Chapter 61, subchapter B, sections 6103 and 6108 address confidentiality of tax returns and how information could be shared with other agencies. Sections 6103(j)(1)(A) and (B) specifically address statistical use of tax return information by the Census Bureau, stating that upon request in writing by the Secretary of Commerce, the Secretary of the Treasury shall furnish tax return information to the Bureau of the Census. Other agencies authorized to receive tax data include the Bureau of Economic Analysis, the Federal Trade Commission, the Department of Treasury, the Department of Agriculture (to conduct the Census of Agriculture), and the Congressional Budget Office (for long term models of the Social Security and Medicare programs). Section 6108 authorizes the IRS to publish annual statistics on income and conduct special statistical studies using tax data. Sections 6103 and 6108 both state that no publication or published information

www.manaraa.com

disclosure can be associated with or identify a particular taxpayer. Section 7213A of Chapter 75, Subchapter A, Part 1, prohibits any unauthorized person from inspecting tax return information, and section 7431 sets damages for disclosure. These sections were both enacted in 1997.

CMS is governed by the Privacy Act of 1974 (5 U.S.C. 552(a) and the Department of Health and Human Services regulations (45 U.S.C. 552a) that implement the act by setting policies and procedures for the maintenance and release of records. Section 5b(9)(b)(4) specifically allows release of individual records to the Bureau of the Census for purposes of planning or carrying out a census or survey or related activity pursuant to Title 13, without obtaining the consent of the individual whose record is being shared.. In addition, section 5b allows release of records to the National Archives, to another government agency for civil or criminal law enforcement activity, to either House of Congress, and to the Comptroller General and the Government Accountability Office (GAO) without obtaining additional consent. The Health Insurance Portability and Accountability Act (HIPAA) also affects the handling of records but primarily covers the health care industry. Covered entities under HIPAA are health care providers, health plans, health care clearinghouses, and Medicare drug plan providers.

The Census Bureau is governed by Title 13 of the U.S. Code. Chapter 1 Subchapter 1, Section 6 authorizes the Secretary of Commerce to call upon any other department, agency, or establishment of the federal government, or of the government of the District of Columbia, for statistical related information. The Secretary may acquire, by purchase or otherwise, from states, counties, cities, or other units of government, or from private persons and agencies, copies of records, reports, and other material required

www.manaraa.com

for the efficient and economical conduct of the censuses and surveys provided for in this title. The Secretary is also directed to acquire and use information available from any source referred to in subsection (a) or (b) of this section instead of conducting direct enquiries to the maximum extent possible and such that it doesn't compromise the quality and timeliness of the data.

Subchapter 1, Section 9 prohibits using information furnished under title 13 for any purpose other than the statistical purposes for which it is supplied; or to publish data in which an individual or establishment can be identified; or to permit anyone other than the sworn officers and employees of the Department of Commerce or Census to examine the individual reports.

Early History of Privacy Protection in the U.S.

Privacy statutes and the government's attitude towards privacy protection and sharing data among government agencies have evolved over time. Statistical confidentiality was a concept that developed alongside the development of U.S. official statistics during the 19th century, originally for the purpose of distinguishing statistical work from law enforcement and administrative record keeping (Bohme & Pemberton, 1991). Confidentiality was particularly important to businesses worried about release of proprietary information. However, the lack of technology at that time made it difficult for the government to amass large, centralized, retrievable databases, so the primary concerns were about individuals leaking sensitive information. By the time of the 1910 census, the Census Bureau, through a proclamation issued by President Taft, was assuring the public that census data would not be used for law enforcement or tax collection purposes (Barabba, 1975). However, during World War I, the U.S.

www.manaraa.com

government built up a greater centralized statistical capacity, including the creation of the Selective Service in 1917 to register young men for the draft, and the creation of the Central Bureau of Planning and Statistics in 1918 to gather economic statistics. The Census Bureau provided information from the 1910 census on the location and distribution of young men to aid in the planning of where to locate draft offices. Eventually, the Census Bureau provided local draft boards, the courts, and the Justice Department with the names and addresses of young men in order for those agencies to determine who had failed to register for the draft (Anderson & Seltzer, 2004). At that time, there were no laws prohibiting the disclosure of names and ages of individuals that had been collected in the census. For the 1920 census, Congress clarified that confidentiality applied to both businesses and individuals, but only specifically prohibited the Census Bureau from sharing business data. The Census Bureau shared lists of illiterates taken from the 1920 census with government and private organizations. The Census Bureau also furnished information to the Internal Revenue Bureau on the ages of children employed in business establishments for tax enforcement purposes (Bohme & Pemberton, 1991).

The authorizing legislation for the 1930 census reaffirmed the confidentiality principle, but also maintained exceptions at the discretion of the Director of the Bureau of the Census that allowed the bureau to share information on individuals as long as the information wasn't used to harm the person to whom the information related. However, information requests for shared census data abated, as other federal agencies were increasing their statistical capacity and the Roosevelt administration created the Central Statistical Board (Anderson, 1988). After the start of World War II, concerns about

www.manaraa.com

confidentiality gave way to national security and intelligence concerns. Section 1402 of the Second War Powers Act of 1942 specifically gave the Secretary of Commerce the ability to share census data, both business and individual, with other federal agencies upon request if the data were needed for use in connection with the conduct of war. As a result, the Census Bureau shared considerable data with other agencies. The shared data often related to businesses' wartime production, but also included the names and addresses of Japanese Americans identified in the 1940 census who might be considered security threats. The Act expired in 1947, as did the discretion to share identifiable records (Seltzer & Anderson, 2007).

Post-World War II to Present

After World War II ended and throughout the 1950s, policies surrounding the sharing of data between federal statistical agencies were unclear. For example, many agencies continued to request information from the Bureau of the Census, as they had under the Second War Powers Act of 1947. At the same time, statistical agencies were beginning to realize that it would be difficult to collect information from the public under a pledge of confidentiality if data were freely shared with law enforcement, regulatory and other agencies. The agencies began to turn down requests for data and develop an institutional culture that gave priority to protecting confidentiality (Anderson & Seltzer, 2004). A Supreme Court case in the early 1960s, *St. Regis Paper Company vs. the United States*, 368 US 208 (1961) brought high level government focus to the issue. The court ruled that a private company needed to turn over to the Federal Trade Commission (FTC) forms it had sent to the Census Bureau in response to the 1958 Census of Manufacturing. While the Census Bureau itself was not required to share the forms with

www.manaraa.com

the FTC, the court interpreted Title 13, the governing statute for Census Bureau data collection, as not applying to information held by private companies. One result was that in 1962, Congress amended Title 13 to provide a confidentiality guarantee for the file copies retained by companies filing census reports (Title 13, U.S.C, Section 9; Public Law 87-813). In addition, the government started to examine the concepts of privacy and confidentiality within the overall federal statistical system in greater depth.

During the 1970s, confidentiality of government records became an issue of concern to the American public when, during the Watergate period, it was revealed that the White House had received tax information on political opponents. A series of commissions were established to look at confidentiality of government records. A history of the evolution of current government views on protecting privacy and confidentiality is contained in the OMB guidance for implementation of Title V of the E-Government Act, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA) (U. S. OMB, 2006).

The OMB-produced historical notes describe the creation of three federal commissions that examined data confidentiality during the 1970s: the President's Commission on Federal Statistics in 1971, the Privacy Protection Study Commission in 1977, and the President's Commission on Federal Paperwork in 1977. These three commissions recommended several principles that ultimately became part of the Paperwork Reduction Act of 1980. The recommendations included the concepts that: (1) the term confidential should always mean prohibiting the disclosure of data in a manner that would allow public identification of the respondent or would in any way be harmful to him; (2) agencies should not promise to hold data in confidence unless

www.manaraa.com

the agency has legal authority to uphold such a promise; (3) no record or information collected or maintained for a research or statistical purpose under federal authority should be used in individually identifiable form to make any decision or take any action directly affecting the individual to whom the record pertains; (4) information collected or maintained for statistical purposes should only be shared with another statistical agency with assurances that it will be used solely for statistical purposes; and (5) information collected for administrative and regulatory purposes must be made available for statistical use, with appropriate confidentiality and security safeguards, when assurances are given that the information will be used solely for statistical purposes.

In addition, the Tax Reform Act of 1976 restricted presidential authority to access tax records, as well as putting strict limitations on how tax records could be shared with other agencies for nontax purposes. The restriction resulted in situations such as the Department of Agriculture (a non-statistical agency) being blocked from using tax return information to construct an address list to survey farmers, while the Census Bureau was allowed to use tax data to construct an address registry of all U.S. businesses.

In 1991, the Economic Policy Council Working Group of the Council of Economic Advisers, known as the Boskin Commission after its chairman, Michael Boskin, issued a report recommending improvements to the nation's economic statistics. Included in the quality improvement were confidentiality issues as well as concern about access to data (Economic Policy Council, 1991).

www.manaraa.com

In 1993, a National Academy of Sciences (NAS) panel on confidentiality and data access recommended that "Statistical records across all federal agencies should be governed by a consistent set of statutes and regulations meeting standards for the maintenance of such records, including the following features of fair statistical information practices: (a) a definition of statistical data that incorporates the principle of functional separation as defined by the Privacy Protection Study Commission, (b) a guarantee of confidentiality for data, …(g) legal sanctions for those who violate confidentiality requirements."(Duncan, Jabine, & Wolf, 1993a) The NAS report is discussed in more detail in chapter 2.

In 1997, OMB issued an "Order Providing for the Confidentiality of Statistical Information" (OMB, 1997). The order applied to twelve designated federal statistical agencies and applied the principles of protection of confidential information collected for statistical purposes. A key element of this was to separate the statistical units functionally from other operational parts of agencies, a principle known as functional separation. After 1997, Congress introduced a number of bills to continue to strengthen the government's ability to protect confidentiality of individual data. These bills culminated in the enactment of CIPSEA in 2002 as part of the EGovernment Act. CIPSEA establishes uniform principles for protecting confidentiality of information collected in surveys and censuses, and allows some limited data sharing between agencies, but it only applies to OMB-designated federal statistical agencies. However, the E-Government Act of 2002 and the Office of Management and Budget (OMB) Circular No. A-11, Exhibit 300, specifically mandate that Federal Agency Privacy Impact Assessments be completed before: (1)

www.manaraa.com

developing or procuring information technology that collects, maintains, or disseminates information that is in identifiable form, or (2) initiating a new collection of information that will collect, maintain, or disseminate information using information technology and includes information in an identifiable form permitting physical or on-line contacting of specific individuals or businesses

Table 2 below summarizes the privacy and confidentiality statutes and regulations that apply to the IRS, Census Bureau, and CMS.

Table 2 Summary of Statutes and Regulations

Authority	Explanation	IRS	Census Bureau	CMS
13 U.S.C. 6	Allows the Secretary of Commerce to get information under title 13 from any other department, agency, or establishment of the Federal Government, or of the government of the District of Columbia.		**XX**	
15 U.S.C. 1552	Authorizes the Secretary of Commerce upon the request of any person, firm, organization, or others, public or private, to make special studies on matters within the authority of the Department of Commerce and to furnish transcripts or copies of its studies, compilations, and other records; upon the payment of the actual or estimated cost of such special work		**XX**	
CIPSEA	Establishes uniform principles for protecting confidentiality of information collected in surveys and censuses, and allows some limited data sharing between agencies,	**XX**	**XX**	**XX**
5 U.S.C. 552(a) (Privacy Act of 1974)	Defines and governs the release and sharing of administrative and statistical records with individual identifying information between agencies and with the public	**XX**	**XX**	**XX**
HHS Regs.(45 U.S.C. 552a)	Implements section 3 of the Privacy Act of 1974 at HHS, by establishing agency policies and procedures for the maintenance of records.			**XX**
26 U.S.C. 6103, 6108	6103 authorizes IRS sharing of tax records with the Census Bureau for purposes under title 13. 6108 requires that statistics reasonably available with respect to the operation of the income tax laws shall be prepared and published annually by the Commissioner.	**XX**	**XX**	
OMB Circular A-11, Exhibit 300	Governs security and privacy issues for government IT systems	**XX**	**XX**	**XX**

www.manaraa.com

Conceptual Framework

The research is a qualitative case study exploration intended to describe the life cycle of data pools that are created by sharing of data records between federal agencies and to illuminate the policy issues raised when the new data pools are created. The life cycle of the data pools will be examined in order to better understand their behavior and characteristics. The research will inform issues surrounding how data are shared and protected.

The rationale for creating a better understanding of the life cycle characteristics of data pools is to contribute to the body of knowledge that will help keep public data available for researchers and public policy analysts while protecting the rights of individuals to keep their information confidential. In addition, there are many opportunities to create new data sets that could provide important information for developing public policies if the barriers to creating and maintaining data pools can be identified and overcome. Data pools are created when data are collected, combined, and stored in a retrievable manner. A data pool may have porous boundaries and a nonfinite life. While pools have been observed empirically for a long time, the theory, behavior, and development of data pools have not yet been defined. A conceptual apparatus describing the data pools needs to be developed in order to be able to develop theory regarding data pools in the future.

Every data pool has a planner, subject matter, and participants. For example, a planner may be trying to determine characteristics of new immigrants to the U.S. by conducting a survey. The subject of the data pool would be the immigrant characteristics, and the participants would the individuals who respond to the survey and become part of

www.manaraa.com

the data pool. The planner often develops a mechanistic approach to data pool management. Thus the pool becomes a function of nonrandom development related to a self conscious setting. Another type of data pool is created by collecting applicant information from individuals eligible for Medicare. This data pool is more complex, with multiple inputs, as original data are collected by states, then sent to CMS. Through ongoing contributions from healthcare providers seeking reimbursement for treating Medicare recipients, the data pool continues to grow. But in this instance, rather than consisting of data collected directly through questioning respondents, data on individuals are collected through multiple third parties. Control over the data characteristics, such as quality, becomes very diffuse, and there are multiple data owners during the life cycle of the data pool.

In addition, data from pools sometimes live on in a residual form, although the pool itself has seemingly been eliminated. For example, the European Union passed Directive 2006/24/EC in March 2006 requiring that member countries ensure that private telecommunications data be retained (including phone, internet, VOIP) for two years by telecommunications companies and be retrievable in order to be made available to law enforcement agencies if necessary (EU, 2006). One may assume that at certain points during that two year period, data will leave one pool (created by the private telecommunications company) and enter another pool (initiated by law enforcement bodies). At the end of two years, the company's pool may be eliminated, but residual data may live on in another form at the law enforcement agencies. Many of the European Union countries have postponed implementing this directive because of these unresolved

www.manaraa.com

issues and their concerns that the directive conflicts with privacy laws enacted by individual European Union member countries (EU, 2006).

Data as a Public Good

One approach to finding the balance between protecting privacy and creating pools of high quality data is to consider data as a public good. Because these data are collected by the government, they may be considered to be a public good. From an economics standpoint, public goods are defined as having one or both of two characteristics: they are nonrival in consumption, and they are nonexclusive (Friedman, 2002). Government-collected data fit this definition. That is, consumption of government-provided data by one researcher does not decrease the amount available for the next researcher. However, data are an impure public good, because they are not nonexclusive. That is, individuals can be excluded from access to certain data. It is because of this characteristic that the privacy and confidentiality of the individuals who provide data can be protected.

As more information becomes available electronically, there is an increased demand worldwide for access to data. Balancing the often conflicting interests of all stakeholders in scientific research—including researchers, publishers, corporations, government, and society can be difficult, especially where the public benefit isn't always clearly defined or doesn't always come first (Romero, 2003). One area that illustrates this conflict is health research, where clinical patient data can be used to provide research statistics that could improve public health (Setness, 2003). For example, the State of New York began collecting data on hospitalized patients in 1979 in order to give report cards on performance to hospitals. This information was used to improve hospital care, and in the area of cardiac bypass surgery, the death rate fell 41% in four years (Lerner, 2002). As a

www.manaraa.com

side result, 37 states began collecting patient data for research purposes, using billing records. Although public health was improved in New York through use of patient records for research, a 1993 survey found that 64% of patients preferred that their medical records *not* be used for research purposes (USMIHS, 2001).

There is a tradeoff between privacy or individual control of personal patient information and data as a public good (Gostin, 2003). Privacy allows the individual to control their personal information. But there are many uses for the information that can benefit the public good, such as informed consumer choice, quality assurance, monitoring fraud and abuse, tracking utilization of health care services, research, and public health activities such as epidemiological investigations. Thus, determining who has access to data is a question of balancing needs (Gostin & James G. Hodge, 2002).

Gostin and Hodge propose three cases that illustrate a method of determining how far disclosure might go in a health environment. The first case is when privacy interests are strong and public interests are weak. In this case, disclosure would only be made to family, friends, the insurer, and possibly the employer. Informed consent on the part of the patient is a key step of disclosure. The second case is when the public interest is strong. This would include research and public health. In this second case, there would need to be a legitimate purpose, there would be no other way to achieve that purpose, and the privacy and security safeguards would need to be strong. The third case is for law enforcement or emergency services that would override the right to privacy.

Steeves (2004) has identified six myths that surround the debate on access to health care records for researchers. They are as follows:

1: Data protection laws restrict access to health information for research

www.manaraa.com

purposes.

2: Research is an unencumbered public good free of any private interest.

3: Privacy is an individual right and so must give way to research as a public good.

4: Observational research data collected without the patient's knowledge and consent will lead to unbiased data.

5: Privacy is a road block to better health.

6: De-identified health information does not pose a risk of harm to the patient.

Of particular interest is her refutation of the argument put forth by researchers, hospital administrators and pharmaceutical companies, who argue that individual privacy rights must not be allowed to constrain medical research because research is a social good that overrules the individual interest in privacy. However, she points out that much research today is in pursuit of economically exploitable intellectual property rights rather than pure science. This raises serious questions about research as a public good. She points out that the worlds' largest seller of health information, IMS, reported revenues of $1.3 billion in 2003, and claims "just about every major pharmaceutical and biotech company in the world" as a client (IMS, 2004).

In addition, she found that many articles published in the British Medical Journal and the Lancet are ghost written by pharmaceutical companies, and clinicians are paid to publish the articles under their own names. This commodification of research information, which creates public distrust and actually limits access to data is currently being investigated by Senator Charles E. Grassley, a member of the Senate Finance Committee, who recently requested information from Wyeth, a large pharmaceutical

www.manaraa.com

company, on drug company payments to ghostwriters and clinicians who agree to lend their names to the articles (Rubenstein, 2008). An example of limiting access to information occurred in Iceland. In 1998, the Icelandic government created the Icelandic Health Sector Database, which contained the genetic information, genealogical history, and health records of everyone in Iceland. Although the database was invaluable to researchers, only one company, deCodeGenetics, was given exclusive access to the database. That company then sold the rights to an American company which licensed access to the Swiss pharmaceutical company Hoffman-LaRoche. Because of these business arrangements, other researchers cannot gain access to the database for 12 years. (Hloden, 2000). This commercialization of research data undercuts arguments that health data are truly a public good.

Organization of Study

The second and third chapters of this study are comprised of a literature review and description of the study's methodology, respectively. The literature review first discusses data sharing and access for research purposes. Then current authorities and activities regarding data sharing among the federal agencies in the case study are reviewed. Literature on the evolution and current state of data stewardship within the Federal government is explored next. Finally, some international approaches to privacy protection and data sharing for statistical and research purposes are discussed for comparative purposes. Chapter three, on methodology, explains how the case studies were designed and describes the limitations of this case study methodology.

Chapter four consists of the results of the data collection and analysis. The chapter describes the life cycle of the data pools created by sharing administrative records

www.manaraa.com

through discussions of the case studies, in order to compare and contrast the two methods for sharing administrative records and protecting privacy. Significant policy issues relating to the creation of the data pools were identified. Chapter five consists of the conclusions and recommendations of the study. The recommendations include ideas for how the U.S. approach for sharing administrative records might be improved by successful practices identified during the research. Areas for additional research are also identified.

Summary

Advances in technology and a hunger for more information have been combining to make possible the creation of huge data pools with information on individuals and businesses. With the creation of these pools have come concerns about protecting privacy and confidentiality and the quality of data being collected. Pressing public policy issues such as universal health care, quality of health care treatments, the state of the economy, and the quality of education have increased the demand for more information arising from combined data sets. When these data sets are generated and tended by the federal statistical system, there are many safeguards in place to prevent the data from being misused. But concerns remain even as the pressure increases for more and better data. This study contributes to the literature on sharing of administrative records and combining these records with survey data and suggests areas for future action and study.

www.manaraa.com

Chapter 2: Review of the Literature

Introduction

In this chapter, literature on protecting privacy and confidentiality of individuals and businesses in the context of data sharing for research purposes is reviewed. In addition, literature on sharing of administrative records between U.S. federal agencies is reviewed, as well as the underlying principles and approaches for data stewardship by federal agencies and researchers with access to data. Finally, this chapter will explore the literature about the formation of the Canadian system for sharing data records and protecting privacy and privacy protections found in the United Kingdom, Australia, and the European Union. Most of this literature is practical and applied in nature, rather than theoretical, that being the primary reason exploratory research is reported in this dissertation.

Data Sharing and Access for Research Purposes

Sharing of data has been a topic of discussion among researchers for the past several decades. The Committee on National Statistics (CNSTAT) of the National Academy of Sciences convened a conference on this issue in 1979, at a time when the use of personal computers was on the rise (Feinberg, Martin, & Straf, 1985). The increased use of computing power in research led to many more opportunities to share and manipulate large data sets. CNSTAT found that there were many problems, controversies, and other consequences of sharing research data, such as the possibility of breaching the confidentiality of individual respondents when data are demanded by law

www.manaraa.com

enforcement authorities (Carroll & Kerr, 1976). Some other examples include the use of business proprietary data in research when the participating business does not want the information revealed, and when researchers plan to market their research and do not want to share their methodology or results because it could impact future profitability. CNSTAT also cited data sharing benefits such as: (1) improving measurement and data collection; (2) developing theoretical knowledge and analytical technique; and (3) encouraging more appropriate use of empirical data in policy formulation and evaluation, among others.

CNSTAT identified the various parties who have a stake in collecting and sharing research data (Feinberg et al., 1985). These parties are the:

(1) initial investigators who first collect the data;

(2) subsequent analysts who analyze one or more data sets collected by others;

(3) scientific community consisting of all scientists engaging in research;

(4) public agencies and foundations that fund research through grants or contracts;

(5) organizations that conduct research such as universities, nonprofit institutions, and commercial enterprises;

(6) respondents to surveys and participants in experiments who have an expectation that their confidentiality will be respected; and

(7) general public that benefits from the result of the research.

Each of these stakeholders has different interests that may, at times, conflict. For example, one of the benefits of sharing data is that several different data sets can be linked to create new data sets against which theories can be tested. One early example of this is the quarterly tapes from the National Crime Survey that were linked to develop

www.manaraa.com

longitudinal data. These data were then used by Reiss (1980) and Eddy, Fienberg, and Griffin (1981) to develop new models and analyses of crime victimization. However, while this most likely contributed to the public interest, it was not the specific use for which respondents originally gave their consent. It's possible that the survey respondents would not have wanted their personal information used in the secondary research. Informed consent by data providers is a topic that will be developed further later in this chapter.

The need for research data to aid in public policy analysis is increasing. The demand for increasingly complex and detailed data is driven and enabled by the rapidly decreasing cost of computing power over the last decade. In 1992, the cost of one terabyte of data storage was approximately $1,000,000. By 2000, the cost had dropped to less than $ 23,284, and by 2007 to $867. It is projected to drop to $211 by 2010 (Gilheany, 2000). This lower cost both allows agencies to collect and hold more data and researchers to conduct more complex studies using large data sets. The additional computing power also affects data confidentiality, as it allows a large number of variables about individuals, linked to geographic data, to be combined and accessed by a wide range of people. In addition, these data can be linked to several open source databases that contain overlapping variables and allow for easier identification of individuals using readily available matching software (Sweeney, 1997) (Winkler, 1998). On the other hand, the computing power also enables sophisticated masking of the data to protect individual identities without compromising the quality of the data (Lane, 2005) (Domingo-Ferrer & Torra, 2001) (Steele, 2005).

www.manaraa.com

Legal Protections

Cecil and Griffin (1982) identified three circumstances involving legal standards governing access to research information. The first circumstance, access to research records maintained by a private researcher supported by private funds, is characterized by the absence of federal support for the research. The second circumstance, access to research records developed with public funds that are maintained by private researchers, describes much university based research. However, it is the third circumstance, access to federal research records maintained by federal agencies, which is of most interest here.

There are several statutes governing the collection and protection of government records containing personal information on businesses and individuals. The Federal Records Act of 1950 (44 U.S.C. §2901 *et seq*), the Privacy Act of 1974 (5 U.S.C. §552a), and the Records Disposal Act of 1976 (44 U.S.C. §3314), as modified by the Freedom of Information Act in 1976 (5 U.S.C. §552) apply to federal executive branch agencies, as well as independent regulatory agencies. (They do not apply to either the judicial or legislative branches of government or to the Executive Office of the President.) Provisions in these laws prevent federal agencies from releasing identifiable research data to researchers if the information is confidential and commercial or financial in nature, or if release of the data might impair the government's ability to obtain necessary information in the future.

In the case of identifiable records, the Privacy Act of 1974 is the primary authority that restricts access to data held by federal agencies. Research and statistical uses of data do not get special treatment under the Privacy Act. However, a *statistical record* is defined by the act as, "a record in a system of records maintained for statistical or reporting purposes only, and not used in whole or in part in making any determination

www.manaraa.com

about an identifiable individual, except as provided by Section 8 of Title 13 (authorizing certain research activities by the Bureau of the Census.)" [5 U.S.C.§552a(a)(4)(1976)] An *administrative record*, on the other hand, is defined as "any item, collection or grouping of information about an individual that is maintained by an agency…and that contains his name, or identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or photograph." [5 U.S.C. § 552a (a) (4) (1976)] While recognizing these differences, the Privacy Act limits the disclosure of identifiable information to third parties *without the prior written consent of the individual for both administrative and statistical records* [5 U.S.C. §552a(a)(6)(1976)].

The notable exception to obtaining written consent, as mentioned above, is in Title 13, which governs data collection by the Census Bureau. Thus, the Census Bureau, when collecting information under Title 13, only needs to inform survey respondents that the information being collected is confidential and will only be used for statistical purposes. Another exemption to the Privacy Act permits disclosure if an agency receives assurance in writing from the recipient that the record will be transferred in a form that is not individually identifiable. Finally, records can be disclosed without consent for a routine use, defined as "a purpose that is compatible with the purpose for which it was collected" [5 U.S.C.§552a(a)(7)(1976)]. Federal agencies have expanded the circumstances under which data can be disclosed without specific prior written consent by using waivers in the original request for information. As mentioned earlier, the issues surrounding informed consent are explored later in this chapter.

More recent legislative attempts to regulate the sharing of personal information include enactment of the Health Insurance Portability and Accountability Act of 1996

www.manaraa.com

(HIPAA), the USA Patriot Act of 2001, and Title V of the E-Government Act of 2002, known as the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). In addition, OMB issued the Federal Statistical Confidentiality Order of 1997, which provided a consistent government policy for protecting the privacy and confidentiality interests of statistical data providers. The Order provides guidance on the content of confidentiality pledges that Federal statistical programs should use under two conditions - first, when the data may only be used for statistical purposes; and second, when the data are collected exclusively for statistical purposes, but the agency is compelled by law to disclose the data. OMB recognized 12 statistical agencies or units in its 1997 Confidentiality Order:

- Department of Agriculture: Economic Research Service and National Agricultural Statistics Service;
- Department of Commerce: Bureau of Economic Analysis and Census Bureau;
- Department of Education: National Center for Education Statistics;
- Department of Energy: Energy Information Administration;
- Department of Health and Human Services: National Center for Health Statistics;
- Department of Justice: Bureau of Justice Statistics;
- Department of Labor: Bureau of Labor Statistics;
- Department of Transportation: Bureau of Transportation Statistics;
- Department of the Treasury: Statistics of Income Division of the Internal Revenue Service; and the
- National Science Foundation: Division of Science Resources Statistics.

www.manaraa.com

Since this guidance was issued in proposed form in October 2006, OMB has recognized two additional statistical organizational units:

- Department of Health and Human Services: Office of Applied Studies within the Substance Abuse and Mental Health Services Administration; and the
- Board of Governors of the Federal Reserve: Microeconomic Surveys Unit.

Subpart A of CIPSEA states that information gathered by federal statistical agencies may not be disclosed in identifiable form for nonresearch purposes without the consent of the respondent. This includes law enforcement, court proceedings, and administrative determinations. These protections extend to data collected by contractors or designated agents on behalf of statistical agencies. Both statistical and nonstatistical agencies can use CIPSEA to protect information they acquire directly from respondents, including State and local governments. However, only statistical agencies or units are authorized under CIPSEA to designate agents to perform exclusively statistical activities, which include data collection, subject to CIPSEA limitations and penalties.

In addition, Subpart B of CIPSEA allows three agencies, the Census Bureau, the Bureau of Economic Analysis, and the Bureau of Labor Statistics to share identifiable business records as long as confidentiality of the records is protected. The exception to this is the Census Bureau's business register, which is constructed from tax records provided by the IRS and protected under section 6103(j)(1)(A) of Title 26 of the United States Code (USC).

www.manaraa.com

While CIPSEA strengthens confidentiality protections as well as allows increased data sharing among agencies, the USA Patriot Act of 2001 greatly decreased the confidentiality protections for educational records kept by the National Center for Educational Statistics (NCES), a part of the Department of Education. According to the NCES website page on data confidentiality and the USA Patriot Act, which can be seen at http://nces.ed.gov/statprog/conflaws.asp:

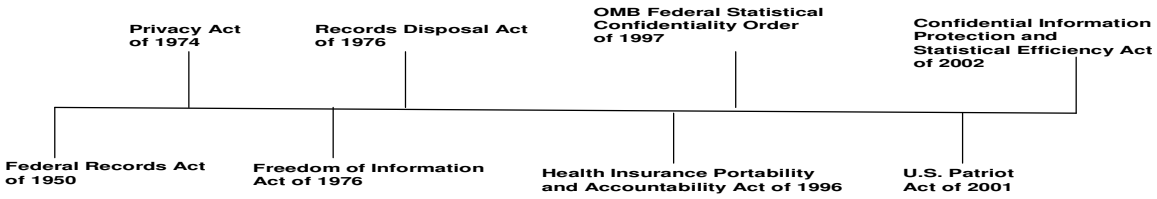
> *"This law amended the confidentiality provisions of NESA 1994 by permitting the Attorney General to petition a Judge for an ex parte order requiring the Secretary of the Department of Education to provide NCES data that is identified as relevant to an authorized investigation or prosecution of an offense concerning national or international terrorism to the Attorney General. Any data obtained by the Attorney General for these purposes must be treated as confidential information, "consistent with such guidelines as the Attorney General, after consultation with the Secretary, shall issue to protect confidentiality.*
>
> *As a result of the Patriot Act, the intended use clause of the NESA of 1994 was amended. That is, the portion of the NESA of 1994 that specified that data collected by NCES may only be used for statistical purposes was amended by the fact that the data may now be used with a judge's order for matters relevant to an offense concerning national or international terrorism. This amendment was incorporated into ESRA 2002.*
>
> *It is important to note that the confidentiality of data collected by NCES is protected in all instances, since even in the case of a judge's order for*

www.manaraa.com

matters relevant to an offense concerning national or international terrorism, the Attorney General must protect the confidentiality of the data."

Thus the NCES stretches the boundaries of confidentiality by asserting that even though confidential information collected for statistical and research work can be shared with the Department of Justice for law enforcement actions against individuals in the database, confidentiality is protected because the Attorney General is protecting the confidentiality of the information. It is not clear in either the legislation or on the website how this is accomplished nor is this new definition of confidentiality explored by NCES. Figure1 shows a time line of when significant privacy legislation was enacted.

Figure 1 Privacy Legislation Timeline



Federal Administrative Record Sharing

The Census Bureau took the lead in exploring the use of administrative records within the federal statistical system after the 1990 decennial census. Record sharing was one avenue that was explored as a means of trying to lessen the large undercount of

www.manaraa.com

certain segments of the U.S. population that occurred (Eddy et al., 1981; Obenski & Prevost, 2004). During the 1990s, the Census Bureau held several conferences on this topic, conducted a survey to identify administrative records files that could be of use, developed a prototype called the Statistical Administrative Records Database (StARS), and conducted a simulated administrative records census. StARS integrated files from IRS and CMS, as well as the Indian Health Service (IHS), Housing and Urban Development (HUD), the Selective Service, and the Social Security Administration (SSA). The StARS database, which contained demographic and address information, was used to simulate a census that would answer the questions on the decennial census short form that is given to every U.S. household, such as number of people in the household; their relationship to each other; age; sex; and race and ethnicity. The 1990 census long form, consisting of about 56 questions, was given to a much smaller subsample of the population. According to Obenski and Prevost (2004), this effort, called the Administrative Records Experiment or AREX, matched about 85% of the addresses in the administrative records files to addresses in the census, using a sample from two counties in Maryland and three counties in Colorado (about one million households). StARS continues to be refreshed by the Census Bureau and currently is used as the basis for numerous research projects.

The 2007-2012 strategic plan of the Census Bureau, available on its website at www.census.gov/main/www/strategicplan/strategicplan.pdf , contains a strategic objective on pg 7 to "support innovation, promote data use, minimize respondent burden, respect individual privacy, and protect the confidentiality of respondents' information." A sub-objective within that strategic objective is to, *"Minimize reporting burden and cost*

www.manaraa.com

to taxpayers by acquiring and developing high-quality data from sources maintained by other government and commercial entities." (p. 7). The strategic plan states that the Census Bureau is required by law to use existing information whenever possible, rather than conducting primary data collection. In addition to the exemptions from the Privacy Act, 13 U.S.C.§ 6(a)(c) requires of the Secretary of Commerce that, *"To the maximum extent possible and consistent with the kind, timeliness, quality and scope of the statistics required, the Secretary shall acquire and use information available from any source referred to in subsection (a) or (b) of this subsection instead of conducting direct inquiries."*

Further, the Census Bureau strategic plan asserts that using administrative records collected from other agencies enhances data quality, improves data products, saves taxpayer money, and minimizes reporting burden. The strategic plan lays out the following actions for the Census Bureau to undertake in order to develop its administrative records capacity:

- Develop a Census Bureau-wide plan on the role of administrative records in censuses and surveys;
- Develop and disseminate Census Bureau-wide policy guidance and security/disclosure avoidance procedures that ensure both the appropriate acquisition and use of administrative records and the delivery of products that incorporate administrative record information; and
- Establish and maintain relationships with administrative record source agencies, program sponsors, the statistical community, and the general public that support

www.manaraa.com

the Census Bureau's expanded use of administrative records to produce timely, high quality, low-cost statistics.

A second sub-objective is to, *"Foster trust and cooperation of the public by respecting privacy and protecting the confidentiality of respondents' information.*" (p.8) This includes:

- Enhancing the Census Bureau-wide privacy and confidentiality program to fully integrate data stewardship policies and practices across all programs; and
- Continuing to assess possible disclosure risks in data products and develop methodologies to address any concerns.

The strategic plan reflects the widespread usage of administrative records data by the Census Bureau. According to Prevost (2001) the Census Bureau uses information from administrative records for both business and person and household information. On the business side, administrative record data are used to manage respondent burden, improve survey quality, and reduce costs by eliminating the need to classify industries before the economic census takes place. The largest use is to build the business registry of all the businesses in the U.S. to construct a sampling frame. The data for the business registry, known as the Standard Statistical Establishment List (SSEL) are provided by the IRS from tax forms.

On the person and household data side, the Census Bureau relies less on administrative records. However, one major use of administrative records was the sharing of postal addresses by the U.S. Postal Service (USPS) to help create the list of the geographic location of every address in the United States that was used to conduct the 2000 decennial census of population and housing (Census 2000). This list is the Master

www.manaraa.com

Address File or MAF. While the sharing of the electronic postal address list (called the Delivery Sequence File or DSF) was intended to save tens of millions of dollars by eliminating the need to send people on foot to canvass neighborhoods to gather addresses, the USPS lists were not of consistently high quality across the United States, so the Census Bureau still had to collect and verify addresses manually by walking through neighborhoods and mapping addresses with paper and pencil

Another major use of administrative records has been for the Longitudinal Employer-Household Dynamics Project (LEHD), now part of the Census Bureau's Local Employment Dynamics program. As mentioned in Chapter 1, the LEHD combines state and federal data on employers and employees with other Census Bureau data to create entirely new integrated data sets that are longitudinal and provide substantial research opportunities to explore labor economics-related areas (Lane & Stephens, 2006). The LEHD program will be explored in more depth as part of the case study.

The Census Bureau is not the only agency that shares administrative records. A major source of such data is CMS, which sponsors and conducts hundreds of research projects that provide outside researchers access to CMS Medicare and Medicaid records. During 2007, CMS sponsored more than 600 active individual research, demonstration, and evaluation projects (CMS, 2007).

According to the CMS website (http://www.cms.hhs.gov/MedicaidDataSourcesGenInfo), the primary data sources for Medicaid statistical data are the Medicaid Statistical Information System (MSIS), the Medicaid Analytic eXtract (MAX) files, and the CMS-64 reports. MSIS is the basic source of state-submitted eligibility and claims data on the Medicaid population, their characteristics, utilization, and payments. The Medicaid

www.manaraa.com

Analytic eXtract (MAX) data – formerly known as State Medicaid Research Files (SMRFs) – are a set of person-level data files derived from MSIS data on Medicaid eligibility, service utilization and payments. Data are available for all states and the District of Columbia beginning with calendar year 1999. Data are available for selected states prior to 1999. The CMS-64 reports are products of the Medicaid and SCHIP Budget and Expenditure Systems (MBES/CBES), the financial budget and grant systems.

MAX data are developed to support research and policy analysis initiatives for Medicaid and other low-income populations. MAX data for 1999 have been used to develop a series of research products related to pharmacy benefit use and reimbursement in Medicaid. These products include a Statistical Compendium of detailed statistics, by state; a Chartbook of Medicaid pharmacy benefit use and reimbursement; and a summary of major Medicaid pharmacy benefit features for 1999, by state. In response to the high demand from researchers for access to microdata, CMS has established the Research Data Assistance Center (ResDAC), a CMS contractor that provides free assistance to researchers interested in using identifiable Medicare and/or Medicaid data for their research.

IRS is highly restricted by law in how it can share administrative records containing identifiable information. Much of its research is conducted through the Statistics of Income Division, or SOI, which shares statistics on individuals, businesses, estates, nonprofits, trusts, and foreign investment. The information is used by a variety of federal agencies, academics, researchers, and the public. It's used to analyze tax policy, project tax revenues, and estimate the overall impact of tax law changes and their effects on tax collections. The primary clients of SOI are the Office of Tax Analysis

www.manaraa.com

(OTA) in the Secretary of the Treasury's Office and the Congressional Joint Committee on Taxation (JCT) – each of whom is entitled to receive detailed tax return files. Most other agencies and individuals can only access data in the aggregate to protect individual privacy as described in Section 6103 of the Internal Revenue Code. The SOI website lists some of the clients as follows:

- The Department of Commerce's Bureau of Economic Analysis, the Federal Reserve Board, the General Accounting Office (currently named the Government Accountability Office), the Social Security Administration, and the Health Care Financing Administration (currently named the Centers for Medicare and Medicaid Services or CMS); and
- Tax practitioners, policy researchers, demographers, economic analysts, consultants, business associations, State and local Governments, universities, public libraries, and the media.

Interestingly, the IRS website does not specifically mention the Census Bureau. A look at some of the presentations at IRS annual research conferences indicates that many researchers from a wide variety of institutions are granted access to IRS microdata files to conduct research. These papers are available on the IRS website (IRS, 2006).

Data Stewardship

In 1993, CNSTAT and the Social Science Research Council asked the Confidentiality and Data Access Committee (CDAC) to examine data stewardship among federal statistical agencies and develop guidelines for protecting confidentiality and privacy (Duncan et al., 1993a). CDAC identified four broad categories into which it

www.manaraa.com

placed its recommendations: (1) statutory protection against mandatory disclosure of individually identifiable data; (2) barriers to data sharing within government; (3) privacy concerns and declining cooperation in surveys; and (4) statistical procedures to protect confidentiality. The panel affirmed that data collected for research or statistical purposes should not be made available for administrative actions involving an individual.

The panel recommended development of a consistent set of statutes and regulations that guaranteed confidentiality of statistical data, which played heavily into the subsequent enactment of the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). The panel recommendation built on the work done previously by groups such as the Ad Hoc Committee on Privacy and Confidentiality of the American Statistical Association (1977), which recommended that legislation be enacted covering all federal agencies providing full and overriding protection against compulsory disclosure of survey and administrative record data that could identify individuals when the data were collected for statistical purposes.

Around the same time, the Privacy Protection Study Commission (1977a) recommended that a clear functional separation be maintained between the use of information for research and statistical purposes and its use for administrative purposes. In addition, the Commission recommended that an independent agency be established within the federal government to monitor privacy issues and issue rulings on the Privacy Act of 1974. In 1978, the Office of Federal Statistical Policy and Standards (OFSPS), then a part of the Department of Commerce, recommended in its report, *A Framework for Planning U.S. Federal Statistics for the 1980's* (1978), that agencies collecting statistical and research data have statutory protection for maintaining the confidentiality of both

www.manaraa.com

commercial and personal data. The report additionally called for the establishment of what it called "protected enclaves" in selected statistical and research agencies within major departments that would be insulated from political intervention and unauthorized access to data. To help assure this, agency employees would be subject to strict ethical standards regarding data handling and to penalties for unauthorized disclosure of confidential data.

The 1993 CNSTAT panel report also supported the concept of an independent federal advisory body that would be charged with fostering a climate of enhanced protection for federal data, including data dissemination and promoting the principle of functional separation embodied by the "protected enclaves". The panel recommended that each agency review its staffing and management procedures and policies and assign responsibility for these areas to qualified, identified individuals. Agencies were also urged to train staff in confidentiality issues such as informed consent, disclosure limitation, etc.

Other recommendations: (1) encouraged the sharing of potentially identifiable personal data among federal agencies for statistical and research purposes, including development of sampling frames; (2) encouraged statistical agencies both to continue widespread release of microdata sets with as few restrictions as possible and to increase access for outside researchers; (3) sought strengthened legal sanctions for violations of confidentiality requirements; (4) directed statistical agencies to undertake continuing research to monitor the views of data responders and the general public on informed consent, data sharing, and related issues; (5) supported ongoing research into statistical disclosure limitation analysis coordinated by OMB; and (6) directed that agencies make

www.manaraa.com

sure that disclosure limitation wasn't so effective that the microdata made available to researchers was not useful (ascertained through their own research and by consulting outside stakeholders).

The 1993 panel also stated that data providers needed to be informed if their data were to be used for anything other than statistical purposes. Basic information to be given to data providers included information to meet Privacy Act requirements; a statement on the amount of time required to supply the requested data; no false promises, such as a 100% guarantee of nondisclosure; information about any *planned or potential* nonstatistical uses of the data provided; information about any *planned or anticipated* record linkages for statistical and research purposes; a statement to cover all unanticipated future uses of the data; and information about how long the data would be retained by the government in identifiable form. However, the panel agreed with the view of the1978 OFSPS report that if interagency transfers of identifiable data were prohibited unless explicit consent were obtained from the data providers, many valuable studies would be eliminated. Thus, the CNSTAT panel recommended a set of guidelines to help determine reasonable levels of informed consent.

A decade later, CNSTAT revisited the issue of expanding access to research data, particularly linked longitudinal microdata, while preserving respondent confidentiality in its report, *Expanding Access to Research Data: Reconciling Risks and Opportunities* (2005). The report focused on assurances given to data providers about how their data would be used and how well protected they are against unauthorized disclosure of personal information. Several factors were cited to demonstrate why a reexamination of the issues in the 1993 report was warranted. Among other things, the report mentioned

www.manaraa.com

increased public concern over privacy issues; public unease over commercial technology gathering large amounts of personal information into large databases; the need for such databases in order to conduct longitudinal research to answer complex policy questions; new types of data becoming available in large databases such as genetic and financial information; the inability of statistical agencies to conduct in-depth research into social issues, therefore requiring the involvement of outside researchers; advances in information technology raising fears of privacy intrusion as well as providing greatly improved research tools; development of new techniques for altering data to protect individual identities; and a changed legal framework.

More recently, the National Science Foundation, through its Information Technology Research Program, funded a five year project that began in 2004 called Privacy, Obligations, and Rights in Technology of Information Assessment (PORTIA)(Portia, 2004). The project has many participants from academia, the federal government, private industry, and the nonprofit sector, including Stanford, Yale, New York University, Rutgers, the Census Bureau, the Secret Service, the National Institutes of Health, Microsoft, Google, Citigroup, and the Electronic Privacy Information Clearinghouse (EPIC). The project consists of multiple research and education projects that examine both technological methods for protecting personal identities on-line when using search engines and visiting web sites, and the philosophical and policy approaches to determining how and to what extent individual privacy should be protected. The theme is to preserve both the ability to collect and mine very large data sets and respect individual privacy.

www.manaraa.com

Although most of the projects are studying situations outside the federal statistical system, such as when an individual is web surfing and unbeknownst to the user, various web service providers are collecting information on the sites visited and activities carried out, a few of the papers were relevant to how agencies handle records of individuals. A case study of the Census Bureau (Weber, 2005) examined the relationship between the Census Bureau, technology available for processing large amounts of data collected in the decennial census of population, and understandings of entities outside of government regarding uses of the data. The case study used the public outcry that occurred when the Census Bureau provided data to the Department of Homeland Security that it requested on urban areas with more than 10,000 inhabitants reporting Arab descent and further refined that with tabulations by country of origin and zip code. Although this information was actually publically available, there were numerous objections from the Arab-American community and others (Clemetson, 2004). In fact, in order to quell the outcry, the Census Bureau issued a new policy on providing custom data tabulations that restated what the Census Bureau already was doing (Census, 2005).

The incident highlighted an instance where information was given by members of the public for one purpose (the census) but was going to be used for an entirely different purpose (homeland security). While the information was available to any member of the public, because identities of individuals could not be determined from the data, the incident demonstrated how technology makes it very easy to provide information out of the original context in which it was provided, and that the context in which the data are being used can be very important to the perception of whether privacy norms are being violated.

www.manaraa.com

The Portia Project also examined the concept of privacy as contextual integrity (Nissenbaum, 2004). This philosophical approach asserts that the notion of adequate protection of privacy is tied to the specific context in which the information is provided and used. Nissenbaum proposes four constructs for this alternative benchmark: (1) informational norms, (2) appropriateness, (3) roles, and (4) principles of transmission. Informational norms describe how society might expect that the information be handled. Appropriateness and roles describe what type of information is being transmitted and who the transmitters are. For example, it may be perfectly all right to discuss personal health problems with your physician, but you would not discuss these same things with the human resources director at work. Similarly, a doctor may ask a patient the sort of questions that it would be inappropriate to ask the doctor's office receptionist. Finally the principles of transmission address how the information is transmitted including past and future actions by the subject and the user. That is, did the subject give permission for the information to be used? Is confidentiality a future requirement of the user? Taking these concepts further, she and her colleagues developed a mathematical model that is much more complex and flexible than current access control systems (Barth, Datta, Mitchell, & Nissenbaum, 2006). These current systems, including the Platform for Privacy Preferences (P3P), Enterprise Privacy Authorization Language (EPAL), and Role Based Access Control (RBAC) do not use information about the past or require future restrictions. By contrast, the contextual norm model is more complex because it goes beyond such information as who "owns" the data or whether it is public or private. Rather it assigns roles to entities which become key variables, such as the need to know, whether there is a two-way transmission of information, confidentiality, whether the

www.manaraa.com

respondent is forced to share the information, and whether the respondent knows the information is being shared. Using a contextual model could be illuminating in the situation where federal agencies are sharing administrative records for statistical purposes, distinguishing these activities from law enforcement, for example.

Census Bureau Data Stewardship

Data collected by the Census Bureau under Title 13 have special protection, which allows the Census Bureau to acquire limited consent from respondents to its surveys and censuses. However, the Census Bureau also collects information from the public on behalf of other federal agencies, using other authorities, primarily Title 15. The Census Bureau uses similar consent language on both the reimbursable surveys conducted on behalf of other agencies and the surveys it conducts on its own behalf, such as the Survey of Income Participation (SIPP) and the Survey of Program Dynamic (SPD). An example of the language provided to respondents of the National Crime Victimization Survey (NCVS), funded by the Bureau of Justice Statistics, is included in the U.S. Census Bureau, NCVS Interviewing Manual for Field Representatives from 2003 and is worded as follows:

> *"The Census Bureau is conducting the National Crime Victimization Survey for the Bureau of Justice Statistics of the United States Department of Justice. The survey's purpose is to provide information on the kinds and amount of crime committed against households and individuals throughout the country. All survey information will be used for statistical purposes only. This survey is authorized by Title 42, Section 3732 of the United States Code."*

www.manaraa.com

Table 3 shows a summary of surveys conducted by the Demographic Surveys Division of the Census Bureau during 2006 and 2007, as well as the authorities under which the data were collected (DSD, 2007). The costs of these surveys are reimbursed by the sponsoring agencies. The table does not include surveys and censuses conducted of business establishments by the Economic Directorate of the Census Bureau. The surveys below are listed to illustrate three things: (1) data are collected from a wide variety of respondents of all ages and include collections from administrative records; (2) the topics covered are of interest not only to federal policy makers but to a wide variety of researchers and other stakeholders; and (3) the data are collected under different authorities, even when the sponsors are the same for multiple surveys. Detailed information about each of the surveys is in Appendix I.

Table 3 Census Bureau Reimbursable Demographic Surveys Summary

Agency Sponsor	No. of Surveys	Sponsoring Authority	Census Bureau Authority
Housing and Urban Development (HUD)	2	Title 12	Title 13
New York City	1	Local Code	Title 13
Bureau of Labor Statistics (BLS)	4	Title 29	Title 13
Bureau of Labor Statistics (BLS)	2	Title 29	Title 15
National Center for Education Statistics (NCES)	5	Title 1	Title 15
National Center for Education Statistics (NCES)	1	Title 42	Title 13
National Science Foundation (NSF)	2	Title 42	Title 13
Bureau of Justice Statistics (BJS)	2	Title 42	Title 13
Bureau of Justice Statistics (BJS)	2	Title 42	Title 15
National Center for Health Statistics (NCHS)	4	Title 42	Title 15
Fish and Wildlife Service (FWS)	1	Title 16	Title 13
National Institutes of Health (NIH)	1	Title 42	Title 13
National Institutes of Health (NIH)	1	Title 42	Title 15
Corporation for National and Community Service (CNCS)	1	Title 45	Title 13

www.manaraa.com

In keeping with the spirit and intent of the E-Government Act of 2002 and the Office of Management and Budget (OMB) Circular No. A-11, Exhibit 300, which mandate preparation of Federal Agency Privacy Impact Assessments (PIA), the Census Bureau unveiled its Privacy Principles to the public in 2003, setting the ethical standards for data collection, handling, and dissemination (Census Bureau, 2006). The Privacy Principles apply to all phases of a project or activity (planning, design, collection, processing, dissemination, and archiving) involving censuses and surveys authorized by Titles 13 and 15, United States Code, for all types of economic, demographic, and decennial census data. They are listed in Appendix II. In addition, Appendix III shows how these principles are presented to the public, particularly survey and census respondents. The Census Bureau also makes its PIAs available to the public on its website, found at http://www.census.gov/po/pia .

CMS Data Stewardship

CMS data stewardship differs from that of the Census Bureau, driven in part because it is not a formally designated statistical agency. Rather, its primary mission is to administer the Medicare and Medicaid programs. Thus, the data it collects are, in large part, records provided by the 50 states that participate in the program. One example of how CMS protects data is in how it handles its Long Term Care Minimum Data Set (LTCMDS). The purpose of this system of records is to aid in the administration of the survey and certification and the payment of Medicare Long Term Care services, which include skilled nursing facilities, nursing facilities, and hospital swing beds, and to study the effectiveness and quality of care given in those facilities. To remain in compliance with the Privacy Act, in 2002, CMS updated information originally published in the

www.manaraa.com

Federal Register in 1998 (CMS, 2002) on how it planned to provide access to identifiable information in its LTCMDS database. According to CMS, information in the system is used to support: (1) regulatory, reimbursement, and policy functions performed within the Agency or by a contractor or consultant; (2) another Federal or state agency, agency of a state government, an agency established by state law, or its fiscal agent; (3) Peer Review Organizations (PRO); (4) other insurers for processing individual insurance claims; (5) research on the quality and effectiveness of care provided, as well as payment related projects; (6) constituent requests made to a congressional representative; (7) litigation involving the Agency; (8) combating fraud and abuse in certain health benefits programs, and (9) national accrediting organizations. The routine uses of the records, which are circumstances under which CMS may release information from the LTCMDS *without the consent of the individual* include:

1. to CMS contractors, or consultants assisting in accomplishment of a CMS function relating to the purposes of this system

2. to another Federal or state agency, agency of a state government, an agency established by state law, or its fiscal agent to:

   a. Contribute to the accuracy of CMS's proper payment of Medicare benefits,

   b. Enable such agency to administer a Federal health benefits program, or as necessary to enable such agency to fulfill a requirement of a Federal statute or regulation that implements a health benefits program funded in whole or in part with Federal funds, and/or

   c. Assist Federal/state Medicaid programs within the state.

www.manaraa.com

3. to PROs in connection with review of claims, or in connection with studies or other review activities, and in performing affirmative outreach activities to individuals for the purpose of establishing and maintaining their entitlement to Medicare benefits or health insurance plans.

4. to insurance companies, underwriters, third party administrators (TPA), employers, self-insurers, group health plans, health maintenance organizations (HMO), health and welfare benefit funds, managed care organizations, other supplemental insurers, non-coordinating insurers, multiple employer trusts, other groups providing protection against medical expenses of their enrollees without the beneficiary's authorization, and any entity having knowledge of the occurrence of any event affecting (a) an individual's right to any such benefit or payment, or (b) the initial right to any such benefit or payment, for the purpose of coordination of benefits with the Medicare program and implementation of the Medicare Secondary Payer (MSP) provision at 42 U.S.C. 1395y (b).

As shown above, the "routine uses" clause in the Privacy Act allows a broad array of people in different organizations inside and outside of government to have access to records containing personal medical information with identifiers attached without the consent of the individuals whose records are being shared. Prior to the 2002 revision in the Federal Register, CMS had specifically mentioned the Census Bureau as a routine user of CMS data. However, citing Exception 4 to the Privacy Act, which allows release of data to the Census Bureau under Title 13, sharing of records was subsumed under

www.manaraa.com

routine use 2a. Thus, Census Bureau use of CMS data must contribute to the accuracy of CMS benefit payments in addition to any other statistical purposes of the Census Bureau.

In the statement to Medicare beneficiaries that is required by the Privacy Act (known as the Privacy Act Statement and found at www.cms.hhs.gov/MinimumDataSets20/Downloads/MDS%20Privacy%20Act%20Statement.pdf), CMS cites the Social Security statutes that allow it to collect social security numbers as well as describes the routine uses of the information, and specifically mentions that the information will be shared with the Census Bureau. The Privacy Act Statement specifies that it is not a consent statement, but simply sharing information. CMS has posted its general privacy principles on its web site. http://www.cms.hhs.gov/PrivacyOffice/03_Privacy_BasicPrinciples.asp#TopOfPage

IRS Data Stewardship

Unlike the Census Bureau, whose primary mission is to collect information about the people and economy of the U.S., the IRS mission is to collect taxes and enforce tax laws. Research and data stewardship are important, but secondary activities. However, data stewardship at the IRS is very important to the taxpaying public. The IRS web site, http://www.irs.gov/privacy/index.html , states that it is committed to protecting the privacy rights of America's taxpayers. But the IRS continues to be plagued with problems of data privacy breaches. Sixty percent of Internal Revenue Service employees in an audited sample of employees were duped into giving control of their passwords to unauthorized callers posing as help desk employees, according to a 2007 inspection report on computer security conducted by the Treasury Inspector General for Tax Administration (WebCPA, 2007). Additionally, IRS employees are periodically charged

www.manaraa.com

with illegally examining the confidential tax records of individuals, sometimes on hundreds of occasions ((TIGTA, 2008).

Nevertheless, the IRS 2000-2005 Strategic Plan (IRS, 2001) includes agency guiding principles, one of which is to demonstrate effective stewardship of assets and information entrusted to the IRS. According to page 14 of the IRS plan, this means, "We must accurately account for taxpayer funds, use our budget funds efficiently and for the purpose intended, manage and account for our inventory of property and equipment, and safeguard taxpayer information." This translates into an IRS strategy to promote effective information stewardship, partially demonstrated by improving internal processes for information management. In addition, the IRS acknowledges that privacy and security are major concerns for both the IRS and the taxpaying public. On page 77 of its strategic plan the IRS states that:

> "We are committed to recognizing taxpayer privacy to the maximum extent possible in all Service initiatives. Given the vulnerability of modern electronic information systems to cyber attacks, hacking, and natural disaster, we are focusing resources on: risk management processes; secure messaging and authentication; physical security; cyber attack response capability; and disaster recovery measures."

The IRS also states on page 77 that,

> "We will incorporate privacy protection principles into all IRS programs and policies. We will enhance the privacy impact assessment methodology, applying it to all stages of a system's development and requiring it as a part of a system's certification."

www.manaraa.com

Similar to other Federal agencies, the IRS performs the Privacy Impact Assessments (PIAs) required by OMB Circular A-11 on its computer systems and applications in order to evaluate any risks these systems may pose to personally identifiable information. On the IRS website, http://www.irs.gov/privacy/article/0,,id=160742,00.html , there are over 300 PIAs available to the public covering the various IRS automated systems (IRS, 2007a). Many of these systems have nothing to do with data collected from the public, such as the Employee Satisfaction Tracker and the Employee Training Database. However, several do address systems containing taxpayer data that are personally identifiable. In addition, the IRS has a Privacy Advocate who develops policies to protect taxpayer and IRS employee privacy and ensures that they are integrated into all IRS practices and modernization efforts. The Privacy Advocate also ensures that taxpayers and employees are aware of their privacy rights. The IRS on-line privacy policy, found at http://www.irs.gov/privacy/index.html primarily addresses privacy on the website as it pertains to visitors. It does not specifically address protecting the confidentiality of taxpayer data, except by reference to Title 26, the Privacy Act, and FOIA.

Outside Stakeholders

There are a number of privacy advocacy organizations that track government and private sector activities related to privacy of individual records. The organizations listed below are national in scope, well recognized, and particularly interested in government records:

www.manaraa.com

- The Electronic Privacy Information Center (EPIC), which was established in 1994 to focus public attention on emerging privacy issues relating to the National Information Infrastructure, such as the Clipper Chip, the Digital Telephony proposal, medical records privacy and the sale of consumer data. EPIC conducts litigation, sponsors conferences, produces reports, publishes the EPIC Alert and leads campaigns on privacy issues.
- The American Civil Liberties Union (ACLU), which was originally founded in 1920. The ACLU conducts extensive litigation on Constitutional issues including privacy and free speech.
- The Privacy Coalition, a nonpartisan coalition of consumer, civil liberties, educational, family, library, labor, and technology organizations in support of legislation that effectively protects personal privacy.
- The US Privacy Council, a coalition of US privacy groups and individuals founded in 1991 to deal with privacy issues in the US. USPC works in Washington, D.C. monitoring legislation and the activities of government agencies.

However, most of these private organizations are not focused on sharing of government held administrative records for statistical purposes. They are concerned, however, about whether agencies are sharing information for law enforcement or other such purposes. For example, EPIC filed suit against the Department of Homeland Security and the Department of Commerce when it thought that the Census Bureau had shared information with DHS on the location of Arab Americans as reported in the 2000

www.manaraa.com

Census (EPIC, 2007). As mentioned earlier in the context of the Portia Project paper, DHS did, in fact, request this information from the Census Bureau, and the Census Bureau provided publicly available information in a customized table that showed number of people with Arab ancestry by zip code. While this information was publicly available, it raised concerns in the privacy community before it was learned that only public information that had no personal identifiers in it was provided by the Census Bureau to DHS.

International Data Protection

While this study does not include any international case studies, it is worthwhile to briefly review the privacy policies of Canada, the European Union (EU) and Australia. These are useful for contrast, because the countries involved have centralized statistical bureaus, rather than the fragmented statistical system found in the U.S. The centralized approach eases the way for sharing of administrative records between operational agencies and the statistical bureaus, because the context in which the records will be used is clear; that is, for statistical purposes.

Canada

Canada has enacted two federal privacy laws: the *Privacy Act*, and the *Personal Information Protection and Electronic Documents Act (*PIPEDA*)*. The *Privacy Act*, in place since 1983, protects the personal information collected by government institutions. It is overseen by the Privacy Commissioner of Canada, who has the authority to investigate complaints. PIPEDA applies to private sector organizations that handle personal data. In addition, every province and territory in Canada (except for

www.manaraa.com

Newfoundland) has guidelines to protect personal information held by government departments and agencies. These acts are based on a set of voluntary fair information practices which were agreed in 1980 at the Organization for Economic Cooperation and Development in Paris, also known as the OECD Guidelines. The Acts are administered and overseen by an independent commissioner or ombudsperson, with the authority to investigate complaints. Generally, Canada's privacy acts require that personal information is:

- collected by government institutions only in direct relation to operating programs or activities;
- collected from the individual him or herself;
- accurate and up-to-date;
- retained to allow affected individuals the opportunity to gain access to it;
- used only for the purpose for which it was collected or a related purpose (or one of a number of specific purposes); and
- able to be corrected by the individual concerned.

Statistics Canada, the national statistical bureau of Canada, was created by legislation (*Statistics Act*. 1970-71-72, c. 15, s. 1.) in order to collect, compile, analyze, abstract and publish statistical information relating to the commercial, industrial, financial, social, economic and general activities and condition of the Canadian people. To carry out these duties, Statistics Canada collaborates with other government departments to collect information, including statistics derived from the activities of those departments. Part of the mandate of Statistics Canada is to assure that there is not

www.manaraa.com

duplication in the information collected by the government as well as to promote and develop integrated social and economic statistics for each of the provinces and for the country as a whole (1970-71-72, c. 15, s. 3). In addition to collecting information from other national departments, Statistics Canada may also enter into agreements with the statistical agencies of provincial governments to share information and administrative records. However, this is limited to statistical agencies that have the same requirements for safeguarding the confidentiality of the data that are collected (1970-71-72, c. 15, s. 11).

Statistics Canada may enter into agreements to share data with any department or municipal or other corporation collected from a respondent on behalf of both of the parties to the agreement. Some examples of record sharing between Statistics Canada and other departments include: (1) tax returns, certificates, statements, documents or other records, which are acquired from the Minister of National Revenue (R.S., 1985, c. S-19, s. 24; 1990, c. 45, s. 54); (2) returns of imports and exports and details of the means of transportation, which are acquired from the Solicitor General of Canada (1970-71-72, c. 15, s. 23; 1976-77, c. 28, s. 41; 2005, c. 38); and schedules relating to criminal business transacted in courts or tribunals, penitentiaries and reformatories, as well as pardons (1970-71-72, c. 15, ss. 24, 25, and 27). If an agreement is in place, Statistics Canada must inform the respondent at the time it is collecting the data with which agencies the data will be shared. The respondent may object in writing, which would preclude that respondent's information from being shared unless otherwise required by law (1970-71-72, c. 15, s. 12). The statistical information collected and maintained by Statistics Canada is considered privileged under the law and can't be used for law enforcement

www.manaraa.com

purposes or as evidence in legal proceedings. No person sworn to protect the data under section 6 can be required by a court, tribunal or other body to give testimony or to produce documents obtained under the Act (1970-71-72, c. 15, s. 18).

The Act contains safeguards for protecting information that is collected from individuals, businesses, and other agencies. Only employees or contract employees have access to data that is in its original form; that is, individual identifiers are still included in the data. In addition, providers of the information, such as provincial agencies that have entered into agreements with Statistics Canada to collect certain information, may maintain access to the original data. Abusers can be fined or jailed (1970-71-72, c. 15, s. 6). However, the Chief Statistician of Statistics Canada may, by order, authorize disclosure of information relating to a person, business or organization if disclosure is consented to in writing by the person, business owner, or organization concerned. Other exceptions include: (1) information available to the public under any statutory or other law; (2) information relating to any hospital, mental institution, library, educational institution, welfare institution or other similar non-commercial institution as long as any individual patient, inmate or other person in the care of any such institution can't be identified; and (3) lists of individual establishments, firms or businesses, showing names addresses, phone numbers, products or services provided, range of number of employees, and official language for doing business; and (4) information relating to any carrier or public utility.

Record linkage is an important technique used by Statistics Canada to develop and analyze data. In its policy on record linkage (Statistics Canada, 2000), record linkage is defined as bringing together of two or more micro-records to form a composite record.

www.manaraa.com

A micro-record is defined as a record containing information about an identifiable individual respondent or unit of observation (e.g., person, family, household, dwelling, farm, company, business, establishment, institution, etc.). Under its policy, which can be found at http://www.statcan.ca/english/recrdlink/policy4-1.htm, record linkages will be made, "only if the following conditions are satisfied:

- the purpose of the record linkage activity is statistical/ research and is consistent with the mandate of Statistics Canada as described in the *Statistics Act*; and
- the products of the record linkage activity will be released only in accordance with the confidentiality provisions of the *Statistics Act* and with any applicable requirements of the *Privacy Act*; and
- the record linkage activity has demonstrable cost or respondent burden savings over other alternatives, or is the only feasible option; and
- the record linkage activity will not be used for purposes that can be detrimental to the individuals involved and the benefits to be derived from such a linkage are clearly in the public interest; and
- the record linkage activity is judged not to jeopardize the future conduct of Statistics Canada's programs; and
- the linkage satisfies a prescribed review and approval process."

Statistics Canada maintains on its website a list called *Info Source* that lists all the linked databases maintained by the agency. All projects using linked databases are

www.manaraa.com

also included in an annual report to the Parliament that is required under Canada's *Privacy Act*. The guiding principles for programs involving record linkage include requiring the linkage to be in the public interest and to provide insights about a specific issue. The public good to be served is assessed by a series of reviews and must be approved by the Chief Statistician. The linkage will not be undertaken if the interests of a specific group of individuals might be harmed. If a sensitive issue is involved, the agency consults with representatives of the affected group. An example given on its website involves linkage of files involving social welfare recipients in order to look at the effectiveness of various social assistance programs. Before beginning the project, the agency consulted with anti-poverty organizations and the Canadian Privacy Commissioner. The analytic results of studies involving linked records are placed in the public domain and are accessible to the public. No linked record studies are confidential or secret. In addition, no linkages are maintained on an on-going basis. The linked data are destroyed at the conclusion of the project. In the case of ongoing projects, periodic reviews are conducted. The linked data bases are as small as possible, using samples, the databases are maintained on servers with no outside access, and all new projects are discussed with the Privacy Commissioner.

According to its privacy policy, Statistics Canada provides all respondents with, "information about: the purpose of the survey (including the expected uses and users of the statistics to be produced from the survey), the authority under which the survey is taken, the collection registration details, the mandatory or voluntary nature of the survey, confidentiality protection, the record linkage plans and the identity of the parties to any agreements to share the information provided by those respondents. The information

www.manaraa.com

required by this policy must, for all surveys, be prepared in written form and made available to respondents prior to or at the time of collection. In the case of telephone/interview surveys without introductory materials the information shall be provided verbally and shall be provided in writing on request."

The United Kingdom, the European Union, and Australia

Canada's privacy policies are similar in their strength to many of the European privacy principles and regulations. The United Kingdom (UK) and other countries of the European Union (EU) follow EC Directive 95/46 (which was introduced in the UK as the Data Protection Act of 1998). The directive contains a number of key principles which must be complied with. Anyone processing personal data must comply with the eight enforceable principles of good practice that state that data must be:

- Fairly and lawfully processed;
- Processed for limited purposes;
- Adequate, relevant and not excessive;
- Accurate;
- Not kept longer than necessary;
- Processed in accordance with the data subject's rights;
- Secure; and
- Not transferred to countries without adequate protection.

The UK's Data Protection Act established a Data Protection Commissioner, subsequently renamed as Information Commissioner in the Freedom of Information Act of 2000. Other countries of the EU have established posts similar to that of the UK's

www.manaraa.com

Information Commissioner. All the member states of the (EU are also signatories of the European Convention on Human Rights. Article 8 of the ECHR provides a right to respect for one's "private and family life, his home and his correspondence," subject to certain restrictions. EU and U.S. perspectives on data protection and privacy are different. The U.S. uses a sector approach to data protection legislation, relying on a combination of legislation, regulation, and self-regulation, rather than overarching governmental regulations.

The U.K. enacted the Statistics and Registration Service Act of 2007, which restructured its statistical system, creating the Statistics Board. This entity is independent of the Executive Branch of the government and reports directly to Parliament. The United Kingdom's statistical system has historically been decentralized. Although the Office for National Statistics (ONS) was the central producer of statistics, other government agencies also produced a large proportion of statistics. The ONS was an Executive Agency headed by the National Statistician who was concurrently the Registrar General for England and Wales. Consequently, the General Register Office (GRO), which administers the system for the registration of births, deaths, marriages and civil partnerships in England and Wales, was part of the ONS. The ONS was also responsible for the creation and maintenance of the National Health Service Central Register (NHSCR). Prior to enactment of the new act, the statistical system in the UK was governed by the non-statutory *Framework for National Statistics* (Statistics, 2000). The Framework included key structures and concepts, including a National Statistician with operational independence from Ministers; National Statistics to provide an accurate, up-to-date description of the economy and society of the UK, professional standards as

www.manaraa.com

set out in a Code of Practice; and the independent Statistics Commission, which advised on the quality and comprehensiveness of official statistics.

- The Act created a new body as the legal successor to the ONS called the Statistics Board, with a statutory responsibility to promote and safeguard the production and publication of official statistics that serve the public good and composed of a majority of non-executive members. The Board has powers to produce statistics, provide statistical services and promote statistical research, including the preparation and publication of the census.

Similar to the much more limited provisions in CIPSEA, *Section 47* of the Act allows the Minister for the Cabinet Office to create regulations that would allow information to be shared with the Board where this would normally not be allowed (either because of a barrier to sharing in existing law, or because such a public authority would not otherwise have the power to share information with the Board). Information shared under the regulations can only be used for statistical purposes, and cannot be disclosed by the Board other than in the limited circumstances set out in *section 39* and where the regulations provide for further disclosure. *Section 50* of the Act allows the Minister for the Cabinet Office to make regulations to allow the Board to use information it has received where such use would otherwise be prohibited. Under *section 51* of the Act, the Minister for the Cabinet Office may, with the consent of the Minister of the Crown responsible for the relevant public authority, make regulations to allow information to be shared by the Board with another public authority where this would normally not be allowed. Information shared under this provision can only be used for statistical purposes,

www.manaraa.com

and onward disclosure of the information is restricted under *section 39*. While the Act is still relatively new, it should lead to more data sharing between the now centralized statistical agency and other government agencies.

Australia mirrors the EU and Canada more than the U.S. when it comes to privacy policy and centralized statistics. In Australia, the federal Privacy Act of 1988 sets out principles in relation to the collection, use, disclosure, security and access to personal information. The Act applies to Australian Government and Australian Capital Territories agencies and private sector organizations (except some small businesses). The Office of the Privacy Commissioner handles complaints for alleged breaches of the Act. The Australian National Statement on Ethical Conduct in Research Involving Humans, issued by the National Health and Medical Research Council in 1999 includes the requirement to respect persons, including having regard for peoples' welfare, rights, beliefs, perceptions, customs and cultural heritage, for their autonomy and giving priority to respect for persons over the expected benefits to knowledge from the research. According to Thomson (2002), in Australia, the principles that are most relevant to the protection of privacy of research participants are: (1) respect for persons, especially the requirement for respect for individual consent to participation; (2) minimizing the risks of harm that include risk of intrusion of privacy; and (3) prior independent review that may identify and correct deficiencies in privacy protection.

The Australian Bureau of Statistics (ABS) is a centralized agency established by the Census and Statistics Act of 1905 as amended. The Census and Statistics Act prohibits the disclosure of identifiable information of a personal or domestic nature and requires

www.manaraa.com

that information can only be published in a manner that is not likely to enable the identification of a particular person or organization. The Act provides a fine of up to $5,000 and/or a penalty of 2 years imprisonment for an unauthorized disclosure of information collected under the Act. It also prohibits the disclosure of identifiable information of a personal or domestic nature under any circumstances to another government agency. Its underlying approach is to use social science statistics to measure the well being of the nation. Based on guidelines proposed by the Organization for Economic Co-operation and Development (OECD) that wellbeing could be effectively measured using key indicators, such as good health, sufficient income, rewarding work, etc., the ABS measures health, family and community, education and training, work, economic resources, housing, crime and justice, and culture and leisure.(ABS, 2008) Figure 1 below shows the key elements of the ABS approach. The ABS uses a variety of sources to gather data for these key indicators. For example, in the area of health, the ABS uses two mail types of data sources: administrative by product, and survey information. The censuses and surveys include:

- National Nutrition survey
- Mental Health and Wellbeing Survey of Adults
- Survey of Disability, Aging, and Careers
- Children's Immunization and Health Screening Survey
- Allied Health Industries Survey
- Private Medical Practice Industry Collection
- Census of Population and Housing
- National Drug Strategy Household Survey
- National Physical Activity Survey

www.manaraa.com

- The Child Dental Health Survey (Redesign of this collection is being undertaken to improve representativeness of estimates, provide linkage with social and service provision data, and allow longitudinal linkage of unit record files.)
- The Adult Dental Programs Survey
- Commonwealth Disability Services Census

In addition, the following national registries are used:

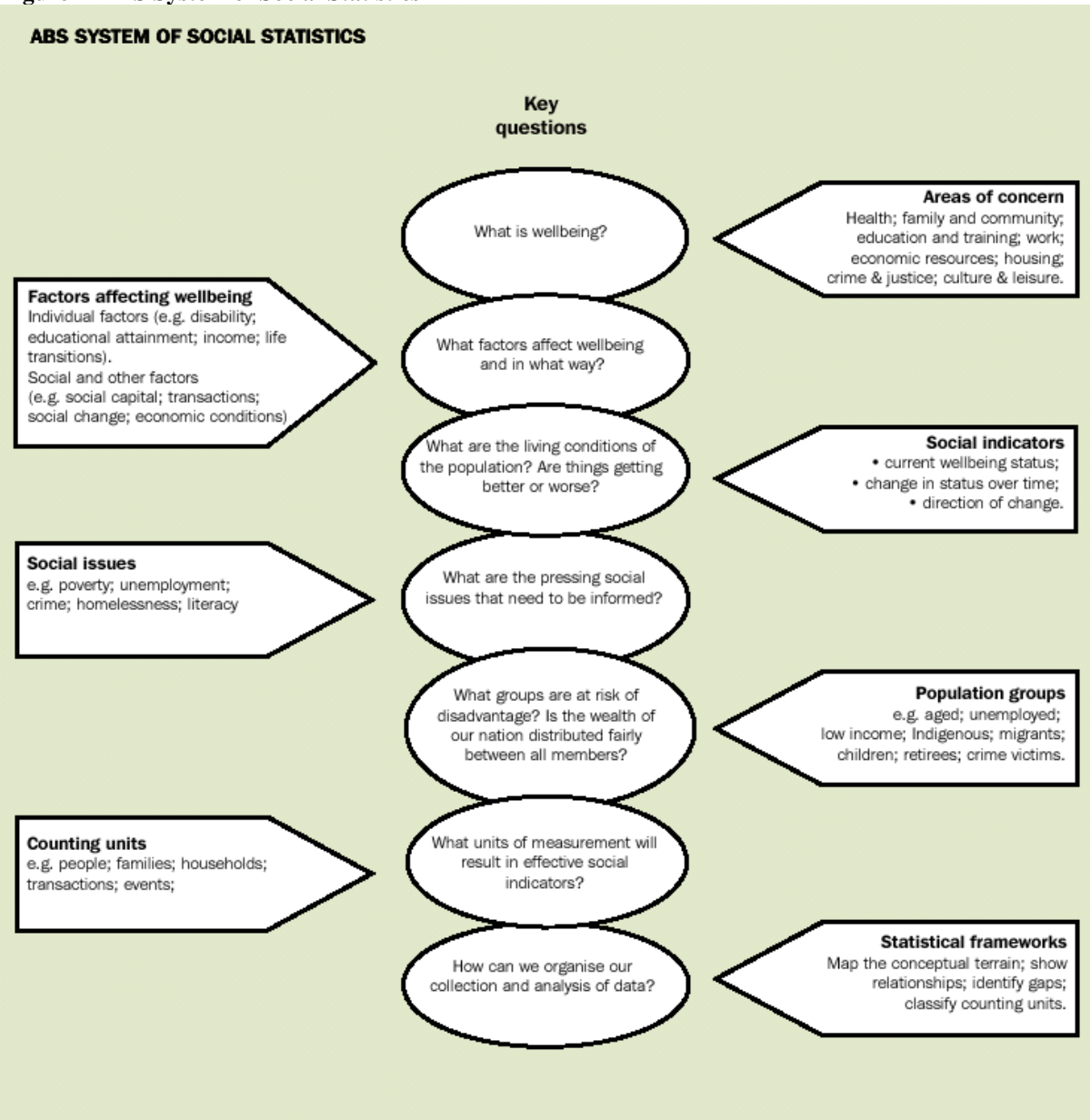
- Cause of Death Collection (registered by State and Territory registrars)
- Perinatal Deaths Collection (supplied by State and Territory Registrars to the ABS)
- Private Hospital Establishment Collection
- National Cancer Registry
- National Diabetes Register
- National Notifiable Diseases Surveillance System
- National Hospital Morbidity Database (compiled from data supplied by the State and Territory health authorities)
- Australian Childhood Immunisation Register
- Medicare Benefits Schedule Data Collection
- Pharmaceutical Benefits Schedule Data Collection
- National Midwives Collection (compiles perinatal data from midwives and other staff, obtained from mothers and from hospital or other records)

Data are also collected from these national studies:

- Australian Diabetes, Obesity and Lifestyle Study (AusDiab)
- Australian Longitudinal Study on Women's Health
- Australian Study of Health and Relationships

www.manaraa.com

Figure 2 ABS System of Social Statistics



Source: Australian Bureau of Statistics, accessible at www.abs.gov.au

Summary

The literature shows that there are significant differences between the Census Bureau, IRS, and CMS regarding their governing statutes and regulations for protecting data confidentiality. The differences are overlaid by the privacy laws governing all

www.manaraa.com

federal agencies, such as the Privacy Act. This contrasts with the Canadian, European and Australian experience, where the laws governing privacy are stronger than in the U.S. Chapter 3 will describe the methodology used to explore how these differences across agencies in the U.S. affect the data pools being studied.

www.manaraa.com

Chapter3: Methodology

This chapter explains the methodology that was used to study the sharing of government administrative records data in the U.S. It begins with a discussion of the research questions that were explored in the study. Next it addresses the data collection strategies. Third, it discusses how the data were analyzed. Fourth, the chapter covers the limitations of the methodology.

Research Questions

The study addressed three primary research questions related to the sharing of administrative records between U.S. Federal agencies, specifically IRS, the Census Bureau, and CMS. This first question was: what is the life cycle flow of administrative records data on individuals and businesses between IRS, CMS, and the Census Bureau? This question had three subparts:

a. What are the laws, rules and regulations guiding the sharing of these records?
b. To what uses are the data put, and how does that affect the handling of the records?
c. What are the business processes that guide the sharing and use of combined data including: 1) agency policies for internal handling; 2) training received by the people who handle the records; 3) compliance measurement; and 4) granting external access to combined data.

www.manaraa.com

Second, what are the significant issues that have arisen as a result of sharing administrative records related to the need to protect privacy and confidentiality? This question also has three subparts:

d. Where do the laws, rules, and regulations overlap or conflict?

e. Who "owns" the combined data?

f. What are the barriers to achieving the intended benefits of data sharing among agencies?

Third, what insights and potential solutions can be learned from the case studies that might be applied to help address the significant data-sharing issues that have been identified?

Definitions

For the purposes of this study, administrative records are defined as microdata records contained in files collected and maintained by administrative (i.e., program) agencies and commercial entities. Government and commercial entities maintain these files for the purposes of administering programs and providing services. Administrative records are distinct from systems of information collected exclusively for statistical purposes. The latter are defined as statistical records. However, when administrative and statistical data are combined, the new records are defined as combined administrative records. Other key definitions are included in Table 1.

Case Study Methodology

The case study methodology was chosen because the research questions seek to answer questions best investigated in a real-life environment. According to Yin (2003), a case study should be used when three conditions are met: (1) the research question asks

www.manaraa.com

how, why or sometimes what; (2) the investigator is not required to have control over the events being studied; and (3) the focus is on contemporary events. In addition, if the what questions are exploratory in nature, such as asking what can be learned from a study of a particularly effective system, that is a justifiable rationale for conducting an exploratory study in order to develop hypotheses for future inquiry. Yin's technical definition of a case study is… "an empirical inquiry that investigates a contemporary phenomenon within its real-life context, especially when the boundaries between phenomenon and context are not clearly evident…the case study inquiry copes with the technically distinctive situation in which there will be many more variables of interest than data points, and as one result, relies on multiple sources of evidence, with data needing to converge in a triangulating fashion…" (Yin, 2003 pp13-14).

Case studies have advantages for theory development including potential for achieving high conceptual validity and strong procedures for fostering new hypotheses; value as a useful means to closely examine the hypothesized role of causal mechanisms in the context of individual cases; and capacity for addressing causal complexity (George & Bennett, 2004). There are six different kinds of theory-building research objectives as identified by George and Bennett (2004), Liiphart (1971), and Eckstein (1975). These are:

1. Atheoretical or configurative idiographic, which provide descriptions that can be used in subsequent studies for theory building;
2. Disciplined configurative, which use established theories to explain a case, impugn established theories, or highlight the need for new theory in neglected areas;

www.manaraa.com

3. Heuristic, which inductively identify new variables, hypotheses, causal mechanisms, and causal paths;
4. Theory testing, which assess the validity and scope conditions of single or competing theories;
5. Plausibility probes, which are preliminary studies on untested theories and hypotheses to determine whether more intensive testing is warranted; and
6. Building block studies of particular types or subtypes of a phenomenon, which identify common patterns or may be parts of larger contingent generalizations and typological theories.

This case study had a research objective that was most closely aligned with the atheoretical or configurative idiographic type. The subclasses of phenomena investigated are the life cycle of new data pools that are formed when federal agencies share microdata, as well as the current implementation of laws, policies, and procedures that govern how these data are shared. These data pools were examined through five dimensions: legal, organizational, perceptual, technical, and human. The legal dimension included laws, regulations, and policies that affect the data pool life cycle. The perceptual dimension examined the attitudes and beliefs of both program participants and the public regarding protection of privacy and how that affects the other dimensions being looked at in this study. The organizational dimension described the processes being used to create, maintain, and safeguard new pools of administrative record data. The technical dimension examined the technological changes that have affected how data pools are formed, their utility, and the implications for privacy protection. Finally, the human dimension looked at how the actions of individuals have created situations in

www.manaraa.com

which significant changes have occurred regarding laws, process, and perception surrounding government data pools.

For this dissertation, two case studies were selected, because they offered contrasting strategies for combining administrative and statistical records data and protecting the confidentiality of the records and privacy of the respondents. Both of these case studies employed an embedded design (Yin, 2003) using more than one unit of analysis. The following criteria were used to select two case studies that addressed the goals of the research.

1. There were at least two agencies involved in sharing administrative records and creating a new data set.
2. The privacy and confidentiality laws, rules, and regulations governing the agencies involved overlapped or conflicted.
3. There were questions regarding "ownership" of these new combined data sets.
4. The combined data were used for secondary purposes, such as research or survey frames.
5. The agencies involved had business processes in place to guide internal handling of administrative records, training for employees and others handling combined data, compliance measurement, and granting external access to the combined data.
6. There were important controversies that arose as a result of combining administrative records that affected national policy as well as government and

www.manaraa.com

outside researchers, data quality, data collection costs, etc. (e.g., OMB had to get involved).

7. Controversies arose as a result of external factors such as new legislation or enhanced oversight from Congress or another governmental body.
8. Issues of informed consent played a role due to the secondary uses of the combined data pools.
9. The activities being examined were recent enough that the people involved could be located and interviewed.
10. The activities that make up each case study had an identifiable beginning and end that could be examined as a whole in retrospect.

Two case studies were selected that met the above criteria. The first involves the Internal Revenue Service (IRS) and the U.S. Census Bureau. The second primarily involves the Centers for Medicare and Medicaid Services (CMS) and the U.S. Census Bureau.

The IRS-Census case study was selected because it had a significant effect on record sharing activities that reverberated throughout the federal statistical system. The case study highlighted significant misunderstandings between agencies based on varying interpretations of laws and policies, and the resultant changes made by both agencies that had a ripple effect for their dealings with other federal agencies.

The CMS case study was selected because it highlighted important issues related to data quality. It also showed some successful practices that agencies and outside researchers have used to overcome some of the barriers to creating combined data sets among agencies.

www.manaraa.com

Both case studies demonstrated the breadth of policy issues that remain to be addressed and further research that is needed if the benefits of creating new integrated data sets are to be realized. While there are other instances of agencies combining data, these two case studies were chosen because they included all the significant issues.

The IRS-Census case study consists of the IRS audit of the Census Bureau's use of IRS records that took place between 1998-2001. The IRS had previously audited the Census Bureau in 1994, finding few violations of IRS laws, policies, and regulations. However, by 1998, the IRS had come under intense congressional scrutiny for both lax handling of tax records and for alleged taxpayer abuse. These external factors influenced the audit related activities.

A primary focus of the audit, the Census Bureau's business register, is constructed from tax records provided by IRS and protected under section 6103(j)(1)(A) of Title 26 of the United States Code (USC). It contains the names and addresses of all the businesses in the U.S. that file tax returns, and it is the basis for including respondents in the Census Bureau's economic censuses and surveys. The micro data collected in these censuses and surveys are made available to outside researchers through Census Research Data Centers (RDCs). During the audit, many research projects based at the RDCs were halted or delayed, and OMB had to intervene in negotiations between the two agencies multiple times. And while both agencies had practices and procedures in place for handling tax data, these were often in conflict. In addition, while the Census Bureau developed a Data Stewardship program during this period, it did not deal with informed consent issues related to administrative records

www.manaraa.com

usage for records that came from other agencies. All of these factors made this event ideal for a case study.

The second case study focuses on the period between enactment of the Health Insurance Portability and Accountability Act (HIPAA) in 1996, the issuance by HHS in April 2003 of the Privacy Rule required to implement the Act, and the Census Bureau's work through 2006 to try to develop an administrative records replacement for the Survey of Income and Program Participation (SIPP). The HIPAA Privacy Rule established a category of health information, referred to as protected health information (PHI), which may be used or disclosed to others only in certain circumstances or under certain conditions. The Privacy Rule placed new conditions on the use and disclosure of PHI by covered entities for research. The creation of a research database or repository, and the use or disclosure of PHI from a database or repository for research, may each be considered a research activity under the Privacy Rule. While the rule begins with an assumption that PHI will be treated confidentially, it covers a series of national priority purposes, including public health, law enforcement, national security, medical research, and so forth in which PHI could be used or disclosed without patient consent (45 C.F.R. §512). Implementation of the rule becomes complicated when data are combined and pooled, and the law is not clear in all areas.

Although HIPAA and the Privacy Rule do not apply directly to the Census Bureau, enactment of HIPAA and the subsequent activities to implement it did have an effect on sharing of Medicare and Medicaid records between CMS and the Census Bureau. For example, although records were shared under Exception 4 to the Privacy Act, which allows release of data to the Census Bureau under Title 13, one use to which

www.manaraa.com

the records were put was to measure the number of people covered by health insurance in the U.S. Medicare records were compared with the results of surveys such as the Current Population Survey (CPS) conducted by the Census Bureau to derive these estimates. In the post-HIPAA environment, the sharing of records between CMS and the Census Bureau became more complex.

In addition, secondary uses of micro data by non-Census employees may have involved entities that are covered by HIPAA. Similar to the Census Bureau RDCs, CMS established the Research Data Assistance Center (ResDAC) to provide free assistance to researchers interested in using identifiable Medicare and/or Medicaid micro data for their research. However, a series of overlapping and sometimes conflicting rules and laws governed the sharing of records between Census, CMS, and outside researchers, creating uncertainty and delays in research while the government sorted out the Privacy Rule.

Regarding informed consent from individuals whose records may be shared, CMS specifically mentioned that the information would be shared with the Census Bureau in its statement to Medicare beneficiaries required by the Privacy Act. But these data were combined with data collected by the Census Bureau under different informed consent language, creating still more issues surrounding data stewardship.

. Note that the case studies were not limited to discussing just the Census Bureau, IRS and CMS. That is, other Federal agencies, such as the Social Security Administration (SSA) and OMB were included because they played a role in the data sharing activities being described.

www.manaraa.com

The study first addressed research questions one and two. Then using these results, processes or approaches that might improve the U.S. system and at the same time, add to the body of knowledge that informs theory development on combined data sets were identified.

Although only two examples of administrative records sharing in the U.S. were included in this research, these cases were in many ways representative of data sharing among federal agencies for statistical research purposes. The cases capture situations in which one of the agencies, the Census Bureau, is a statistical agency recognized by OMB. The second agency, the IRS, while not primarily a statistical agency, has within it a statistical organization, the Statistics of Income Division. IRS is highly visible to the public and has been in news reports and has testified before Congress regarding how it protects the confidentiality of the data it collects from the public. The third agency, the CMS, directly serves a large segment of the public, has a large constituency of outside researchers interested in access to its administrative records, and shares large amounts of sensitive, confidential data on individuals with other federal agencies. These agencies represent three distinct profiles of agencies that might share data records under a variety of legal authorities. Issues that surfaced in these case studies could well apply to other agencies that operate within the federal statistical system.

Data Collection

The study's data collection consisted of three phases. These phases were: (1) gathering information about the two systems from laws, regulations, policies, other documentation, and historical data; (2) gathering information about each agency's practices through interviewing knowledgeable current and former employees of each

www.manaraa.com

agency; and (3) re-interviewing some of the individuals interviewed during the second phase to gain insight and reactions to possible ways of improving the system and addressing issues that have been identified through analysis of data gathered during phase one and two.

Gathering of information about the total system consisted of (1) researching public documentation, (2) continuing the literature review begun during the proposal writing stage, and (3) gaining information through the interviews with practitioners. The substance of the interviews was limited to discussions of the participants' work and official duties.

The primary method for gathering information about each agency's practices was through interviewing practitioners in each of the agencies, and reading relevant papers and other documentation that they provided. Many of the practitioners interviewed had written extensively in presented and published papers about their agency's work in combining administrative records. A total of 15 study participants were interviewed. They were selected based on the level of involvement they had in the case studies being researched. All the participants had first hand knowledge of at least one of the case studies through working at their agencies during the periods being examined. Some had knowledge of both case studies. The level of involvement of the interview participants ranged from policy level to researchers, technical personnel handling the records, and process coordinators.

After analyzing the data collected in the initial rounds, preliminary conclusions and recommendations for both future study and possible government actions were synthesized. A key part of the data collection was to go back to a subset of the

www.manaraa.com

individuals initially interviewed and discuss the preliminary findings and recommendations with these individuals. This was an important check on the quality and the feasibility of the recommendations.

The interviews were semi-structured and were guided by a set of questions that are included in Appendix V. They had the following characteristics as identified by Mason (2002): (1) interactional exchange of dialogue; (2) a relatively informal style such as "conversations with a purpose" (Burgess, 1984); (3) a thematic, topic-centered approach; and (4) the responses from participants put into contextual focus in order to better capture perspective, meanings, and understandings of the participants' knowledge. There was not a "one size fits all" approach to the interviewing, in order to give maximum opportunity for the construction of contextual knowledge and to allow each interview to focus on relevant specifics for each respondent. That is, while a standard set of questions was used as a guideline for the interviews, the actual interviews deviated from the standard questions, because the perspectives of each of the participants varied based on their jobs and their agencies.

Gaining access to the research participants was accomplished in a variety of ways. Fifteen interviews were conducted in total, consisting of one at the IRS, one at CMS, one at SSA, two at OMB, four with current Census Bureau employees, and six with former Census Bureau employees. Participants were employees who either worked directly with administrative records, were involved with safeguarding the records, or who were responsible for developing administrative records policy, including one presidentially appointed position. Participants were identified both through this researcher's knowledge of the field and through referrals by participants.

www.manaraa.com

The second round of interviews was conducted with a smaller subset of participants, consisting of four people from IRS, CMS, and Census. The second round participants were identified based on their knowledge and contributions the first round of interviewing. Permission to return for a second round was requested during the first interview, and all participants agreed. During the second round of interviewing, the findings discussed in Chapter 4 were validated, and the descriptions in the case studies were fact checked and validated.

Because of the small number of participants, no pilot test of the interviewing was conducted. The interview questions evolved as the study progressed. Detailed notes were taken during the interviews. The confidentiality of each respondent is being preserved, although due to the small number of participants, each was informed of the possibility that their input may be recognized as part of the informed consent process. The interviews each lasted between one and three hours.

Data Analysis

Data analysis began with a review of the documentation collected and the first round of interview responses. Qualitative data from the open-ended questions asked at or about the U.S. agencies were analyzed using an inductive process to identify, code, and categorize the primary patterns in the data (Patton, 1990). From this, logic models were constructed to diagram the work processes at each agency (McLaughlin & Jordan, 2004). The work process diagrams are included in Chapter 4. These results were folded up into a flow chart and description of the interactions of the whole system. A cross case synthesis was then performed, and the findings aggregated across the findings of both

www.manaraa.com

case studies (Yin, 2003). Both cases were examined to determine whether issues that arose in one system have been addressed in the other system.

The results of the preliminary analyses were used to identify on-going issues and recommendations. These issues and recommendations were discussed with the four participants who were re-interviewed, and a final set of findings and recommendations was constructed using this additional input.

Limitations

As with all research, there are limitations to this study's methodology. Because the study included only two case studies, there are some limits on generalizing the findings to the broader federal statistical system in the U.S. However, the purpose of this research is exploratory, and the findings and recommendations were reviewed and assessed by the participants, therefore, this limitation is not as serious as in a situation involving hypothesis testing. Measurement validity threats arise when the procedures used in the study threaten the researcher's ability to draw accurate data responses from the interview participants. This may occur due to the biases of both the interviewer and the participants. This cannot be eliminated, but by interviewing multiple participants and ensuring anonymity of the responses, this effect may be alleviated. In addition, participants may have had faulty memories of events or procedures. Assuring the validity or accuracy of the information may be characterized as its trustworthiness, authenticity, and credibility (Creswell & Miller, 2000). Creswell recommends eight primary strategies to check the accuracy of the data. These strategies were utilized in this research. They are:

www.manaraa.com

- *Triangulating different data sources:* this was accomplished by examining data from different sources, asking multiple participants the same questions and by pulling data from existing documentation;
- *Using member checking*: this occurred when the findings were taken back to the participants to ascertain whether they thought the findings were accurate;
- *Using rich descriptions to convey the findings*: they are included in the narrative of the report;
- *Clarifying the researcher's bias:* the report findings and recommendations make the biases clear;
- *Presenting negative or discrepant information:* this type of information was included in the report;
- *Spending prolonged time in the field*: which was accomplished by this researcher working for several years at one of the agencies being studied in a position with influence over the topic being studied (also a source of bias);
- *Using peer debriefing to enhance the accuracy of the account*, which was accomplished in this study by discussing the findings and recommendations with knowledgeable colleagues who were not participants; and
- *Using an external auditor to review the entire project*: Outside review is being accomplished through the dissertation committee review.

Another limitation of the study is that the recommendations are likely to be hard to implement. By using member checking and peer review, the chance of developing recommendations that are not feasible were minimized. The report discusses what

www.manaraa.com

the best courses of action in the future might be, even though they may not be easily accomplished in the short term.

Summary

The dissertation addressed three important questions regarding the safeguarding of confidential information when the government shares administrative records for statistical and research purposes. The research contributed to the literature on privacy policy and administrative records sharing, and helped develop theory regarding the life cycle of data pools created by government agencies. By conducting document reviews and face-to-face interviews, the dissertation research provided important findings on the data sharing policies, procedures, and implementation mechanisms of the U.S. government.

In addition to contributing to the development of theory on data pools, the research has practical applications. The recommendations that arose from the findings and conclusions may be used by practitioners and policy makers to make improvements to the U.S. federal statistical system and address identified weaknesses regarding data stewardship in the context of creating integrated data sets of administrative records and other data.

The pressures that are driving the use of administrative records and increased record sharing among agencies are not likely to abate in the foreseeable future. Respondent cooperation may continue to decline as people continue to lead busy, stressful lives and use technologies that put them out of easy reach of data collectors, such as cell phones. Further, as certain government agencies continue to collect more data on individuals in order to identify terrorists and other national security threats, the

www.manaraa.com

public has become much more aware of the ways in which private information is passed among various agencies and the vast data pools of personal information subsequently being created. Government actions could affect the public's willingness to cooperate and to consent to uses of its data. The cost of data collection will continue to increase, even as federal budget pressures increase. In addition, public policy problems are increasingly complex, and with improved technology, researchers are clamoring for more access to microdata and designing studies that take advantage of the ability to combine and analyze large data sets. Understanding data pools and finding ways to document and possibly improve the U.S. federal statistical system for record sharing among agencies should contribute to assuring that the system can meet the challenges of continuing critical public policy research in a changing environment.

www.manaraa.com

Chapter 4: Presentation and Analyses of the Case Studies

Introduction

The purpose of this study was to explore the life cycle of data pools that are created when administrative records are shared between federal agencies, with a focus on three federal agencies, the U.S. Census Bureau (Census), the Internal Revenue Service (IRS), and the Centers for Medicare and Medicaid Services (CMS). For the purposes of this research, the lens for examining record sharing was privacy and confidentiality. The research used two case studies to examine the public policy aspects of this question through five dimensions: legal, organizational, perceptual, technical, and human. This chapter presents and analyzes the findings of the two case studies. The discussion addresses three questions. The first question is what is the life cycle flow of administrative records data on individuals and businesses between IRS, CMS, and the Census Bureau? The second question is what are the significant issues that have arisen as a result of sharing administrative records related to the need to protect privacy and confidentiality? These two questions are addressed through understandings gained by thinking of administrative records data as imperfect public goods and through empirical knowledge.

The third question addresses the normative aspects of data sharing. What insights and potential solutions can be learned from the experiences of those who have worked within the federal statistical system that would help address the significant data-sharing issues that have been identified?

Based on the findings, Chapter 4 includes recommendations on how the record sharing processes that are in place to safeguard data and protect confidentiality and the

www.manaraa.com

privacy of data providers could be improved. The recommendations are considered in a contextual framework, particularly how context may shape the behavior of agencies and individuals.

Chapter 4 is organized into three sections. The first section provides the context for the cases studies by giving an overview of their structure and how they collect and utilize data. The second section presents the case study data in the context of the five dimensions as derived from interviews and document review. The two case studies are presented separately but are then compared for commonalities and differences. The third section summarizes the findings of the analyses of the case studies.

It is important to note that much of the information presented has been gathered through personal interviews of a relatively small group of study participants from their point of view. Therefore, this information is not specifically cited throughout the paper.

The Case Studies Contexts

The Internal Revenue Service

The Internal Revenue Service (IRS) is a bureau of the Department of the Treasury under the immediate direction of the Commissioner of Internal Revenue. The Commissioner has authority over the assessment and collection of all taxes imposed by any law providing internal revenue and accomplishes this through IRS (26 C.F.R. section 601.101(a)). The Office of the Commissioner of the Revenue was created by Congress and President Lincoln in 1862 to support the war effort by collecting income tax revenue. There have been 47 Commissioners since that time. IRS collects approximately $2.4 trillion in tax revenue annually and has over 100,000 employees nationwide. The current stated IRS mission is, “to provide America's taxpayers top quality service by helping

www.manaraa.com

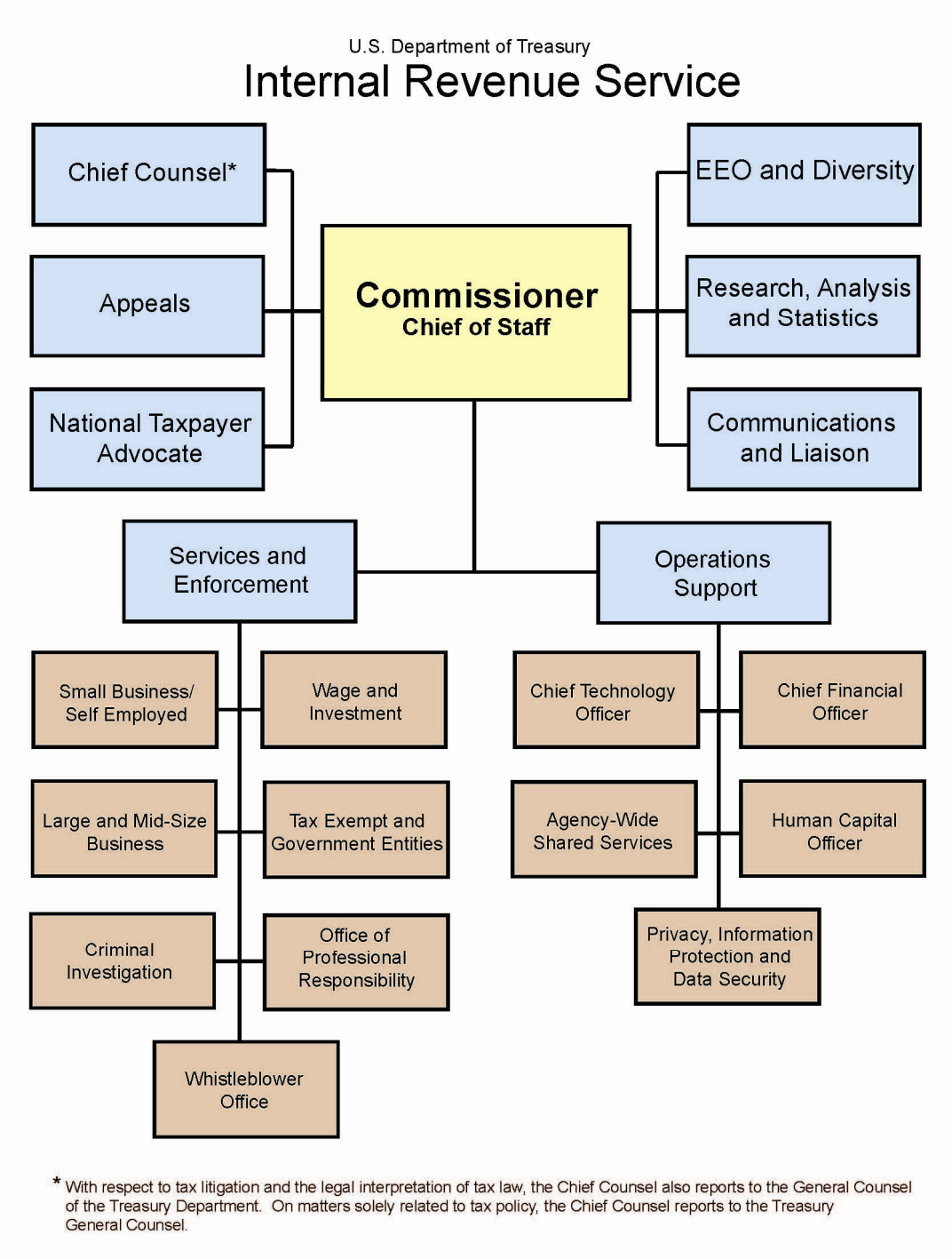
them understand and meet their tax responsibilities and by applying the tax law with integrity and fairness to all." (IRS, 2009)

The modern structure of IRS was created by the Internal Revenue Restructuring and Reform Act of 1998 (Pub.L. 105-206, 112 Stat. 685). IRS is organized to carry out the responsibilities of the Secretary of the Treasury under section 7801 of the Internal Revenue Code. IRS currently has four main divisions, Large & Mid-Size Business (LMSB), Small Business / Self-Employed (SB/SE), Wage and Investment (W&I), and Tax Exempt & Government Entities (TE/GE). In addition, there are several other offices to support these operating divisions, such as the Office of Legal Counsel. Of most interest to this study is the Statistics of Income (SOI) Division, located within the Office of Research, Analysis, and Statistics in the Office of the Commissioner. An organizational chart for the IRS is shown in Figure 3 and is also available at www.irs.gov/pub/newsroom/irs_org_chart_1-09.pdf .

SOI has four branches: (1) individuals and sole proprietorships, (2) corporations and partnerships, (3) special studies (including international, tax exempts, and estates), and (4) statistical computing, which provides support to the other three branches. The information SOI gathers, analyzes, and publishes is used by a variety of federal agencies, academics, and researchers to analyze tax policy, project tax revenues, and estimate the overall impact of tax law changes and their effects on tax collections. The source of data is tax forms filed by individuals and businesses. A primary client of SOI is the Office of Tax Analysis (OTA) within the Office of Tax Policy (OTP) in the Secretary of the Treasury's Office.

Figure 3 IRS Organization Chart

www.manaraa.com



OTP is headed by the Assistant Secretary of the Treasury for Tax Policy. OTP assists the Secretary in developing and implementing tax policies and programs; provides the official estimates of all Government receipts for the President's budget, fiscal policy

www.manaraa.com

decisions, and Treasury cash management decisions; establishes policy criteria reflected in regulations and rulings and guides preparation of them with IRS to implement and administer the Internal Revenue Code; negotiates tax treaties for the United States and represents the United States in meetings and work of multilateral organizations dealing with tax policy matters; and provides economic and legal policy analysis for domestic and international tax policy decisions. Within that office is OTA, which advises and assists OTP by assessing, from an economic and policy perspective, all major tax initiatives, including Administration and congressional tax proposals, and studies the effects of the existing tax law and alternative tax programs. OTA develops and operates several major micro simulation models and maintains large statistical databases to analyze the economic, distributional, and revenue effects of alternative tax proposals and tax systems. Many of the large microdata files used in OTA's models are developed from samples of tax returns prepared by SOI. Both SOI and OTA maintain a close relationship with the congressional Joint Committee on Taxation. The special relationship enjoyed by these three entities rests in no small measure on the access they share to tax-related microdata from filings. This access is closely guarded, although the Census Bureau, by statute, also has access for limited purposes.

The Centers for Medicare and Medicaid Services

The Centers for Medicare and Medicaid Services (CMS) is part of the Department of Health and Human Services. Its mission is "to ensure effective, up-to-date health care coverage and to promote quality care for beneficiaries" of Medicare and Medicaid (CMS, 2009). These two programs were established in the Social Security Act of 1965, in Title XVIII and Title XIX. Until 1977, Medicare was managed by the Social Security

www.manaraa.com

Administration (SSA) and Medicaid was managed by the Social and Rehabilitative Services Administration (SRSA), at which time the Health Care Financing Administration (HCFA) was established to manage and operate both programs. In 2001, HCFA was renamed CMS.

Medicare extends health coverage to almost all Americans aged 65 and older, as well as people with long term disabilities and end stage renal disease. Medicaid provides health care services to low-income children, their caretaker relatives, elderly, blind and disabled people, and pregnant women and their children up to 6 years old if their incomes do not exceed 133% of the poverty level. The Balanced Budget Act of 1997 established the State Children's Health Insurance Program (SCHIP), which provides health insurance to uninsured, low-income children 18 years of age or younger, including those who are homeless. The Census Bureau receives a $20 million mandatory annual appropriation to produce statistically reliable annual data for each state on the number of low-income children who do not have health insurance coverage. Census-provided data are used to allocate funds to states based on the number of children without health insurance living in low-income families. Allocations are based on statistics from the Annual Social and Economic Supplement to the Current Population Survey (CPS), which is conducted by the Census Bureau.

CMS makes data available to several entities for research purposes including:

- Academic Institutions/Private Sector
- Congressional Entities
- HHS Federal Agencies/Contractors
- Non-HHS Federal Agencies

www.manaraa.com

- Health Care Providers
- State Government Agencies
- State Medicaid Agencies

CMS has a number of rules and policies in place to govern approval of research projects. It has a contractor-operated Research Data Assistance Center (ResDAC) through which many requests are reviewed and which also offers assistance to researchers. Data must be able to be disclosed under the Privacy Act of 1974 and published as a System of Records. The Privacy Act of 1974 and the System of Records are CMS's legal authorization to release data, and these legal requirements protect the confidentiality of individually identifiable data. Also required is a strong research design with a scope and subject matter that assist CMS in monitoring, managing, and improving the Medicare and Medicaid programs or services provided to beneficiaries. Researchers must sign a Data Use Agreement (DUA) that requires the researcher to get CMS permission before linking to other data files and specifies the process for destruction or return of data to CMS at the conclusion of the project.

All research conducted, including products or tools that are developed, must be shared with the public but have to be reviewed by CMS prior to publication. The purpose is to ensure that those products meet CMS standards for privacy, as they relate to small cells in tables or small person counts elsewhere in reports.. According to the interviewees, reviews had been conducted by the Division of Privacy Compliance. However, that Division is no longer conducting their own reviews, because the volume of the work was too much for the staff. The Division is now requiring approved users to

www.manaraa.com

certify that they are meeting CMS standards in the release or publication of research products.

CMS also shares administrative record data with states to help them administer the programs. For example, Medicare's Enrollment Database (EDB) contains information that is updated daily on all individuals entitled to Medicare, including demographic information, enrollment dates, third party buy-in information, and Medicare managed care enrollment. States can send CMS a file of Medicaid beneficiary Social Security Numbers (SSNs) that will be compared with information in the EDB to determine which Medicaid beneficiaries are also eligible for Medicare.

Another data set, the Long Term Care Minimum Data Set (LTCMDS) is the core set of screening and assessment elements of the Resident Assessment Instrument (RAI), which provides an assessment of each long-term care facility resident's functional capabilities, and helps staff to identify health problems. This health status assessment is performed on all residents in a Medicare and/or Medicaid-certified long-term care facility, and is made available to state agencies upon request if the agencies meet the privacy protection requirements.

Also available to state agencies is the Outcome and Assessment Information Set (OASIS), which is a group of data elements that represent core items of a comprehensive assessment for an adult home care patient and form the basis for measuring patient outcomes for purposes of outcome-based quality improvement (OBQI). The comprehensive assessment is performed on every patient receiving services of home

www.manaraa.com

health agencies that are approved to participate in the Medicare and/or Medicaid programs.

The Medicaid Analytic eXtract (MAX) data consist of person-level data files on Medicaid eligibility, service utilization, and payments. The MAX data, organized onto annual calendar year files, are created specifically to support research and policy analysis. MAX data are extracted from the Medicaid Statistical Information System (MSIS). The MAX development process combines MSIS initial claims, interim claims, voids, and adjustments for a given service into this final action event. Because MAX data contain individually identifiable data, they are protected under the Privacy Act. Only projects approved by the CMS Privacy Board and certain other users, such as the Census Bureau, Department of Justice, and Congressional Budget Office are entitled to obtain MAX data.

The Census Bureau works through the CMS Office of Research, Development and Information (ORDI) to gain access to Medicare records. The Director of ORDI reports to the CMS Administrator. Within ORDI is the Research and Evaluation Group, which houses the Division of Research on State Programs & Special Populations. The Medicaid MAX system manager, who is in that office, along with the Information Methods Group in the Division of Survey Management & Data Release, which houses the CMS Privacy Board, works closely with the Census Bureau on its requests to obtain administrative records. The mission of ORDI is to provide analytic support and information to the CMS Administrator and Executive Council, perform environmental scanning of emerging trends in health care delivery and financing, design and conduct research and evaluations of health care programs, coordinate CMS demonstration

www.manaraa.com

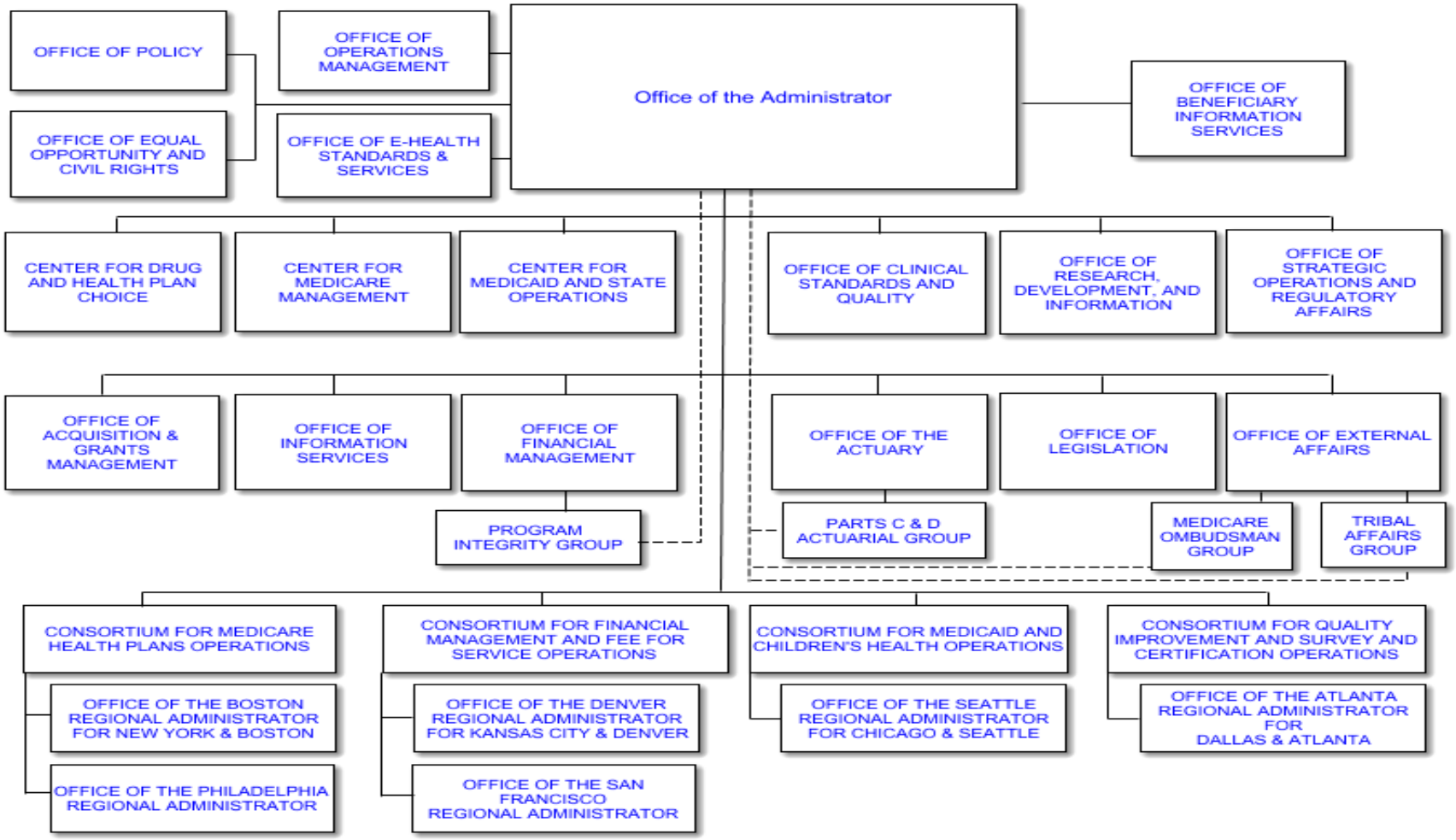
activities, manage assigned demonstrations, including federal review, approval, and oversight; and develop research, demonstration, and other publications and papers related to health care issues. Also playing a role in review of research project proposals is the Office of Information Services, which houses the Division of Privacy Compliance within the Enterprise Architecture and Strategy Group. Among other things, this office provides Medicare and Medicaid information to the public within the parameters imposed by the Privacy Act, performs information collection analyses to satisfy the requirements of the Paperwork Reduction Act, directs CMS' IT security program, and directs and advises the Administrator, senior staff, and agency components on the requirements, policies, and administration of the Privacy Act. A CMS org chart is in Figure 4 and can be seen online at: http://www.cms.hhs.gov/CMSLeadership/50_OrganizationalChartASP.asp#TopOfPage

The Census Bureau

The Census Bureau was originally established as a permanent agency in 1902, for the purpose of conducting the decennial census of population and housing, which had been conducted every ten years since 1790, as mandated by the Constitution. It is currently part of the Department of Commerce, where it was moved in 1903. The Census Bureau employs about 5,000 people at its headquarters in Suitland, Maryland and another 6,000 permanent employees in 11 regional offices and three telephone call centers around the U.S. Its temporary staff swells to about 900,000 people during the year of the decennial census. The temporary employees are assigned to about 530 temporary local offices throughout the country. The organization chart for the Census Bureau can be viewed in Figure 5 or at this link: http://www.census.gov/aboutus/orgchart.png .

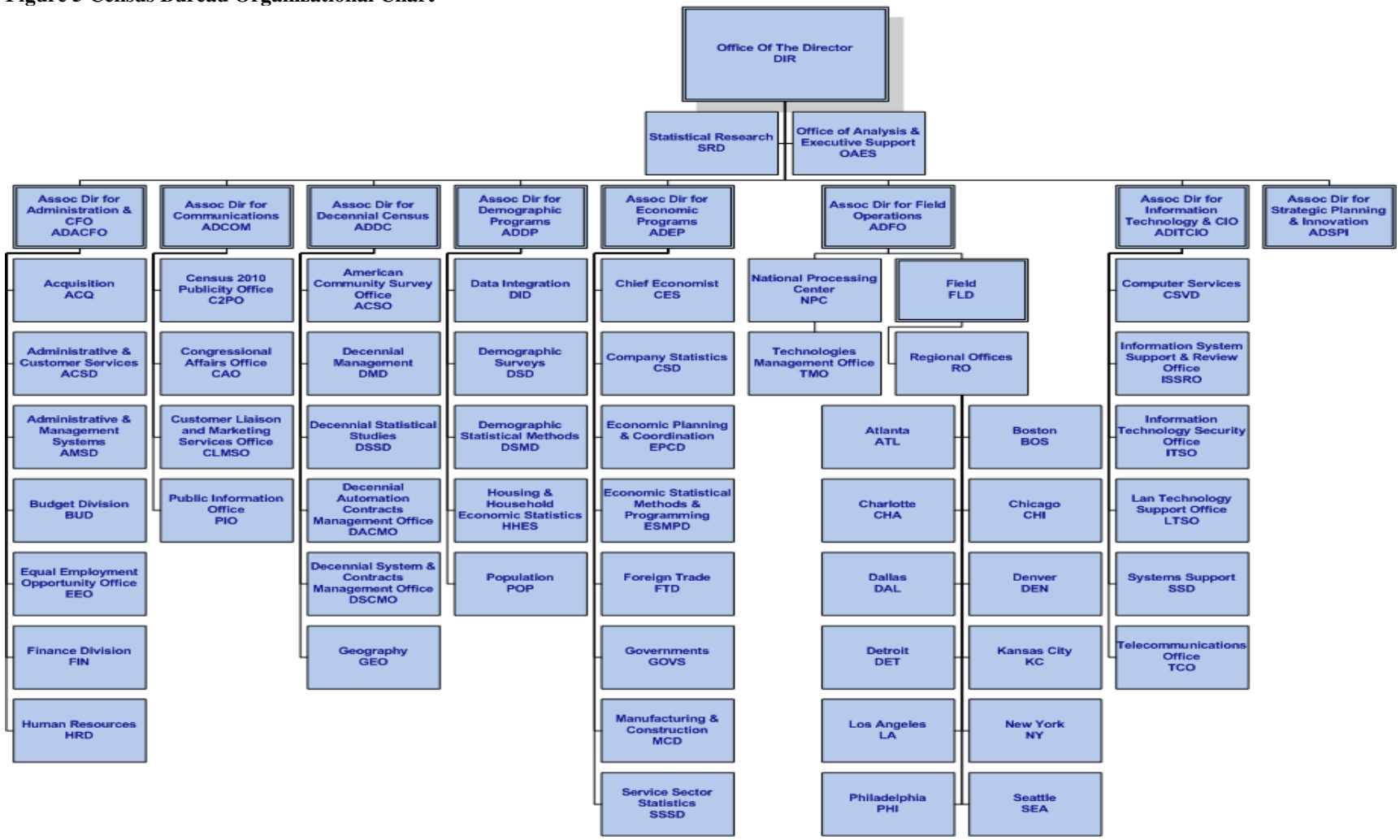
www.manaraa.com

Figure 4 CMS Organizational chart



www.manaraa.com

Figure 5 Census Bureau Organizational Chart



www.manaraa.com

Although conducting the decennial census is the activity for which the Census Bureau is best known, the bureau conducts several other surveys and censuses as well as analytical work and research and development into statistical and survey methodology. In 1954, legislation was enacted that combined existing laws governing Census Bureau programs and codified them in Title 13 of the United States Code. The modern day Census Bureau is the premier, although not the only, statistical agency of the U.S. government. It conducts more than 200 annual demographic surveys on behalf of other agencies. Appendix 1 summarizes the major surveys conducted on a cost reimbursable basis. The Census Bureau receives directly appropriated funds to conduct the American Community Survey (ACS), which replaced the long form of the census. The ACS is very large – during its first four years it included 800,000 households in 1,203 counties. The Census Bureau also receives direct appropriations to conduct the Survey of Income and Program Participation (SIPP), which collects detailed information on cash and non-cash income, taxes, assets, liabilities, and participation in government transfer programs for individuals. These surveys are conducted by the Demographic Programs Directorate of the Bureau.

The Census Bureau began its current administrative records research program after the 1990 decennial census. At the time, there was a lot of interest in whether administrative records could be used in the design of the 2000 census. The Census Bureau sponsored an information gathering conference in 1993 and began conducting tests, such as the Administrative Records Test for the 2000 Census (Obenski, 2006). The testing on the feasibility of an administrative records-based decennial census began in

www.manaraa.com

1995 with the development of a prototype of the Statistical Administrative Record System (StARS). In 1997, an Administrative Records Research staff was established in the Standards and Methodology Directorate to formally establish a point of responsibility for this work. StARS was created by merging IRS and Medicare files and then using the SSA's SSN Transaction File, known as the NUMIDENT to validate the SSNs and ascribe personal characteristics, such as race and gender, to the merged files. In 1997, the Census Bureau reached an agreement with SSA for it to provide its NUMIDENT file on a regular basis, which includes every transaction made for an SSN. StARS, containing the merged files with validated SSNs and personal information contained similar information to that collected during the decennial census on what was then known as the census short form.

As part of the testing program for Census 2000, two different models for using merged administrative records were developed and compared with the census results. Some of the quality shortfalls that were discovered as a result of the tests were that children were missing from the administrative records, and there was an overcount of the elderly (Obenski, 2006) . StARS is an ongoing program at the Census Bureau, with the goal of improving data quality.

Census Bureau Sharing of Records with SSA

SSA and Census have been sharing information since 1939, when the two agencies began sharing industry codes. They started linking records during the 1960s. The linkages have been beneficial to both agencies.

www.manaraa.com

During the 1990s, SSA wanted to begin research on how Social Security benefits affected the well being of special populations, such as the disabled. In order to carry out that analysis, SSA needed information on individuals from Census Bureau surveys, SIPP and CPS, which are protected by title 13. SSA could not gain direct access to the survey microdata. However, SSA could get the data more indirectly by allowing certain record linkages at the Census Bureau. SSA agreed to start send the NUMIDENT file to Census in order to create linkages to survey data by SSN. The Census Bureau had addresses from IRS and used them to create SSNs for survey respondents (who had not been asked to provide SSNs for the surveys). This comingling of IRS data and the NUMIDENT enabled the linkage of the survey data with SSNs. Census returned the linked files to SSA, which then linked the quality checked SSNs with administrative data such as lifetime earnings and Social Security payments of the survey respondents. SSA also received information on marital status, number of children, other sources of income, assets, and race from the surveys. The information was used strictly for statistical purposes such as developing models to project the effect of changes in the law. This arrangement created benefit for both SSA and Census.

Internal Uses of Linked Data at Census

In addition to sharing records with other agencies, the Census Bureau uses linked records in many internal programs. Within the Demographic Programs Directorate, the Population Division and the Population Estimates Branch annually produce estimates of the population for states and counties. Estimates of the population for the 36,000 general purpose units of local government are produced biennially. The estimates are produced

www.manaraa.com

using administrative records that track births, deaths, international migration and internal migration. For example, the Census Bureau geographically codes current year individual income tax returns and matches them to the prior years' geographically-coded file to track people who have changed addresses. Estimates are compiled for people who have moved into a designated area or moved out of that area. The individual income tax returns provide street address, post office name, 9-digit ZIP Code and mailing state abbreviation. The Population Division uses SSA records to track deaths, Customs and Immigration Service (CIS) records to track numbers of people entering the country, and hospital records to track births.

The Census Bureau also houses an Economic Programs Directorate that conducts censuses and surveys of U.S. businesses and relies heavily on administrative records. The information collected by the Census Bureau is used by the Bureau of Economic Analysis (BEA), also in the Commerce Department, to calculate economic indicators such as the Gross Domestic Product (GDP) and ongoing indices such as new housing starts, retail sales, and others. The Census Bureau began collecting economic data in 1930, and currently conducts the Census of Governments and the Economic Census every five years. Ongoing surveys conducted by the Census Bureau include the Survey of Business Owners, the Commodity Flow Survey, the Business Expenditures Survey, the Monthly Building Permits Survey, the Monthly Retail Sales Survey, and the Monthly Wholesale Trade Survey, which are described in Appendix 3.

The Economic Directorate relies heavily on IRS tax files of businesses to construct its Business Register, which it has been updating continuously since its

www.manaraa.com

inception in 1972. The Business Register is a listing of all domestic businesses, covering more than 160,000 multi-establishment companies, representing 1.8 million affiliated establishments, 5 million single establishment companies, and nearly 21 million non-employer businesses (Census, 2009a) . Business Register information includes business location, organization type (e.g., subsidiary or parent), industry classification, and operating data (e.g., receipts and employment). The Business Register consolidates and links administrative, Census, and survey data. Records from multiple sources are used, based on Internal Revenue Service Employer Identification Numbers (EINs). The system includes unique Census-assigned identification numbers, EINs, and industry classifications assigned by the Social Security Administration. Information for single establishments and EINs is updated continuously; including employment and payroll data based on payroll tax records, and receipts data based on income tax records from the IRS. Information for establishments of multi-unit companies is updated annually based on responses to the company organization survey and annual survey of manufactures conducted by the Census Bureau. Other sources of update information include other Census Bureau current surveys and the Economic Census.

Housed within the Economic Programs Directorate is the Center for Economic Studies (CES). This center manages the Census Bureau's remote Research Data Centers (RDCs), which were established in 1994 to provide access to confidential Title 13 and administrative data to researchers. There are nine RDCs at the following locations:

- Boston, MA – National Bureau of Economic Research
- Berkeley, CA – University of California, Berkeley
- Los Angeles, CA – University of California, Los Angeles

www.manaraa.com

- Suitland, MD – Census Bureau Center for Economic Studies
- Chicago, IL – Federal Reserve Bank of Chicago
- Ann Arbor, MI – University of Michigan
- New York, NY – Baruch School of Public Affairs
- Ithaca, NY – Cornell University
- Durham, NC – Duke University

From a legal perspective, researchers accessing data at an RDC are sworn to uphold the confidentiality provisions of Title 13. Wrongful disclosure of confidential Title 13 data is punishable by fine not exceeding $250,000.00 and/or imprisonment of no more than five years (13 U.S.C. Section 214; 18 U.S.C Section 3571). Taking the oath confers upon researchers Special Sworn Status under 13 U.S.C. § 23(c), which authorizes the Census Bureau to have temporary staff to work on projects that will benefit the bureau. Organizationally, researchers at RDCs additionally must have approved projects that benefit the Census Bureau. RDCs must have a permanent Census Bureau employee on site but are operated with academic and non-profit partner institutions. Because some of the records include tax data provided by IRS, RDCs must comply with not only Census Bureau information and physical security requirements, but also IRS requirements.

According to the guidelines issued by CES (Census, 1997), research projects are reviewed and approved based on five major review standards:

1. A benefit to Census Bureau programs conducted under Title 13;
2. Scientific merit in that the research will contribute to existing knowledge;
3. A clear need for non-public data;
4. Feasibility of success; and

www.manaraa.com

5. Acceptance of all confidentiality protection and disclosure avoidance review requirements.

The benefits to programs conducted under title 13 that researchers must demonstrate have been defined as follows:

1. Evaluating concepts and practices underlying Census Bureau statistical data collection and dissemination practices, including consideration of continued relevance and appropriateness of past Census Bureau procedures to changing economic and social circumstances;
2. Analyzing demographic and social or economic processes that affect Census Bureau programs, especially those that evaluate or hold promise of improving the quality of products issued by the Census Bureau;
3. Developing means of increasing the utility of Census Bureau data for analyzing public programs, public policy, and/or demographic, economic, or social conditions;
4. Conducting or facilitating Census Bureau census and survey data collection, processing or dissemination, including through activities such as administrative support, information technology support, program oversight, or auditing under appropriate legal authority;
5. Understanding and / or improving the quality of data produced through a Title 13, Chapter 5 survey, census or estimate;
6. Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census or estimate;
7. Enhancing the data collected in a Title 13, Chapter 5 survey or census; for

www.manaraa.com

example:

a. Improving imputations for non-response;

b. Developing links across time or entities for data gathered in censuses and surveys authorized by Title 13, Chapter 5.

8. Identifying the limitations of, or improving, the underlying business register, household Master Address File, and industrial and geographical classification schemes used to collect the data.

Researchers at RDCs also may be accessing data provided by agencies other than IRS. In those instances, the Census Bureau also must follow all laws and regulations governing the use of those data, and researchers may have to sign special nondisclosure forms. The other agencies also get joint approval authority of the research projects as is the case with IRS when projects include tax data.

Projects initiated internally by Census Bureau employees also must pass a review process, described in more detail in the IRS case study in the next section of this chapter. However, internal projects use a streamlined proposal review process. The Census Bureau division sponsoring the project must attest in a memorandum to the Chief of CES that the projects meet all five of the criteria. However, if the projects use tax data, they must also be reviewed by the IRS. The proposals that link records must comply with the *Administrative Records Handbook (DS-001)* dated May 16, 2001, as well as Census policies.

www.manaraa.com

IRS and Census Bureau Case Study

The IRS and Census Bureau case study investigates the 1999 Safeguard Review. The Safeguard Review is conducted by IRS to ensure "that policies, practices, controls, and safeguards employed by agencies and their agents and contractors adequately protect the confidentiality of information they receive from the IRS" (IRS, 2007b). It is an on-site evaluation of the agency receiving Federal Tax Information (FTI). Reviews are conducted by the IRS Mission Assurance & Security Services Office of Safeguards within the Office of Privacy.

Legal Dimension

Title 26, Subtitle F Chapter 61, Subchapter B, Section 6103(J)(1), states that:

"Upon request in writing by the Secretary of Commerce, the Secretary shall furnish—

***(A)** such returns, or return information reflected thereon, to officers and employees of the Bureau of the Census, and*

***(B)** such return information reflected on returns of corporations to officers and employees of the Bureau of Economic Analysis,*

as the Secretary may prescribe by regulation for the purpose of, but only to the extent necessary in, the structuring of censuses and national economic accounts and conducting related statistical activities authorized by law."

The Treasury Department is statutorily responsible for promulgating regulations to determine how Federal Tax Information (FTI) is to be provided to the Census Bureau. Much of the interaction between Census, Treasury, and IRS has been because of varying interpretations of this language and how it conflicts with Title 13. That is, Title 26 allows

www.manaraa.com

Treasury to share information, but "only to the extent necessary". Title 13, on the other hand, directs the Commerce Secretary, *"To the maximum extent possible and consistent with the kind, timeliness, quality and scope of the statistics required, the Secretary shall acquire and use information available from any source referred to in subsection (a) or (b) of this subsection instead of conducting direct inquiries." [13 U.S.C. §6(a)(c)]*. As the case study shows, the conflict between "the extent necessary" and "the maximum extent possible" led to serious repercussions for the research community.

According to the interviewees, Treasury traditionally held a very narrow interpretation of 6103(J) (1). The overall approach had been to limit the total amount of FTI released to authorized agencies. If one agency got expanded access to FTI, then Treasury attempted to limit availability somewhere else to make it a zero sum game. In addition, the statute specifically refers to statistical activities rather than to analysis. The Treasury view up until the 1999 Safeguard Review had been that researchers should not get access to FTI. Rather, analysis should be limited to Treasury's Office of Tax Policy and the Congressional Joint Committee on Taxation in order to limit the possibility of blindsiding the Administration and Congress with outside analyses and recommendations on tax policy or related policy issues. Thus, Treasury's interpretation of "to the extent necessary" in the legislation had been quite different than the Census Bureau's, causing considerable disagreement that will be discussed later in the chapter.

According to interviewees, the Census Bureau's Business Register, and its use in expanded research projects conducted at the RDCs, was at the heart of much of the controversy surrounding the 1999 Safeguard Review. The Business Register is created

www.manaraa.com

from tax records and other data, and it is a key element in the Census Bureau's ability to carry out its mission and provides the frame for the Economic Census and most economic surveys. In addition, the Business Register is used to update samples and in edit and imputation activities. It is also used to respond to requests for special reports and reimbursable tabulations from the Bureau of Economic Analysis; the Departments of Defense, Energy, and Housing and Urban Development; Small Business Administration; state and local economic development agencies; and private businesses. Business Register information is confidential under both the Census Bureau's Title 13 and Title 26 of the US Code; therefore, access is restricted to persons specially sworn to uphold the confidentiality provisions of both titles**.**

Title 26 Section 6103(p)(4) governs the safeguarding of FTI. Other federal agencies besides the Census Bureau, as well as state, and local agencies are authorized to use FTI only for specific purposes described in law. If any agency uses FTI for a purpose other than the one specifically authorized and approved by IRS, then IRS may discontinue supplying FTI and impose civil or criminal penalties on the responsible officials. This legal authority underlies the enforcement of findings from the IRS Safeguard Reviews.

In 1988, the Taxpayer Bill of Rights (P.L. 100-647) was enacted in response to widespread reports of the IRS abusing taxpayers. The Act shone a spotlight on IRS enforcement, and as a result of continuing, high visibility problems at IRS throughout the 1990s, the Taxpayer Browsing Protection Act of 1997(P.L. 105-35, 110 Stat. 1104) was enacted. The Act made it unlawful for federal employees, state employees, or other

www.manaraa.com

specified persons to willfully inspect, *except as authorized,* any tax return or return information. It provided for a monetary penalty, imprisonment, or both for violators. It also permitted civil damages for the unauthorized inspection or disclosure of tax returns and return information, and it required the taxpayer to be notified as soon as practicable if any person was criminally charged by indictment with inspecting or disclosing the taxpayer's return or return information. The following year, Congress passed the IRS Restructuring and Reform Act of 1998 (P.L. 105-206, 112 Stat. 685, 778), which expanded taxpayer rights and called for reorganizing the agency into four operating divisions aligned according to taxpayer needs.

These legal changes and the resultant oversight from Congress and the (then named) General Accounting Office (GAO) caused major changes in the Safeguard Reviews being conducted by IRS. As described by some of the Census Bureau study participants, these changes caught the Census Bureau off guard, and resulted in a 1999 Safeguard Review that had a significantly different outcome than the previous Safeguard Review that had been conducted during 1992. That is, while the Census Bureau was aware of the increased oversight and negative publicity at IRS, the staff at Census did not expect the Safeguard Review to change significantly as a result.

Perceptual Dimension

Safeguard Reviews are supposed to be conducted every three years. When IRS had conducted the 1992 Safeguard Review of the Census Bureau, it appeared that the Census Bureau was using FTI for improving the Business Register and other statistical purposes that supported Census' authorized activities under Title 13. Another review

www.manaraa.com

was due to be conducted in 1995, but, according to an interviewee, due to a lack of resources at IRS, it was delayed until 1999. However, during the1992-1999 period, both IRS and the Census Bureau changed significantly - IRS regarding the stringency of research project reviews and Census regarding the greatly increased availability of comingled data sets for internal and external research purposes. These changes were to cause serious and long lasting repercussions in the relationship between Census and IRS regarding sharing of FTI. As described by interviewees, the IRS staff felt caught off-guard because they were not explicitly consulted about the expansion of research projects. The Census Bureau staff, on the other hand, thought that referencing the projects in the Annual Safeguard Activity Reports to IRS was sufficient notification.

Many areas within the Census Bureau rely on FTI, including the Center for Economic Studies (CES). Although CES had been started during the 1980s in order to give data access to researchers, it was really just getting off the ground in 1992. According to the interviewees, during the 1992 Safeguard Review, CES reported to the IRS that it was not using tax data. However, it appears that at that time there was not a clear understanding between IRS and the Census Bureau as to what constituted FTI covered by section 6103.

Because the Census Bureau had been comingling the IRS information with its own survey data and other administrative record data, and had assigned its own unique identifying numbers to establishments, it didn't consider research projects that used the Business Register or information based on the Business Register to be FTI. Thus it reported to IRS in 1992 that CES was not using FTI. The 1992 Safeguard Review went

www.manaraa.com

smoothly, primarily because IRS did not actually validate this information as reported, and the Census Bureau was found in compliance. Interviewees indicated that the review was short and not controversial at all.

By 1999, IRS was taking a much tougher stance on the reviews, reflecting the significant congressional and public pressure on IRS to improve its own performance. During the early 1990s, GAO carried out numerous studies and investigations of IRS to monitor implementation of the 1988 Taxpayer Bill of Rights Act. In a 1992 report on implementation of the Act, GAO raised concerns about the need for IRS to clearly delineate responsibility for protecting the privacy of taxpayer information (GAO, 1992). During 1993, GAO again identified weaknesses in IRS' general controls over its computer systems which resulted in various problems, such as unauthorized access to taxpayers' account information by IRS employees (GAO, 1993).

According to GAO (GAO, 1994), an internal audit conducted by IRS (IRS, 1992) found that 368 IRS employees in one region alone had used the IRS automated tax record system to gain access to nonwork-related taxpayer accounts. IRS responded to its GAO and congressional critics by stating that there would be "zero tolerance" for such snooping in the future. To prevent unauthorized access to taxpayer accounts, IRS was planning to limit some employees' access to only specified accounts authorized by a manager for official purposes. IRS also indicated that it planned to build security controls into the new automated tax information management system that was replacing the old system in place during the early 1990s, in order to minimize unauthorized access of taxpayer information

www.manaraa.com

However, IRS was still coming under severe congressional criticism during the late 1990s. According to the Senate Republican Policy Committee (Craig, 1997), during various 1992-1994 probes, GAO had found that 1,300 IRS employees were suspected of "snooping" in confidential taxpayer files. Additionally, GAO found that of the 1,515 reported cases of "snooping" during 1994-1995, only 23 employees were fired. The Senate Republicans cited an April 3, 1997, *Wall Street Journal* article about an IRS contract employee who had browsed taxpayer records, was convicted in December 1995 of 13 counts of wire and computer fraud but whose conviction was thrown out by the First U.S. Circuit Court of Appeals on grounds that "it wasn't a crime because the prosecution didn't prove that he had used the information or disclosed it to anybody. (WSJ, 1997)"

In quick succession, Congress enacted the Taxpayer Browsing Protection Act of 1997(P.L. 105-35, 110 Stat. 1104) and the IRS Restructuring and Reform Act of 1998 (P.L. 105-206, 112 Stat. 685, 778. Then, in 1999, IRS conducted the Census Bureau's Safeguard Review, which took place in this atmosphere of heightened scrutiny and intense outside criticism of IRS.

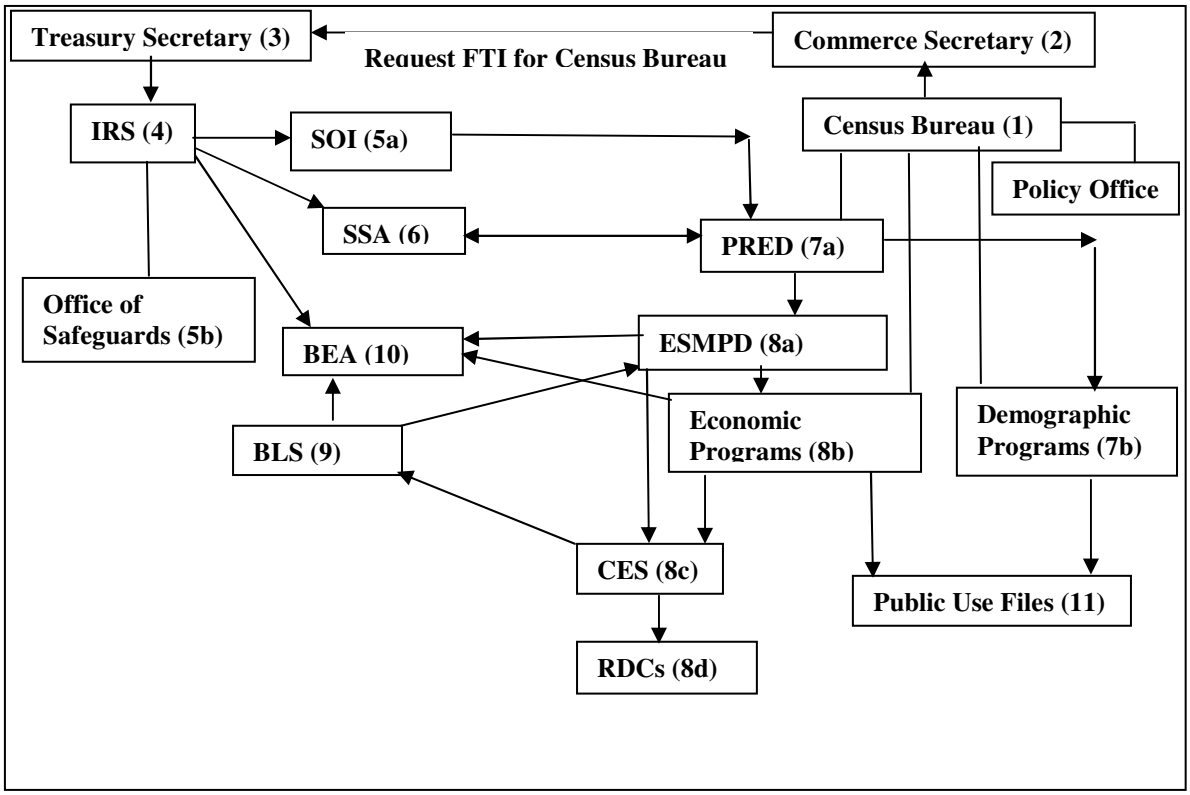
The Census Bureau, in the meantime, was aware of these changes at IRS, but not making the connection to the Safeguard Reviews and its own use of FTI. According to the interviewees, Census was continuing to develop its research program at the RDCs, and still operating under the belief that research based off the Business Register was one step removed from FTI and not covered by Title 26. The differences in the Census and IRS interpretations of regulations and laws governing FTI, and the resultant activities created the perfect storm conditions that occurred during the 1999 Safeguard Review.

www.manaraa.com

Organizational Dimension

The process for sharing FTI between the IRS and Census changed significantly after the 1999 Safeguard Review. Figure 6 below was developed from oral input from the interviewees that was subsequently validated by a subset of them, and shows the process flow that was in place prior to 1999 for sharing of FTI between the Census Bureau and IRS.

Figure 6 Pre-1999 Process Flow



As Figure 6 shows, the request for FTI was initiated by the Census Bureau (1). The Census Bureau would send a request letter to the Secretary of Commerce (2), who would sign it and send it to the Secretary of Treasury (3). (The lack of a number in the Census Policy Office box is indicative of its role serving as the staff that prepared the requests for the Census Bureau Director. In addition, the Policy Office had the lead role

www.manaraa.com

in staffing the Census Bureau's response to IRS Safeguard Reviews and in producing and enforcing the bureau's privacy policies.)

IRS (4) was authorized to share FTI with other agencies. The task of reviewing Census Bureau projects for appropriateness was delegated to the Statistics of Income (SOI) Division (5a). However, prior to 1999, IRS did not explicitly require that it review each individual project before it was approved. IRS shared data files with SSA (6), Census (7a), and BEA (10). SSA was authorized by both Title 26 section 6103 and the Social Security Act to receive FTI. SSA processed W2 forms to collect lifetime earnings histories and merged these with self employment records from IRS. This so-called Master Earnings File was then shared with the Census Bureau.

Data files from SSA and IRS were physically or electronically received by the Census Bureau Program Research and Evaluation Division (PRED), which validated the records and removed the personal identifiers, a process described in more detail in the technical dimension section (7a). Various files were then distributed to several areas of the Census Bureau to carry out a variety of projects, including the Economic (8b) and Demographic (7b) Program Directorates. The Economic Statistical Methods and Programming Division (ESMPD) (8a) within the Economic Directorate created the Business Register, comingled administrative and survey data, and shared files with both CES(8c) and BEA (10). CES, in turn, made data files available to researchers in the RDCs (8d) and to BLS (9) in order to compare the two agencies' Business Registers. Both the Economic and Demographic Program Directorates prepared files that were scrubbed through a Disclosure Review process to remove identifiable microdata in order

www.manaraa.com

to create Public Use files that could be made available to all researchers and the public. The role of the IRS Safeguard Office (5b) was to carry out an audit of the Census Bureau every three years to assure that FTI was being handled properly.

The 1999 Safeguard Review, which was expected to take about four months from start to finish, ended up continuing for 18 months. Numerous problems were discovered. According to the interviewees who were directly involved in the review, some of the most significant included:

- There were many lapses in protocols for requesting FTI. (Over time, there had been some redelegations of request authority at Commerce and Census that IRS challenged.)
- Census had shared FTI with other agencies without IRS permission and without auditing how FTI was being protected - most notably the Bureau of Labor Statistics (BLS)
- Census had received FTI from SSA in order to match it with survey data but had not informed IRS
- FTI was embedded in numerous internal files being used by Census employees without required Title 26 safeguards (although Title 13 safeguards were being observed).
- Census was unable to tell IRS exactly how many and which projects were using FTI within the bureau.
- IRS had not been involved in reviewing any of the proliferating research proposals being carried out at the RDCs.

www.manaraa.com

According to the interviewees who participated in the 1999 review, many of the instances above arose as a result of significant RDC growth since 1992, under the direction of CES, which had expanded use of FTI in the RDCs without proper IRS approval. Although the Census Bureau had been sending documentation to IRS listing CES projects that were using tax data, IRS had not followed up to verify that the Census Bureau was in compliance with all laws and regulations, and technically, had not approved the projects. An important point for the Census interviewees was that the Census Bureau did not believe it was sharing FTI with other agencies. Rather it had agreements to share Title 13 data with people from other agencies who had been given Special Sworn Status. Interviewees indicated that IRS, on the other hand believed both that the comingled files did constitute FTI, and that Treasury and IRS were somewhat uncomfortable with the number of non-Census employees who had been given Special Sworn Status to gain access to the microdata.

In another instance, the approvals that Census had earlier received from IRS to match IRS earnings data provided by SSA with data from the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP) had lapsed. However, neither SSA nor Census informed IRS that the work was continuing, nor did they ask for continuing authorization to use IRS supplied data. In addition, Census first claimed that an MOU existed allowing this project, but subsequently could not produce an MOU and had to acknowledge that one didn't exist. This exacerbated the lack of trust that was emerging between Census and IRS.

www.manaraa.com

In yet another instance, Census and BLS had embarked on a joint project to understand the differences in the business lists used by both agencies. Both BLS and Census maintain their own business registers created from independent sources, and both registers are used to supply aggregate data inputs for certain national statistics generated by the Bureau of Economic Analysis (Fixler & Landefeld, 2006). The differences in the two Business Registers could affect programs that rely on BEA statistics such as per capita state personal income, which is used to determine the federal share of Medicaid funds allocated to each state.

The BLS register includes monthly data on employment and quarterly data on total wages (payroll), industry classification, and geographic location. BLS incorporates information from the State Workforce Agencies (SWA) and the Multiple Worksite Report (MWR), which collects monthly employment and quarterly wage information on establishments associated with multi-establishment firms within a state.

As mentioned earlier, the Census Bureau Business Register includes data for mid-March employment from the CPS, annual and first quarter payroll, industry classification, and geographic location. FTI provided by IRS is the primary source of information on the existence, location and operating status of businesses from the Business Master File and the Business Income Tax and Payroll Tax forms. In addition, the Census Bureau receives other data, primarily industry classification codes, from SSA and BLS.

Another difference between the two registers is that BLS data include all workers and payroll items covered by federal and state Unemployment Insurance (UI) laws. However, organizations, employees and payroll items that are not covered in the UI

www.manaraa.com

system may be included in the Census list. Also, the Census Bureau updates its Business Register continuously, while BLS updates quarterly on a flow basis.

Census Bureau and BLS designed and began a three part Business List Comparison Project that required sharing of microdata from the two registers. According to the interviewees, the purpose was to better understand the differences between the two registers and improve economic statistics as a result. The Census Bureau sent its data files to BLS, which processed the files in a facility run by a contractor. Unfortunately for Census, BLS, and the project, when IRS conducted the 1999 Safeguard Review, it found that (1) the project had never been approved by IRS; (2) IRS had never authorized BLS to receive FTI, even if it was completely comingled with other data in the Census Bureau Business Register; and (3) the contractor facility used by BLS was not secure, and the Census Bureau Business Register tapes were discovered sitting out in an open area at the facility by IRS investigators. Interviewees recall the last point as being the main issue.

All the interview participants concurred that the outcome of the 1999 review fundamentally changed the processes and procedures governing the sharing of tax records between the Census Bureau and IRS. The combined effect of all the negative findings was that IRS decided to immediately pull back all FTI from the Census Bureau. However, both IRS and Census recognized the consequent effect on important work. In addition, while IRS found that Census had not adequately followed the reporting requirements and had embarked upon joint projects with other agencies, there had not actually been any breaches or leaks of information from Census staff, the RDCs , BLS, or the BLS contractor. This is most likely because Title 13 is as strict as Title 26 about protecting data, and the Census Bureau had been applying Title 13 protections to the

www.manaraa.com

comingled data. As a result, IRS let existing projects continue, although it put a hold on approval of new research projects and stopped the joint Business Register project Census had begun with BLS.

The interviewees indicated that after protracted negotiation, on September 19, 2000, the two agencies entered into an agreement establishing the processes and criteria for review and approval of projects containing FTI. The so-called "Criteria for the Review and Approval of Census Projects that Use Federal Tax Information" (Barron, 2000) form the basis of the current processes and practices. The IRS agreed to supply FTI to Census, and Census agreed to let the IRS approve all internal and external projects using FTI, regardless of whether or not data were comingled and identifiable as tax data.

In addition, under the agreement, projects seeking to employ FTI also needed to demonstrate that their predominant purpose was to benefit Census Bureau programs and that they met at least one of the following nine criteria:

- Understanding and/or improving the quality of data produced through a Title 13, Chapter 5 survey, census, or estimate;
- Leading to new or improved methodology to collect, measure, or tabulate a Title 13, Chapter 5 survey, census, or estimate;
- Enhancing the data collected in a Title 13, Chapter 5 survey or census. For example:
    - Improving imputations for non-response;
    - Developing links across time or entities for data gathered in censuses and surveys authorized by Title 13, Chapter 5;

www.manaraa.com

- Identifying the limitations of, or improving, the underlying Business Register, Master Address File, and industrial and geographical classification schemes used to collect the data;
- Identifying shortcomings of current data, collection programs and/or documenting new data collection needs;
- Constructing, verifying, or improving the sampling frame for a census or survey authorized under Title 13, Chapter 5;
- Preparing estimates of population and characteristics of population as authorized under Title 13, Chapter 5;
- Developing a methodology for estimating non-response to a census or survey authorized under Title 13, Chapter 5; and
- Developing statistical weights for a survey authorized under Title 13, Chapter 5. (Census, 2000)

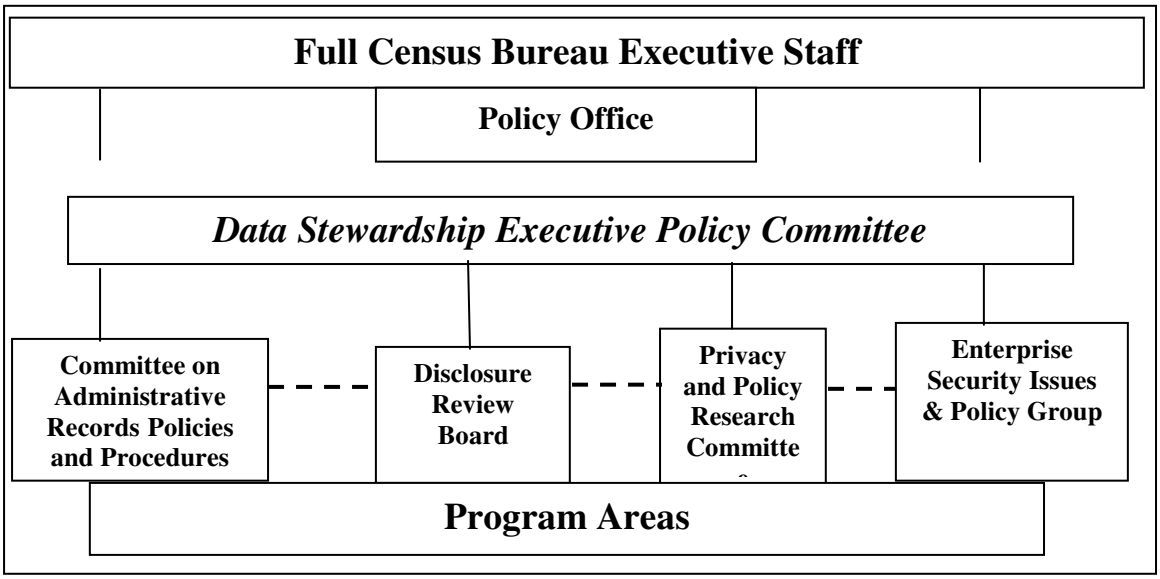
Under the new procedures, the Census Bureau and IRS jointly determine whether the predominant purpose of a project proposing to use FTI meets these criteria. Other post- 1999 Safeguard Review Census Bureau publications that govern the use of data at RDCs include:

- DS-002, *Articulating the Title 13 Benefits of Census Bureau Projects (DS-002)*, dated October 10, 2002
- DS-001, *Administrative Records Handbook (DS-001),* dated May 16, 2001
- DS-014, *Record Linkage Policy (DS-014)* dated February 5, 2004

www.manaraa.com

In addition to executing the agreement with IRS, the Census Bureau undertook several actions after the 1999 Safeguard Review to strengthen its own internal processes. Most significantly, the Census Bureau Data Stewardship Executive Policy (DSEP) Committee was created. The DSEP Committee was established to help the Census Bureau achieve a systematic and integrated balance between business decisions and privacy and confidentiality constraints. (To provide full disclosure for this study, note the author was instrumental in establishing the DSEP Committee and also chaired it.) The goal for data stewardship was to develop policies that balanced data quality and use against both the legal constraints for confidentiality and privacy and ethical standards as promulgated by professional organizations such as the American Statistical Association, the American Association of Public Opinion Research, and others. Figure 7 shows the data stewardship structure that was established at the Census Bureau in 2000 (Potok, 2002).

Figure 7 Data Stewardship Structure at the Census Bureau



According to Census Bureau internal documentation, the Data Stewardship Executive Policy (DSEP) Committee was composed of a subset of the Census Bureau's Executive staff and chaired by the Principal Associate Director. It was staffed by the Policy Office. Reporting to the DSEP Committee were four subcommittees with

www.manaraa.com

membership that cut across organizational lines. These included the Committee for Administrative Records Policies and Procedures, which recommended policies to the DSEP Committee; the Disclosure Review Board, which reviewed procedures and associated data files prior to making them public to assure that individual data could not be identified; the Privacy and Policy Research Committee, which specifically focused on privacy issues; and the Enterprise Security Issues and Policy Group, which focused on related IT Security issues.

To begin to implement the new agreement, Census prepared a Safeguard Procedures Report, which is an IRS required record of how FTI is processed and protected from unauthorized disclosure. The required Annual Safeguard Activity Report, which advises IRS of minor changes to procedures or safeguards, was changed as well so that Census began reporting activities at the project level, rather than at the program level.

Census also established a point of accountability in the Policy Office to coordinate the process of getting IRS approval of projects prior to commencement of those projects. According to the interviewees, a workgroup led by the Policy Office staff developed internal processes for handling project reviews using both IRS and Census criteria. A tracking database was created for all such projects that could be shared with IRS, called the Administrative Records Tracking System (ARTS). ARTS documents all projects and the data sets they are using. Census also created a new process to facilitate reviews on both internal and external projects. For internal projects, Division Chiefs must certify against a checklist that a project meets all established criteria. The project is then reviewed by the data custodians. Initially, the custodial divisions were (1) the

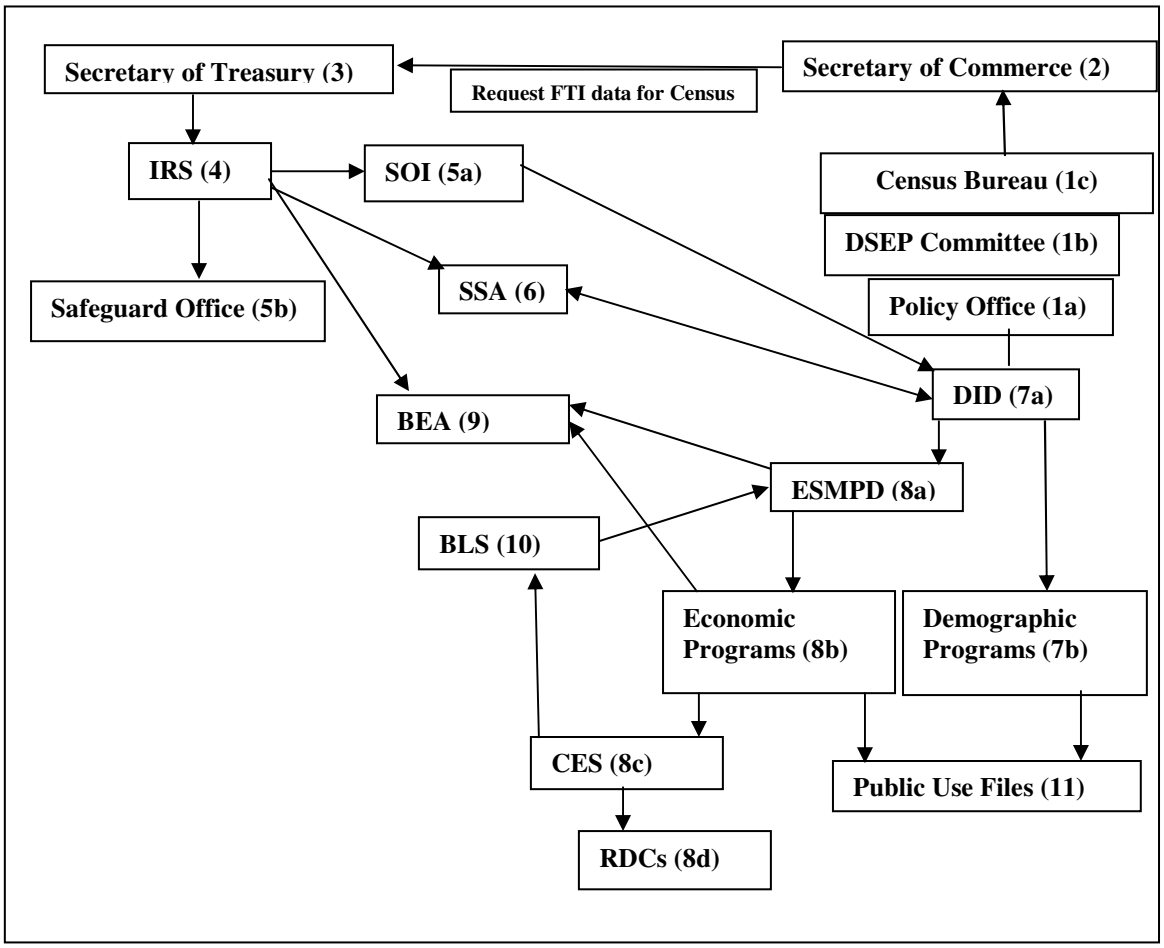
www.manaraa.com

Administrative Records Research, Specifications, and Operations Staff of the Planning, Research & Evaluation Division (PRED) in the Standards and Methodology Directorate, and (2) the Economic Statistical Methods & Programming Division (ESMPD) in the Economic Programs Directorate. However, the Census Bureau has since reorganized, and the Research Directorate and PRED were eliminated. The current data custodians are ESMPD and a newly formed Data Integration Division (DID) in the Demographic Programs Directorate.

Once internal reviews are completed, projects are sent to IRS for review and approval. Based on information provided orally and subsequently validated by the interviewees, Figure 8 was developed to show the new flow of FTI that was established after the 1999 Safeguard Review. The strengthened role of the policy office and the newly formed DSEP Committee are key elements in satisfying IRS requirements for approving and tracking projects. In addition, each division now has a designated IT security officer who is responsible for preparing for the Safeguard Review.

www.manaraa.com

Figure 8 Post 1999 flow of FTI



In addition to establishing the internal review process and the ARTS database, the interviewees indicated that the Census Bureau also developed two key administrative records related policy documents during 2002. The first provides guidance on controlling non-employee access to Title 13 data (Census, 2002a), particularly when it is appropriate to confer Special Sworn Status under Title 13 to individuals working on protected data, and clarifies when it is appropriate for the Census Bureau to provide access to protected data offsite. In order for a nonemployee to receive Special Sworn Status, the project must:

www.manaraa.com

- Require access to confidential data,
- Benefit Title 13 programs, and
- Be a viable project that is feasible, abides by use restrictions in interagency agreements, adhere to disclosure review requirements, and be consistent with the bureau's privacy principles. (Census, 2002a)

In order to provide off-site access to the data, the project must:

- Provide a technical and logistical advantage,
- Meet the required security model for off-site access,
- Have a legal or regulatory functional separation of the data collected for statistical purposes if a government agency, and
- Obtain approval from the DSEP Committee (Census, 2002a).

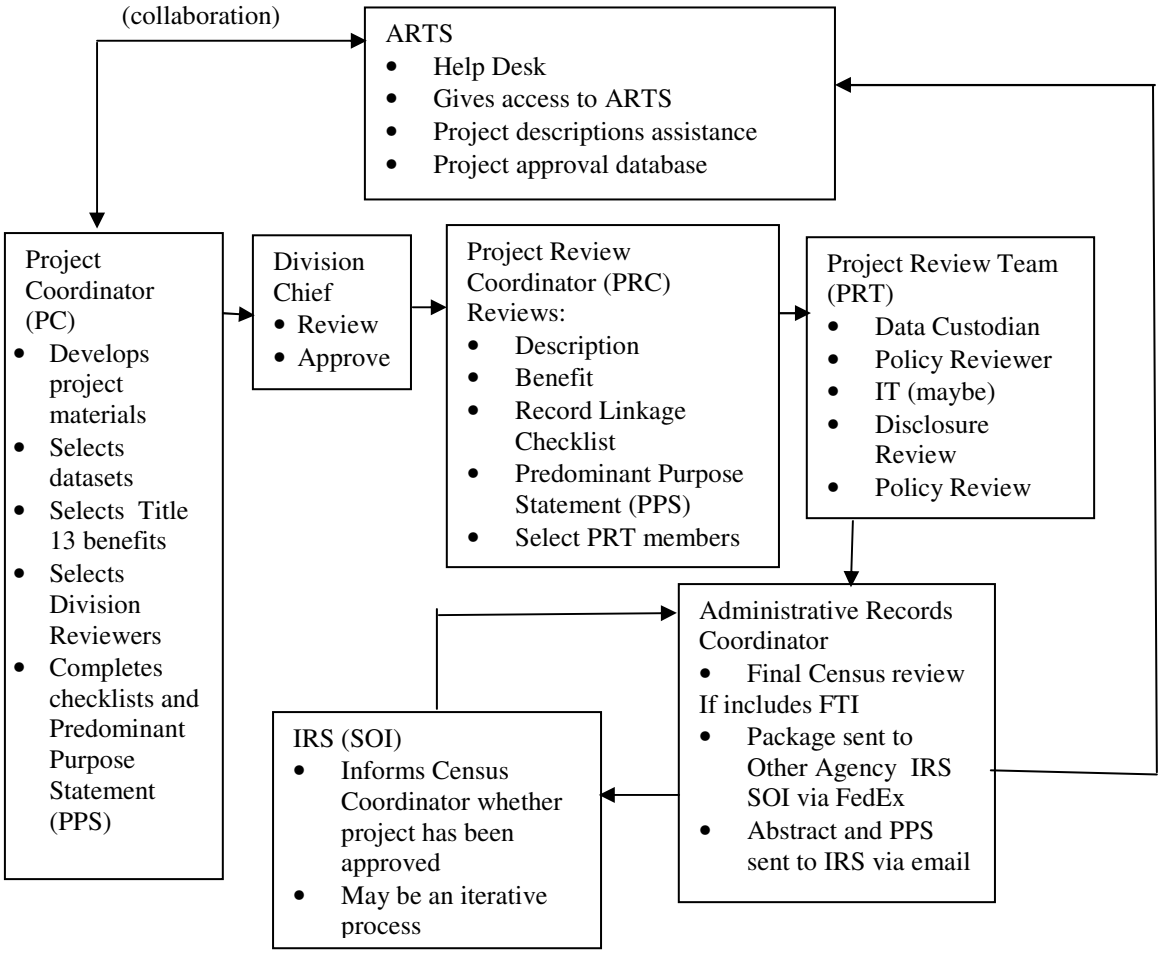
The second policy provides guidance on negotiating collaborative arrangements with other agencies in order to acquire administrative record data for Title 13 purposes (Census, 2002b). The policy establishes principles that ensure that projects using the records are legal and ethically appropriate. The policy covers the following areas:

- acquisition of the records,
- Census Bureau deliverables to agencies providing the source data (including data products, data enhancements, statistical models, access to data, file editing or other technical services), and
- Ongoing arrangements, such as Census Bureau's practice of processing IRS data and then providing back to IRS aggregated counts on the number of IRS returns filed for various forms(Census, 2002b).

www.manaraa.com

Based on narrative documentation provided by the Census Bureau, Figure 9 was developed to illustrate the review process for Census Bureau internal projects.

Figure 9 Census Bureau Project Review Process



Technical Dimension

Prior to 1999, FTI files at Census were processed by a special administrative records unit in the Planning, Research, and Evaluation Division (PRED) that comingled and took out the individual identifiers in the data. The IRS data tapes were sent to the Census Bureau's headquarters and later in the decade to the new secure computer processing facility in Bowie, Maryland, where quality checks were run on the number of

www.manaraa.com

records and the readiness of data. The records were validated by running them through the SSA's NUMIDENT file. This allowed Census to compare the names, addresses, and SSNs from IRS and SSA.

The Census Bureau uses a unique process it developed especially for record linkage called the Personal Identification Validation System (PVS). The PVS was not developed as a result of the 1999 Safeguard Review, but is a part of the ongoing administrative records research program at the Census Bureau, and is managed by the Administrative Records Research staff. It uses probabilistic matching to verify SSNs in survey or other data files against the NUMIDENT (Obenski, 2006). SSNs are matched using names, dates of birth and gender to verify the matches. Weights are used to define acceptable matches. If there is no SSN provided with the record, the file is sent to the search phase, which consists of an address-based search followed by a name search. The address-based (or geokey) search is conducted by logically grouping the geographic data and then subsequently relaxing the geographic criteria as the records are looped through multiple passes. Any remaining unmatched records go through a name search. After the verification and search phases are completed, a new file is created containing the (1) verified and assigned SSNs, (2) inconclusive SSNs, and (3) original records that didn't go through the PVS process because the records were blank or a respondent refused to give an SSN on a survey.

Under the Census Bureau's privacy policy, if a respondent refused to give an SSN when asked, the bureau cannot independently search for the SSN and add it to the file. Only the validated records with verified and assigned SSNs are used in record linkage.

www.manaraa.com

Once a match is found, the SSN is removed from the IRS or other file and replaced with a unique but randomly generated identifier, called a Personal Identification Key (PIK). Every SSN receives a PIK through this process. Thus, the actual SSNs are never used in the research projects themselves. Original files are kept on a server that can't be downloaded, and access is only granted to two Census employees and two backups.

Another area of keen interest related to safeguarding confidentiality of records is disclosure review, which was not a direct part of the 1999 Safeguard Review but is a key element of sharing microdata files. These techniques use statistical methods to ensure that data files released to the public do not identify individuals or businesses, or allow others to manipulate data in such a way that individually identifiable information may be revealed. There are a number of disclosure avoidance methodologies, such as data suppression and modification, data-rounding, top-coding, data swapping, thresholds, random noise, cell suppression, and complementary cell suppression. Differing methods are used for different types of data releases (FCSM, 2005).

Title 13, U.S.C., Section 9 prohibits the publication or release of any information that would permit identification of any particular establishment, individual, or household. The Census Bureau has an internal Disclosure Review Board that sets the confidentiality rules for all data product releases. A checklist approach is used to ensure that potential risks to the confidentiality of the data are considered and addressed before any data are released. All data collected or maintained by the Census Bureau under Title 13 need disclosure protection, including Title 13 information commingled with or enhanced by information from other sources such as survey or administrative records data. Some

www.manaraa.com

commingled information may be subject to disclosure protections from the source agency, as well.

Disclosure avoidance is a much larger topic that goes beyond the linking of administrative records and survey data. It is beyond the scope of these case studies and will not be addressed in this study except to note that the Census Bureau has a rigorous review process in place for any data that will be released.

Human Dimension

According to those interviewees who work directly with data sharing projects, they are generally initiated by an individual or a small group of individuals within an agency, often the Census Bureau. The nature of these projects reflects the specific interests of the person generating the project. These interests may or may not coincide with the interests of the agency providing administrative records. In the case of IRS, it is required by law to provide tax records to the Census Bureau. However, IRS gets very little benefit from providing the records. In fact, IRS has a disincentive to share records, because of the close oversight from Congress and the stringent language of Title 26. For the IRS leadership, there is much risk and little gain. In the case of the Office of Tax Policy, there is also a disincentive to share tax data that may be used to conduct policy research in competition with its own research. Because the only role of IRS is to safeguard the FTI, it focused a lot of attention on the project approval process after the 1999 Safeguard Review. Participants described the review process as unwieldy and time consuming. Each project use was closely scrutinized by IRS before approval was given to proceed. This created a negative reaction from those participating in the review process

www.manaraa.com

and spawned an entire capability in proposal writing at the Census Bureau directed towards describing projects such that they can be approved. Additional guidance was given to outside researchers who aspired to gain approvals for RDC projects. It is not clear whether this resulted in more secure projects or simply better written proposals. Perhaps both ends were achieved.

However, in 2007, a major step forward was taken when the Census bureau director issued a policy statement stating that analytical research is a valid use of FTI under Title 13. By issuing that blanket statement, reviews by the IRS were considerably shortened, because IRS no longer had to make that specific determination. According to interviewees, the IRS reviews now take an average of 1-2 weeks. The review process is still long, but most of the hold ups are now within the Census Bureau.

After the 1999 Safeguard Review was completed, but before it could be closed out, issues had to be satisfactorily resolved between IRS and Census. This included the Census Bureau acknowledging the IRS definition of FTI in comingled data. However, an impasse was reached. At that point, IRS wanted to close off all access to tax data due to the review findings. The Census Bureau believed it had a statutory right to receive tax data regardless. During this period, ongoing projects in the RDCs and at the Census Bureau were allowed to continue. However, no new projects were approved, bringing new research to a halt. This had a significant detrimental effect on numerous researchers in academic settings, as well as Census Bureau staff embarking on new projects. Eventually, the issue had to be resolved by the head of the Office of Information and Regulatory Affairs (OIRA) in the Office of Management and Budget (OMB), who served as a mediator between Treasury and IRS and Commerce and Census, and brokered an

www.manaraa.com

agreement. The result was the September 2002 joint agreement on the project review process.

As mentioned, the process agreed to by the Census Bureau and IRS for using FTI is often characterized by its users as cumbersome and slow. People involved in the process believe that it lengthens project approval by months. Part of the reason for the delays is the difficulty of keeping the attention of the many people in the approval chain focused on the reviews. For most people, the approvals are other duties they carry out in addition to their full-time workload. As a result, approvals can sit on people's desks as lower priority items for long periods before moving on to the next review stage.

Several projects were affected either directly or indirectly by the 1999 Safeguard Review. Most notable was the Longitudinal Household Employer Dynamics (LEHD) project. As described in Chapter 1, LEHD was started by the Census Bureau in 1998 by combining existing data from censuses, surveys, and administrative records to create new data and products. Under this program, quarterly worker and business wage records are supplied to the Census Bureau by states. The Census Bureau merges the state records with other data from sources such as Census 2000, the American Community Survey, IRS summary and detailed earnings records, SSA benefit data, and the Business Register to produce new data and products. These new products include a longitudinal national frame of jobs, and an associated data infrastructure that describes where workers live, where people work, and companion reports on age, earnings, and industries by geographic block. The goal for the program is to create a data infrastructure that captures the complex interactions among households and businesses at the microeconomic level and characterizes the dynamics of the modern economy (Abowd et al., 2004).

www.manaraa.com

LEHD was controversial when it began, both within the Census Bureau and at IRS. It was one of the primary drivers, along with the Safeguard Review, for forming the DSEP Committee. LEHD was much more ambitious in scope than any previous projects linking administrative records. In addition, one of the principal researchers wanted to work offsite and have remote access to data. Within the bureau, there were tensions between the strong privacy advocates on one hand, including those concerned about combining so many types of administrative records together from various sources, as well as the IT security organization, and on the other hand, those who wanted to develop a new data project that public policy researchers would find invaluable. For example, some privacy advocates were concerned about informed consent and whether there should be explicit language informing individuals applying for unemployment benefits that their records would be linked by the federal government. Another concern was whether the linked data could be made disclosure proof, because it contained a lot of information on individuals from several sources.

According to interviewees, many of the IRS concerns were driven by OTA in Treasury's Office of Tax Policy. The LEHD project needed information from W2 forms in order to match employer and employee data, and include income. However, while the Census Policy Office argued that the regulations accompanying Title 26 sections 6103(j) (1) (a) and (1)(b) permitted access to the W-2 data, in fact, Treasury's was correct that those forms were not identified in the regulations. Nor did Treasury particularly want to identify the W-2s in the regulations, in order to limit access to FTI. This was driven by a strong belief that analysis of tax data should be done by OTA and the Congressional Joint Committee on Taxation, not other federal agencies. Thus, providing W2s to Census for

www.manaraa.com

LEHD, specifically for a match to the CPS and the SIPP, required a change in Treasury regulations. Treasury fought hard to keep the W2 information out of the hands of the LEHD researchers. The argument was made that if Treasury allowed access to W-2 data only for a CPS and SIPP match, the result would be a "slippery slope" leading to Census wanting to use W-2 population data, with serious perception consequences for the Treasury and IRS. The poor results of the 1999 Safeguard Review gave additional ammunition to Treasury and IRS to hold firm to their position.

As a result, the Census Bureau was forced to use an alternative path to get the link between workers and firms, negotiating state by state to get unemployment insurance wage record data, which is administered by the Employment and Training Administration at the Department of Labor, as well as ES202, or business data, which is administered by BLS. Ten years later, there are still some states missing from LEHD data, because agreements are not yet in place. The project took much longer and cost substantially more than it would have if Treasury and IRS had shared the W2 files. Even so, the LEHD program calculated that the additional cost of processing administrative records was 2 cents per case per fiscal quarter. This is not costly when compared to an average cost of a face-to-face interview of up to $1,000 per case if primary data collection had been required (Lane, 2009).

Eventually, after two years of effort to clearly define that the only use of W-2 records was to permit linkages to CPS and SIPP data, and after some changes in personnel at Treasury, the Treasury regulations were changed to allow the Census Bureau to get access to W2 data. However, the W2 data could only be used for very limited purposes and the temporary regulations were in place for two years. Eventually, as

www.manaraa.com

personnel changed again, the very tight restrictions in the regulations were changed so that the Census Bureau could get access to the W2 data as long as it was for an "authorized" use. Interviewees indicated that concerns still remained at Treasury about the "slippery slope" of Census Bureau access. However, by the time these changes were put in place, LEHD was too far along to take advantage of this shortcut.

SSA was also leery about LEHD, primarily because it was such a large project, and it wasn't clear whether there would be any privacy concerns that might be raised later. However, SSA had a long history of sharing the NUMIDENT file with Census, and it allowed it to be used for LEHD.

Summary

The Census Bureau has long made use of its statutory authority to request tax data from IRS. The process of requesting, using, and safeguarding data has evolved over time, reflecting changes in the external environment that include technological advances in data processing, and increased sensitivity on the part of the government and the public to privacy and confidentiality issues. Although sharing of tax records reached a nadir in 1999, when the IRS Safeguard Review turned up numerous disagreements, as well as noncompliance issues within the Census Bureau, the painful process of resolving these problems led to significant changes in how FTI is protected at Census. These changes, primarily new documented policies and additional layers of project review, have been very effective to date in assuring that access to tax data provided by the IRS is secure, restricted, and in compliance with laws and regulations. At the same time, the review process is long and burdensome and relies heavily on the judgments of multiple levels of individual reviewers. To try to overcome this barrier, the Census Bureau has developed

www.manaraa.com

expertise in how to coach prospective researchers through the approval process and has lengthy guidance on its website, in particular for the RDCs. However, there is no evidence that the value or relevance of approved projects has improved since 1999. Rather, the project justifications are more complete, the tracking of the projects is more organized, and IRS staff is able to exert control over how tax data are used even after turning them over to the Census Bureau, which seems to satisfy IRS.

www.manaraa.com

CMS and Census Bureau Case Study

The Census Bureau began receiving microdata from CMS in the mid-1990s. The file sharing was the result of recommendations from National Academy of Sciences panels suggesting that using Medicaid enrollment data could help improve Census Bureau survey data (Duncan, Jabine, & Wolf, 1993b). The Census Bureau formed the Administrative Records Research, Specifications, and Operations Staff of the Planning, Research & Evaluation Division (PRED) in the Standards and Methodology Directorate to handle these files and other administrative record files, and, according to the interview participants, received the first file from CMS in 1998. This case study examines the environment in which CMS and the Census Bureau share data and how the passage of the Health Insurance Portability and Accountability Act (HIPAA) has affected that environment.

Legal Dimension

CMS is governed by the Privacy Act of 1974 under 5 U.S.C. 552(a) in regard to sharing of administrative records with the Census Bureau. Section 552(a) defines and governs the release and sharing of CMS administrative and statistical records with individual identifying information between agencies and with the public. In addition, regulations from the Department of Health and Human Services (DHHS) implement section 3 of the Privacy Act by establishing agency policies and procedures for the maintenance of records (45 U.S.C. 552(a)). Section 5b(9)(b)(4) specifically allows release of individual records to the Bureau of the Census for purposes of planning or carrying out a census or survey or related activity pursuant to Title 13, without obtaining

www.manaraa.com

the consent of the individual whose record is being shared. In addition, section 5b allows release of records to the National Archives, to another government agency for civil or criminal law enforcement activity, to either House of Congress, and to the Comptroller General and the Government Accountability Office (GAO) without obtaining additional consent. HIPAA also affects the handling of records but primarily covers the health care industry. Covered entities under HIPAA are health care providers, employer sponsored health plans, health insurers, health care clearinghouses, and Medicare drug plan providers.

As described by the interview participants, the record sharing process between CMS and the Census Bureau historically had fewer layers of review than the IRS-Census process, because the Privacy Act provisions are not overly proscriptive. Even after enactment of HIPAA, the process for exchanging records remained less subject to internal reviews, although Memoranda of Understanding (MOUs) have always been executed between the two agencies to cover the specifics of how files are to be shared, used, and destroyed after project completion. However, in 2003, the Department of Health and Human Services released the HIPAA Privacy Rule, which regulates the use and disclosure of certain information held by covered entities. Following that release, a Privacy Board was established at CMS. The interviewees described how the Privacy Board made several changes in how data were handled at CMS. Even though HIPAA did not apply to records being shared with the Census Bureau, it did create an overall trend towards increased data protection at CMS.

Because the data sharing relationship the Census Bureau enjoys with CMS is governed by a less stringent standard than those imposed by IRS, CMS participates in

www.manaraa.com

joint projects with the Census Bureau. Although CMS is not a statistical agency, shared files are protected under Title 13, and the Census Bureau does not provide to CMS linked survey data that can be identified at the person-level.

The issue of informed consent does arise when combining CMS records and survey data. As mentioned in chapter 2, the "routine uses" clause in the Privacy Act allows a broad array of people in different organizations inside and outside of government to have access to CMS records containing personal medical information with identifiers attached *without the consent of the individuals* whose records are being shared. Prior to a 2002 revision in the Federal Register, CMS had specifically mentioned the Census Bureau as a routine user of CMS data. However, citing Exception 4 to the Privacy Act, which allows release of data to the Census Bureau under Title 13, sharing of records was subsumed under routine use 2a, under which CMS may release information without the consent of the individual to another Federal or state agency, agency of a state government, an agency established by state law, or its fiscal agent to contribute to the accuracy of CMS's proper payment of Medicare benefits. Consequently, Census Bureau use of CMS data must contribute to the accuracy of CMS benefit payments in addition to any other statistical purposes of the Census Bureau.

In the statement to Medicare beneficiaries that is required by the Privacy Act (known as the Privacy Act Statement and found at www.cms.hhs.gov/MinimumDataSets20/Downloads/MDS%20Privacy%20Act%20Statement.pdf), CMS cites the Social Security statutes that allow it to collect SSNs as well as describes the routine uses of the information, and specifically mentions that the information will be shared with the Census Bureau. However, the privacy concerns do

www.manaraa.com

add another layer of complexity and delay to the review and approval process, as it is important to find mutually beneficial applications of the linked data and to craft the justifications carefully for non-researchers in the approval chain.

For the Census Bureau, the issue of obtaining informed consent from survey respondents to link their data to administrative records was made easier with the enactment of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002. The Federal Policy for the Protection of Human Subjects (also known as the Common Rule) requires that research participants give informed consent to their participation. While interpretations have varied depending on the composition and outlook of Institutional Review Boards (IRBs), informed consent could mean that the survey respondent has to agree to any linkage of the survey data with administrative records. CIPSEA doesn't specifically exempt statistical agencies from the Common Rule requirements regarding informed consent. However, one of the interview participants suggested that the confidentiality protections afforded by CIPSEA would be a solid justification for seeking an IRB exemption. The Census Bureau has claimed an overall exemption from IRB review, and uses the DSEP Committee to approve its research. Eliminating IRB approvals greatly reduces potential bottlenecks in the review process.

Perceptual Dimension

The initial steps in sharing administrative record files involved exploratory work on both sides. The interviewees described how the Census Bureau devoted the first six months of file sharing efforts to examining the data files, in order to gain an

www.manaraa.com

understanding of what data were on the files, what the data really represented, and how best to use data to improve the Census Bureau surveys. For example, one of the issues that arose was determining how the records were attached to the payee, that is, whether people were really at the addresses listed in the files. The Census Bureau began receiving additional files from the Medicaid Statistical Information System (MSIS) in 1999. As mentioned earlier, the MSIS contains state submitted eligibility and claims data on Medicaid participants. States submit data to CMS on a quarterly basis through the MSIS.

During this initial period, the process of sharing records did not involve multiple levels of review at either CMS or the Census Bureau, according to interview participants. The Medicaid Research Systems Manager in the CMS Division of Research on State Programs and Special Populations would look at requests on a case-by-case basis, using the Privacy Act as a guideline for both internal and external requests for data.

However, enactment of HIPAA had a spillover effect on administrative record sharing activities both with states and with Census. For example, prior to HIPAA, state Medicare agencies were submitting unencrypted data to CMS, primarily on cartridge tapes that could not be encrypted. There were some isolated incidents described by the interviewees where tapes were lost in transit, although these did not receive much publicity at the time. CMS investigated these incidents and instituted a "locked box" system to try to make the transiting of cartridges more secure. However, this system was unwieldy, and as technology progressed, the CMS Privacy Board was able to require encryption of data submitted on CDs.

In addition, at the distribution end, CMS increased the layers of internal and external security surrounding data shared with researchers. The Privacy Board began

www.manaraa.com

requiring research protocols to be submitted for approval. It then reviewed research project requests for data. The Board wanted to assure that approved projects constituted legitimate research, which is an allowable "routine use" of data, as well as to assure that the research would be useful, benefit the public, and not be duplicative of projects already underway. Once a project was approved by the Policy Board, researchers were permitted to receive data. At the end of the project, the researcher was required to certify that the data files had been either returned or destroyed. Beginning in 2007, the Privacy Board began applying "minimum data required" criteria to projects, which excludes provision of records to researchers that are not deemed necessary.

Additionally, the interviewees stated that there are many repeat users of CMS research data within the research community. Although there is a wide audience for health research, the CMS research community is relatively small. After initial approval of a project, a researcher may come back with a reuse request to use the data for a new project with new funding. However, the Board's policy was to approve one project "use" at a time, in order to give CMS more control over the projects and reduce disclosure risks. This has resulted in pressure from the researchers to allow larger data releases.

While the new CMS procedures for project reviews tightened up the process for approving record sharing with Census, the interviewees described an environment that has remained generally a friendly one. A significant contributor to this may be the perception at CMS that it will gain benefits from joint data products that may be developed. Although, CMS cannot use improved data (such as those with verified SSNs) for program administration and enforcement activities, there are enough research uses for these data that the researchers, at least, view record sharing opportunities favorably.

www.manaraa.com

CMS's favorable attitude carries over into how the records are tracked after sharing. That is, when CMS provides data to the Census Bureau, it doesn't continue to claim ownership. Unlike the IRS regulations and statutes that require IRS to trail comingled tax data forever, CMS does not audit the uses of its data by the Census Bureau. Once the overall data sharing process is approved, the Census Bureau is free to create new products and share data with researchers.

Organizational Dimension

As mentioned, the process for approving Census Bureau access to CMS data records changed after enactment of HIPAA and no longer covered only CMS researchers and system managers. However, according to the interview participants, when the CMS Privacy Board reviews the Census Bureau requests, it is not looking at *whether* to provide data, but *how* to provide data. For example, the interviewees described that in 2006, a Census Bureau data request was refused. The reason was not because of some objection to the use of data at Census, but because data that were requested weren't on encrypted tape cartridges. The tapes could not be transferred securely and the cost of encryption was too high. Eventually, an alternative arrangement was negotiated whereby the Census Bureau was able to get the data it wanted from an MSIS file extract, and the work went forward. In 2008, CMS agreed to grant Census Bureau access to Medicare Analytic eXtract (MAX) data (consisting of person-level data on Medicaid eligibility, service utilization, and payments) to be used for a number of projects, supplemented by MSIS data that were more current than the 2004 data on the MAX file. Because technology had by then advanced enough that the tapes could be encrypted, CMS agreed to provide the MAX and MSIS data to Census. In exchange for CMS providing the files, Census

www.manaraa.com

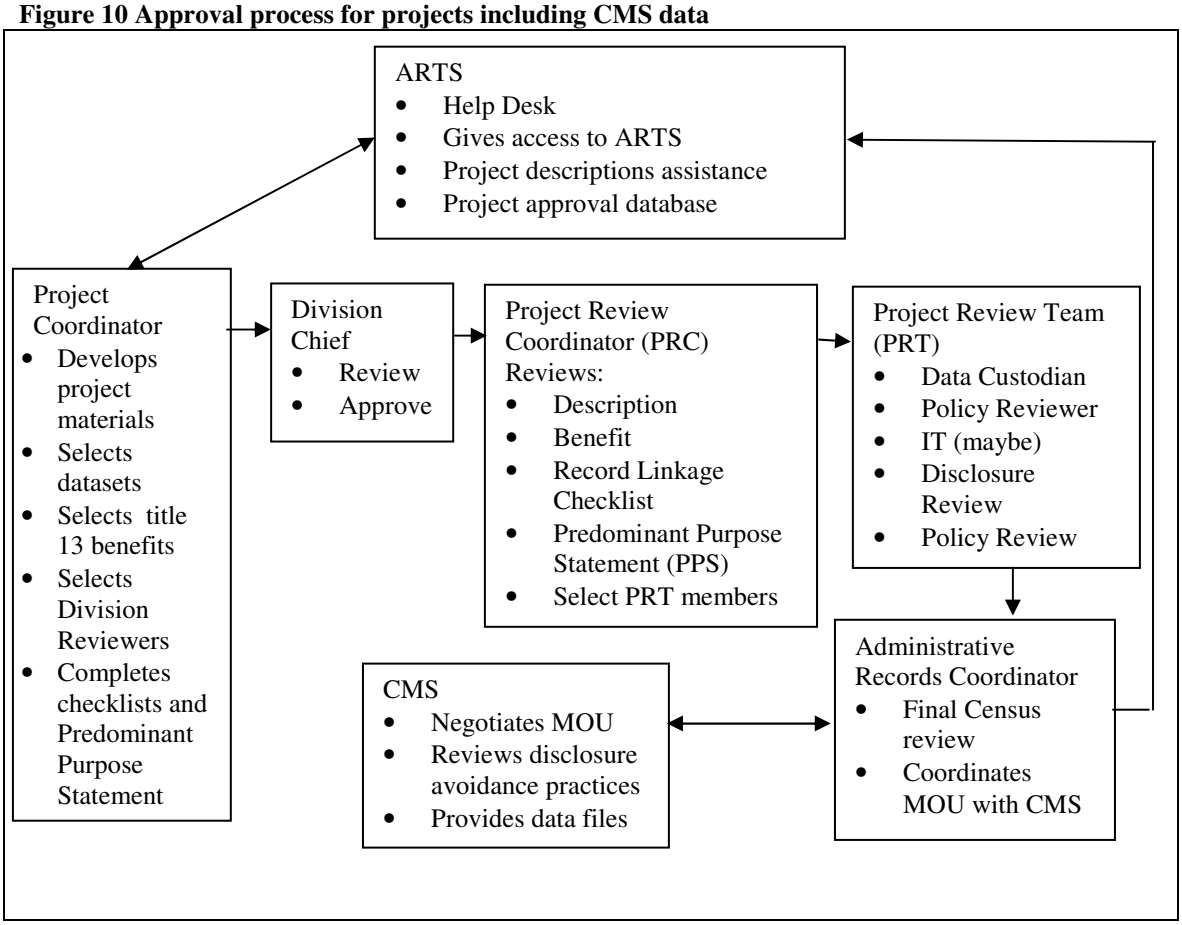
validated the quality of the SSNs that were in data provided by states and returned aggregate information to CMS.

Exchanges between CMS and the Census Bureau are governed by Memoranda of Understanding (MOUs). These are crafted for each separate agreement where files are exchanged. In addition to specifying what files will be provided, the legal basis for sharing files, how files will be provided and handled, who will have access to the data, and the responsible authorities for executing the agreement, the technical privacy protection provisions are also spelled out in the MOUs. The MOUs require that data be securely stored, have a retention date, and that a certificate of destruction be provided at the end of the project. If data are kept beyond the original retention date, extensions have to go through the approval process. Negotiation of MOUs for provision of new files can be a lengthy and exhausting process as described during the interviews. This will be explored more in depth later in this chapter

The Census Bureau's internal process for approving projects using CMS records has many similarities to the process used for projects that include FTI. However, CMS does not require that it approve all individual projects in which CMS data are used. Rather, CMS agrees to provide certain data files with specific variables and negotiates an MOU for that file. Once the file is at the Census Bureau, it is up to the bureau to determine on which projects the files will be used. Figure 10 below was developed with input from the interviewees and subsequently validated in the re-interviews. The figure shows the process within Census for approving administrative records projects that include CMS data.

www.manaraa.com

Figure 10 Approval process for projects including CMS data



Technical Dimension

The technical capability of the administrative records staff at Census has grown considerably since the mid-1990s. Large numbers of records from multiple sources may be linked. The Census Bureau has developed two processes that enable individual identities to be protected, even as it uses SSNs to link records. The first of these processes is the Person Identification Validation System (PVS), which is described in more detail in the IRS case study. This process uses probabilistic matching to verify SSNs in survey or other data files against the NUMIDENT. The other process that is key to safeguarding personal identities is assigning a Personal Identification Key (PIK) that

www.manaraa.com

replaces the SSN in the individual data records. Every SSN receives a PIK. This ability is critical to the success of protecting confidentiality when records are integrated from multiple sources.

An example of a project with significant technical complexity was a four phase project that was undertaken jointly by the Census Bureau, CMS, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) at the Department of Health and Human Services (DHHS), and the State Health Access Data Assistance Center (SHADAC) at the University of Minnesota. This project used several shared files and offered the opportunity to develop important technical expertise in linking records. The purpose of the research was to determine why there is an undercount of people in the Census Bureau's Current Population Survey (CPS) who are insured through Medicaid, with the overall goal of developing guidelines for researchers and analysts using survey data to conduct policy analysis (Davern, 2006). The joint project compared the numbers of people in the MSIS file who were enrolled in Medicaid with the number of people in the CPS who reported being on Medicaid, in order to assess the gap and then try to ascertain the extent of errors and their possible source.

Davern (2006) describes how during the first phase of the project, a massive database was constructed consisting of national health-insurance enrollment. The CMS MSIS files were merged with the CMS Medicare Enrollment Database (EDB) files using SSNs (Davern, 2007). The SSNs first were verified by the Census Bureau using the PVS, and then a separate process was then run to substitute PIKs for the SSNs in order to protect privacy. Then two Census Bureau files were used to assess the quality of the merged enrollment database. These files were the Master Address File Auxiliary

www.manaraa.com

Reference File (MAFARF) and the Person Characteristic File (PCF). The MAFARF is a Census Bureau file that contains a Master Address File Identifier (MAFID) and a PIK for each individual in the file. The MAFID is derived from the Census Master Address File (MAF) that contains an up-to-date inventory of all known living quarters in the United States and Puerto Rico. The PCF holds basic descriptive data for each person with an SSN, including race and gender (sometimes imputed).

The project used data from the CMS-provided MAX and MSIS files, as well as the Census Bureau-provided CPS, PCF, and MAFARF data files, with the purpose of matching CPS respondents to MSIS data for CY 2000 by SSN to examine survey reporting accuracy and to begin to understand the measurement error. All of the files were needed because if CPS and MSIS data didn't match, records were supplemented with information from MAX, MARARF, and PCF to try to gain a match.

The next phase of the project matched the state frame, household, and person data to the CPS using the state MSIS files, CPS, the CPS 2001 Supplemental Survey, MAX, and the MAF. Finally, the researchers tried matching data from the National Health Interview Survey (NHIS) instead of the CPS, in order to compare the data from the two surveys and begin to understand how survey design and implementation affect the quality of survey data. For that part of the project, the National Center for Health Statistics (NCHS) sent files to the Census Bureau, which then ran the files against the PVS to provide SSN links and then assigned PIKs to substitute for the SSNs. However, according to the interview participants, the project was stopped because NCHS wanted the linked files back, and the Census Bureau would not provide them due to Title 13 considerations. The project is still on hold.

www.manaraa.com

The project highlighted a number of issues. One was the quality of the data in both the surveys and in the Medicaid records. Census found duplicate SSNs that appeared in multiple states, as well as MSIS enrollees not enrolled in full benefits (unless they were pregnancy related). Of the 38.8 million Medicaid enrollees left in the MSIS file, 9 percent of the MSIS records did not have a verifiable SSN, 6.1 percent of CPS respondents had not provided a valid SSN, and 21.5 percent of the CPS respondents had refused to give their SSN. In the end, 12, 341 CPS respondents matched into the MSIS by SSN (Davern, 2007). There were errors in both the Medicaid records and erroneous response in the CPS data given by respondents. That is, people responding to the CPS question on whether they had health insurance sometimes responded that they did not, even when the Medicaid records showed they had coverage. There were also errors in the Medicaid records, particularly regarding SSNs.

One issue that remains outstanding for researchers is that while there is a large body of work on how to measure survey error, there is not much work that has been done regarding errors in administrative records. Although this may not directly affect privacy and confidentiality, it is a cause for concern if administrative record linkage continues to foster additional public policy research.

Another program that uses CMS records and has been able to benefit from the technological advances developed by the Census Bureau is the Small Area Health Insurance Estimates (SAHIE) program, which develops model-based estimates of health insurance coverage for counties and states (Census, 2009c). Data are obtained from a variety of sources, including the Annual Social and Economic Supplement of the Current

www.manaraa.com

Population Survey, County Business Patterns, Demographic Population Estimates, Federal tax returns, Food Stamp participation, Census 2000, Medicaid participation, and State Children's Health Insurance Program (SCHIP) participation . CMS provides Medicaid participation records from the MSIS. The SCHIP participation data are from CMS state counts. The Medicaid and SCHIP participation data are collapsed into one variable for use in the model. The data are aggregated to the state and county levels by age and sex for the estimates. In contrast to the SHADAC-led project comparing individual responses and records on Medicaid coverage, no personal identifiers are included on the records used by the SAHIE program.

However, the estimation models require significant data manipulation to produce estimates. In the case of counts of individual participants in the Medicaid, some counties have missing data and some counties have data that don't make sense. In either of these cases the Medicaid count is rejected, and a statistical match is used to replace the rejected Medicaid counts. State counts of SCHIP participation are used because data are not available at the county level for most of the counties. MSIS data are used to determine a ratio for allocating the adjusted SCHIP numbers to the county by sex. Again the quality of data in the records, this time provided by the states, has a significant effect on the ability to measure the number of uninsured in small areas, causing greater reliance on statistical methods to fill in the gaps.

The Human Dimension

CMS gains considerably by sharing data with the Census Bureau. However, as pointed out by the interview participants, because CMS is not covered by Title 13, it has

www.manaraa.com

not been possible so far for the Census Bureau to return verified person-level records back to CMS. CMS would like to find ways to make that happen, perhaps by designating a sub-part of CMS as a statistical unit. At present, there is no systematic verification of SSNs in the MSIS (and derived MAX data). However, improved identification of unique persons and better linkage of data across systems depends, in part, on CMS's ability to verify a person's identity. CMS does receive back cleaned and validated aggregate records.

While CMS does not have stringent requirements comparable to IRS Title 26, there are still many hurdles to pass in its review processes. These hurdles are placed by individuals in the agency, rather than by law or regulation. Negotiating MOUs can take years, sometimes due to the multiple levels of review, and sometimes due to one individual in the review process who can hold up agreements. For example, interviewees noted that it took 10 years of negotiation for SSA to agree to share the NUMIDENT with the Census Bureau. As a result, an individual who wants to initiate a major project faces a long involved process that takes sustained interest and attention over long periods of time.

At CMS, the slowdowns generally occur due to two situations described by the interviewees. The first is that the program administrators who "own" the Medicare and Medicaid data don't view record sharing and related research as their primary mission. Rather, their first priority is to operate their programs. It takes time and other resources to try to accommodate requests for data, and providing these may not be a high priority. Often, the person who is called upon to review requests is already overloaded with their

www.manaraa.com

own programmatic responsibilities. Thus data requests can languish for long periods, and the work can stall, unless there is an obvious and clear benefit to the data provider.

Second, there is a natural tension between researchers and the privacy community, which is concerned that as more records are linked, the opportunities for disclosure increase. This is particularly true in hot button areas such as medical records. As a result, the Privacy Officer at CMS sets very high standards that reflect a strict and narrow interpretation of the Privacy Act.

In addition, turf issues can arise between individuals at agencies. Although this has not been much of an issue between CMS and Census, it was certainly an issue with the Treasury Department and the Census Bureau's desire to expand its ability to conduct analyses using tax data.

Summary

The changed environment at CMS following passage of HIPAA affected the processes and procedures regarding the sharing of administrative record data with the Census Bureau. When CMS revamped its approval processes for outside research projects, it also revamped the procedures for approving Census Bureau projects. The newly established Privacy Board took a larger role in approving how files were to be shared with the Census Bureau. At the same time, the Census Bureau restructured its own internal review processes, primarily as a result of the 1999 IRS Safeguard Review. These new processes were also applied to projects using data from CMS. As a result of these two significant events, data sharing between the agencies became more formalized, and the approval time for projects lengthened. Although the additional organizational

www.manaraa.com

structure at both agencies made it easier to track projects, what is not as clear is whether the process also became more secure from a data disclosure standpoint.

In particular, the key process used to safeguard personal identities in the data files, namely using the Personal Identifier Key or PIK, was not materially affected by either passage of HIPAA or creation of the DSEP Committee. The Census Bureau did not develop the PIK process in response to either CMS rules or IRS pressure. Rather, it was an internal process developed to help the Census Bureau comply with its own confidentiality law, Title 13.

The use of Medicare and Medicaid records by the Census Bureau remains relatively unrestricted by CMS. Once specific variables in files are approved for sharing, CMS does not monitor the projects for which that information is used on an ongoing basis. However, getting new projects started that require additional variables can be a long and arduous process, primarily because of the reviews, and in particular, the process of negotiating the MOUs.

Summary of Case Study Findings

The study found that the IRS, CMS, and Census have set up elaborate systems to control the life cycle of combined data sets. These systems exist within the larger federal framework governing IT security, as well as privacy and confidentiality laws and regulations. The three organizations interact regularly with each other and with other agencies such as the federal Office of Management and Budget (OMB) and the Social Security Administration (SSA) to (1) design combined projects, (2) receive required approvals of the projects from various involved parties, (3) share data files, (4) combine, manipulate, and safeguard data, and (5) share specific portions of combined data with

www.manaraa.com

each other and specific designated parties. Often, the design and approval of projects can take years, even when projects are relatively simple to execute.

This rest of this summary of the case study findings is organized according to two questions. What are the main challenges for additional administrative record sharing among federal agencies? What administrative policy recommendations emerge from the case studies of IRS, CMS and the Census Bureau? Table 4 highlights the key findings of the study when viewed through the five dimensions used to structure the research. The discussion below of the two questions lays the groundwork for the conclusions and a future research agenda for record sharing.

Challenges for Additional Administrative Record Sharing

Despite the clear financial advantages and the opportunity to greatly improve the quality of research data, there are a number of challenges to moving ahead with additional projects that link administrative records with survey data or with other records. The process now in place at the Census Bureau through the RDCs clearly enables new projects to begin with existing file sharing agreements. But the initiation of major new projects, similar in scope to LEHD, faces significant hurdles. These are discussed below.

Advances in Technology

Advances in technology are a double-edged sword for record linkage. The utility of linking administrative record data is evident, and many more uses could be found. For example, technology enables the development and use of models that can use surveys to provide estimates at smaller geographic levels than is possible without the

www.manaraa.com

Table 4 Summary of findings by dimension

Legal	Perceptual	Organizational	Technical	Human
1. Each agency involved in sharing administrative records is governed by a different set of statutes and regulations that mostly do not overlap. . 2. This patchwork of laws and regulations greatly slows down the initiation of record sharing projects. 3. Agencies are quite protective of their statutory authorities and are reluctant to ask for legislative changes that might dilute their unique positions. 4. Elaborate interagency agreements govern the creation, terms of use, and destruction of shared administrative records	1. Participants at the agencies believe that, for the most part, privacy safeguards are adequate 2. Agencies have a sometimes strained relationship due to perceptions that other agencies are bottlenecks and roadblocks to more data sharing. 3. Maintaining public perceptions that personal data are safe when provided to the government is very important to the agencies. 4. Participants at the agencies expend significant effort to assure that data are protected as required by law and by interagency agreements.	1. Each agency has its own distinct internal processes for approving and tracking record sharing projects. 2. There are no mature government-wide shared processes or criteria for reviewing or approving projects involving multiple agencies. 3. The current processes are slow and burdensome and discourage initiation of new projects. 4. Data handling processes and data stewardship training are unique to each agency. 5. Agencies collecting administrative records don't have consistent quality control procedures. Poor quality of administrative record data is problematic.	1. Technology exists to successfully decouple personal identities from submitted data. Expertise for developing and using this technology resides primarily at the Census Bureau. 2. The technological methods used to safeguard the confidentiality of the data are effective at masking identities of respondents. 3. The technological safeguards assure that there is very little danger of confidentiality breaches when sharing micro-data between agencies. 4. Constant research is required in order to maintain the technological ability to foil those who would try to breach privacy and confidentiality electronically.	1. Data sharing projects are usually initiated by an individual or small group within an agency. There are no champions of the overall process. 2. The nature of the projects generally reflects the specific interests of the person generating the project, rather than fitting within a broader, previously determined research framework. 3. Due to the lengthy and involved approval process, initiation of major projects takes the sustained interest and attention of individuals over long periods. Mechanisms don't exist to reward these individuals 4. Turf battles between agencies, driven by individuals, affect the approval process.

www.manaraa.com

quality checks provided by the record linkages ((Lane, 2009; Obenski, 2006). Technology also enables the use of administrative record data to adjust survey weights, obtain characteristics of nonrespondents, and measure accuracy of survey data (Resnick & Obenski, 2006) .

On the other hand, advances in technology can also make people more fearful that private information may be misused and mishandled. During the last decade, privacy concerns have been heightened among the general public by incidents of identity theft, concerns around illegal immigrants, government homeland security activities, and some well publicized losses of equipment containing personal information by government employees in a variety of agencies.

Within the federal government the move to create IT efficiencies by consolidating computer resources also poses some problems for smaller statistical agencies. Records must be turned over to a centralized IT component that may be outside of the statistical agency. Assuring that the records are secure and not comingled with other information adds another layer of complexity that could be daunting for a small agency. If nothing else, sorting out the IT security and data handling procedures adds additional time to an already lengthy approval process.

Looming over these issues is how advances in technology affect the ability of unethical researchers and others to ascertain the personal identities of those whose records are linked. Although existing disclosure avoidance techniques are sophisticated, they must continue to evolve as computing power evolves. Ongoing research in disclosure

www.manaraa.com

avoidance is essential if record linkage is to remain secure, viable and able to protect data confidentiality (Doyle, Lane, Theeuwes, & Zayatz, 2001).

Administrative Record Quality

A significant part of the value of using administrative records as a substitute for survey data is the great cost savings. However, these savings are not worthwhile if the information in the records is not accurate or a lot of data are missing. One quality evaluation factor is coverage, that is, the number of survey respondents that can be linked to administrative records. Another factor is accuracy, which measures how closely administrative records content matches survey content. The large number of mismatches between the MSIS and CPS data on Medicaid reporting found in the SHADAC project indicates that there are a lot of errors in the MSIS as reported to CMS by states and a lot of misreporting on the CPS by respondents (Davern, 2007). Although the Medicare Modernization Act of 1998 requires states to deliver data to CMS, there are no incentives to send accurate and complete data and no enforcement of this by CMS. Exacerbating this is that one out of three people on Medicaid don't have SSNs (cite).

The Census Bureau links other administrative records data with survey data to try to improve quality. For example, one area that needed improvement was race data in StARS (Obenski, 2006). Because the race data had been captured from the NUMIDENT, it was deficient. Because SSA historically classified race as White, Black, or Other, there were gaps with the significantly different categories that the government currently uses to capture race information. In addition, SSA stopped its practice of collecting race data when children are born, so more recent entries on the NUMIDENT do not have race information. The Census Bureau began matching the linked CPS and NUMIDENT files

www.manaraa.com

with decennial census data, using a model to assign race and ethnicity if the data couldn't be matched.

The Census Bureau found that about 94% of CPS records could be linked to administrative records, including 50.2% for SSN verification, 36.4% in an address search, and 6.9% using a name search (Obenski, 2006) About 90% of the CPS and NUMIDENT records could be matched on age, although almost 7% were off by one year. That the data don't match is not surprising. Administrative data are collected and reported in a variety of ways by multiple parties and for multiple purposes. While survey data are collected in a more controlled manner, there can still be misunderstandings. As a result, measuring the amount of error in an integrated data set is very challenging.

The importance of recognizing errors in the records is illustrated by another study that was not included in the case studies. The National Longitudinal Mortality Study (NLMS) is a national, longitudinal, mortality study sponsored by the National Cancer Institute, the National Heart, Lung, and Blood Institute, the National Institute on Aging, the National Center for Health Statistics and the U.S. Census Bureau. It consists of a merged data base of files from the CPS, Annual Social and Economic Supplements and a subset of the 1980 Census combined with death certificate information to identify mortality status and cause of death. The study currently consists of approximately 3.0 million records with over 250,000 identified mortality cases (Census, 2009b). Results of the study seemed to show that Hispanics have higher survival rates than other ethnicities, even though they engage in activities such as smoking and have lower incomes, which tend to lower survival rates. However, researchers thought that the higher rates may be an effect of errors in SSNs.

www.manaraa.com

The challenge for improving the accuracy of administrative records is that there is not much of a tradition in place for correcting these errors. For the most part, agencies do not have processes in place for systematically assessing and correcting errors. Until means are in place to improve the quality of administrative records, particularly those supplied by multiple sources such as states, the value of record linkage will be limited, and significant additional work will be needed during projects just to ascertain the accuracy of the data. While research on data accuracy is a step forward, it is several degrees away from being able to use data to directly research public policy issues and create accurate models to assess the potential effect of proposed policies.

Bureaucratic Culture

While laws are necessary to protect privacy, and agencies need to enforce these protections, sometimes the predominant culture in government becomes a barrier to achieving advances. All of the people interviewed for this study agreed that several aspects of government bureaucracy are serious impediments to advancing administrative records research. One of the major barriers cited was the laborious process of negotiating MOUs, which can take anywhere from 8 months (at best) to 10 years (in one instance). The average time to negotiate an MOU involving the Census Bureau is one year (Obenski & Jones, 2007). Part of the problem with negotiating the MOUs is that there are multiple players in multiple agencies that all take time to review, comment, and negotiate on the MOU content. This includes attorneys, privacy advocates, and data providers. These organizational components are not always supportive of the goals of researchers.

In addition, the speed at which the responsible individuals attend to their review duties can depend on a lot of factors. As mentioned earlier, the research being proposed

www.manaraa.com

is not always viewed as part of the core mission of an agency that is administering a program such as Medicare. Reviews command resources, which may be scarce and directed towards other activities that are deemed a higher priority for an agency than helping the Census Bureau assess survey accuracy and coverage. The larger issues may get lost in the crush of daily demands, and result in MOUs being put aside for long periods.

In addition, turf issues may contribute to the sluggishness of the approval system. Agencies have a natural tendency to protect "their" information and may be reluctant to undertake all the additional work needed to share records simply to provide benefit to another agency that may garner recognition for its work. Thus projects need to be developed that provide mutual benefit to the agencies involved. This can be tricky, particularly in instances where the statistical agency needs to be certain that any shared data will not be used for law enforcement or other such purposes. However, creative approaches have been developed that allow benefits to accrue to data providers. For example, in the case of LEHD, when state data were brought into the Census Bureau, two files were created. One file contained only Title 15 data, and the other contained the Title 13 data. The Title 15 data were cleaned up and returned to the states. The Title 13 data were kept on a different server and were not returned to the states. Another approach used by the Census Bureau has been to use file extracts. For example, the Economic Directorate at the Census Bureau comingles data from the quarterly Financial Report and the Business Register. Because the Business Register contains FTI, IRS does not allow the Census Bureau to directly comingle the data for further dissemination and research. However, so-called "pure" Title 13 data on the Business Register can be extracted and

www.manaraa.com

comingled with the Quarterly Financial Report. This provides the degree of separation that will pass muster with the IRS approvers.

The current system is set up to encourage bureaucratic organizations to say no to record sharing at every point along the way to initiating a project. There is varied institutionalization of the approval process, including the resources needed to move proposals through an expedited process. Success of new proposals often hinges on personal relationships and the commitment of specific individuals to see a project through to fruition. When individuals change jobs, projects can die. What is missing is consistent top level support from agency leaders that includes providing the technical, legal, budget, and other support functions needed to get projects moving ahead. Rather, individuals with an interest in the research are forced to use their own networks and personal relationships to move new projects forward.

In addition, there is insufficient recognition and encouragement for those individuals who devote massive amounts of time and effort to creating innovative projects. Funding is scarce for new projects, and there are no formal mechanisms in place that provide widespread visibility and rewards within the research community. Researchers inside and outside of government need to be highly motivated over a sustained period of time, with very little institutional support in order to initiate major new research projects involving combined data sets.

Legal Barriers

Because many of the laws and regulations are subject to interpretation, the role of the individual is critical. In addition, the patchwork of laws governing each agency

www.manaraa.com

requires lengthy reviews and negotiations to overcome the various levels of protections and uses surrounding shared data.

For example, the National Center for Health Statistics had proposed that the Census Bureau use the PVS for the whole federal statistical system. This would have greatly facilitated record sharing. The barrier, however was that data in the PVS are covered by Title 13. While many agencies were supportive of the proposal in principle (including the Department of Commerce and SSA), a large amount of energy and resources continue to be devoted to figuring out a method for making this a reality.

Although CIPSEA addressed some of the barriers to data sharing, it did not go far enough. By setting a floor for protecting confidentiality equivalent to Title 13, CIPSEA enables dating sharing among the 14 recognized statistical agencies on an equal playing field. However, it did not address administrative records from non-statistical agencies. Further, title B of CIPSEA, which allows sharing of business data between the Census Bureau, BEA and BLS, is very narrowly crafted. BLS still can only get access to data containing FTI for research, and can't use that information to improve its operations. However, there is strong institutional resistance at Treasury to further opening up access to tax data.

Although many new joint projects have been initiated as a result of CIPSEA, some questions remain. For example, if data that were collected before enactment of CIPSEA are combined with data collected after enactment of CIPSEA, is the new combined data set completely covered under CIPSEA protections? These criteria are still being developed by OMB.

www.manaraa.com

Recommendations for Administrative Record Sharing Among Federal Agencies

Several recommendations for administrative record sharing among federal agencies emerged from the case studies and the literature. The recommendations were consistent across the relevant literature and the interviews. A consistent government-wide approach is needed to help break down barriers to administrative record sharing and linkages. The components of this approach are shown below, organized by dimension.

Legal:

1. Amend CIPSEA or move beyond it to new legislation that would enable more data sharing among statistical agencies.

Perceptual:

2. Gain consensus from the leaders of agencies that there is a substantial business reason for agencies to pursue record sharing and that all agencies ultimately benefit from better data on which to base public policy.

Organizational:

3. Undertake systematic reviews of the quality of data contained in administrative records. This should be part of the program performance evaluation process at agencies administering programs that rely on administrative records.
4. Direct sufficient resources to continue research on data disclosure, as well as continue to strengthen agency IT security and data stewardship policies and training for employees handling records with personal information. Make this research more explicit in agency budgets to gain support.

www.manaraa.com

5. At the working level, establish government-wide successful practices, and templates that can be used for developing MOUs and new project proposals so that each effort does not have to "reinvent the wheel". This effort is already underway in a subcommittee of the Federal Committee on Statistical Methodology at OMB.

Technical:

6. Develop a more efficient government-wide approach to some of the technical solutions for addressing privacy and confidentiality issues, such as moving ahead to enable the Census Bureau to provide some government-wide services in the areas of validating and substituting for personal identifiers such as SSNs.

Human:

7. Appoint champions throughout government that can move the record sharing process forward and eliminate bottlenecks.
8. Expand use of research fellowships such as through the American Statistical Association and the National Science foundation to promote administrative records research.
9. Establish formal recognition mechanisms for providing recognition and financial incentives to the "heroes" of new research projects that make major contributions to public policy research.

Summary

The major dimensions examined in these two case studies were legal, perceptual, organizational, technical, and human. The discussion illustrates the successes and challenges within each of these dimensions. The case studies illustrate both the

www.manaraa.com

complicated nature and the benefits of linking administrative records with survey data to advance public policy goals. Experiences suggest that protecting the privacy and confidentiality of the individuals whose information is being linked can be done very successfully. The laws and regulations governing the safeguarding of personal information have driven many of the processes and procedures that are in place, and their goals have been achieved. In order to move forward, it may be necessary to reexamine some of these laws and regulations to achieve a more consistent approach government-wide.

www.manaraa.com

Chapter 5: Conclusions

Introduction

The purpose of this chapter is to review the case study findings from a broader perspective. This chapter incorporates the case study findings presented in Chapter 4, which included a detailed analysis of the case study data, and builds on the ideas in that chapter. In the first section of Chapter 5, we raise two additional related questions that are intended to bridge the case study findings to a future research agenda. In the second section of Chapter 5, we present suggestions for a future research agenda, including two questions that arose out of the study's findings, and discuss related research strategies that have emerged from this study. The third section of Chapter 5 includes a brief summary of other related questions that could be further studied. Finally, in the fourth section of Chapter 5, the implications of the study for public policy and public administration are presented.

Reflections on the Research Questions

The case study analyses focused on answering the research questions posited in the study. The case studies examined the life cycle flow of administrative records data between IRS, CMS, and the Census Bureau. They also identified the significant issues that have arisen as a result of sharing administrative records, examined through the legal, perceptual, organizational, technology and human dimensions. Another objective of the study was to identify insights and potential solutions that could be learned from the experiences of those who have worked within the federal statistical system that would help address the significant data sharing issues that were identified.

www.manaraa.com

1. Should administrative records sharing and data linkage be organized at a government-wide level?

There is strong evidence that a government-wide approach to administrative record sharing would provide significant benefits. Ongoing research has demonstrated that integrated data sets containing both administrative and survey data can be a cost effective approach to addressing major policy issues, as illustrated by LEHD (Lane, 2009). Whether data contain information on individuals or businesses, or the issues are related to health, the economy, or other pressing areas of concern to public policy, the benefits are evident. In addition, ongoing projects have demonstrated that technological solutions are available that can provide the privacy and confidentiality protections that need to be afforded to data providers.

Further, there are three domains that should be recognized that have a stake in public policy research. These include agencies in the executive branch, organizations within the legislative branch (such as the Congressional Budget Office), and outside researchers. A government-wide approach that takes all these domains into consideration, rather than an exclusive focus on the executive branch, would garner support from a broader constituency, particularly if legislation is required.

Additionally, there are increasing calls for evidence-based policy in addressing the nation's problems. Pressure on the government to solve widespread problems that cut across traditional agency lines continues to increase. Increased attention to the important role of combined data sets in advancing public policy research is needed. National health insurance, unemployment, economic recovery, and disaster response all require

www.manaraa.com

government activity that transcends one single agency. In order for the government to develop policies that are responsive to the problems at hand, the best possible information is needed, which is often found in the combined data sets.

A government-wide approach does not necessitate developing a single enormous data base of all information collected by the government, whether it is to administer programs or in surveys and censuses. Although studies could show that approach to be cost beneficial, it is not likely to be palatable to many of the statistical system participants. Neither does it follow that sharing data leads to centralized program management and planning across the board. Rather, an effort to build on the current approach could yield the most benefit for the least cost. That is, agencies should be held accountable for the quality and consistency of the data they collect to administer programs. They should continue to own that data. However, consistency is needed in the criteria used to determine how and in what form data should be shared and linked to other data owned by other agencies. The current haphazard approach is too dependent on the commitment and whims of individuals who can advance or hold up important research at multiple decision points.

The Federal Committee on Statistical Methodology is positioned to spearhead the effort to develop a government-wide approach to record linkage and data integration. The existing Statistical Uses of Administrative Records Subcommittee is just a beginning to what should be an effort that engages agency heads at the highest level. The subcommittee could establish criteria to evaluate worthwhile projects, develop templates for MOUs, and provide other related support as necessary. But more needs to be done,

www.manaraa.com

particularly because the individuals engaged in these activities are already those most committed to advancing their individual projects.

The case studies highlight that there are currently significant barriers to linking various data sets residing in separate agencies that have the potential to provide rich, integrated data sets. These barriers can only be overcome through a systematic, government-wide approach that pushes the existing boundaries beyond the current voluntary, turf oriented, personality-driven approach to which many agencies now subscribe.

2. Should administrative data linked to other records and survey data be delivered through a centralized structure?

Many economies of scale can be achieved by using a centralized delivery mechanism, rather than recreating duplicate capabilities throughout government. Because the Census Bureau has been at the forefront of much administrative records research, it has developed the most sophisticated capability in government to link records and provide privacy protection. From an efficiency standpoint, it makes sense to develop this capability at the Census Bureau, and allow it to become a government-wide service provider.

However, overcoming the fear and mistrust that exists between federal agencies will require political will to enable the creation of a centralized structure to deliver linked records. Buy-in from leadership would be a key element in accomplishing this. In addition, it would also be necessary to overcome the deep distrust of centralized government planning that exists both within government and among the public. It would be important to emphasize that efficient record sharing does not equate to establishing

www.manaraa.com

centralized planning for all government programs. Rather, it builds on the model for statistical agencies already in existence throughout Europe and in Canada.

The concept of creating a centralized statistical agency in the U.S. has been examined several times over the years and has been rejected as many times. It's possible that a centralized system may not work with the form of government in place in the U.S. As with many proposals to consolidate government functions, opposition has centered around turf. However, not having a centralized statistical agency does not preclude finding successful practices being used in other countries and adopting them to the governmental system of the U.S. Putting efficient processes in place for centralized record linkage is a far cry from creating a centralized statistical agency. Rather a network structure that crosses jurisdictional boundaries would be created to deal with the ongoing policy issues that could be researched and analyzed using integrated data sets. The Census Bureau may have survey or census data to contribute to the evolving policy issues facing the nation. On the other hand, it may simply have the expertise to link records and provide data disclosure services to other agencies undertaking such analysis. This would not be unlike central services provided by agencies in the administrative arena for payroll, purchasing and other activities that can best be accomplished within a consistent framework. Of course, it would be essential to have a clear separation of statistical research activities and program enforcement activities within agencies in order to provide adequate privacy and confidentiality protections for linked data sets.

www.manaraa.com

A Research Agenda for Administrative Records Linkage

This section presents two research questions and related research strategies that have emerged from this study. There is also a brief summary of other related questions that could be further studied.

1. What are the policies, processes, and procedures followed in other countries with regard to integrating administrative records and survey data? How do they compare to the system in place in the United States?

As mentioned in Chapter 2, Canada, the United Kingdom (UK) and Australia, among others have centralized statistical bureaus, rather than the fragmented statistical system found in the U.S. The centralized approach eases the way for sharing of administrative records between operational agencies and the statistical bureaus, because the context in which the records will be used is clear; that is, for statistical purposes.

Statistics Canada, for example, collaborates with other government departments to collect information, including statistics derived from the activities of those departments. Part of the mandate of Statistics Canada is to assure that there is not duplication in the information collected by the government. Record linkage is an important technique used by Statistics Canada to develop and analyze data. Under its policy, strict criteria are followed for pursuing record linkages, and a website is maintained that lists all linked databases. The analytic results of studies involving linked records are placed in the public domain and are accessible to the public.

The U.K. enacted the Statistics and Registration Service Act of 2007, which restructured its statistical system, creating the Statistics Board. This entity is independent

www.manaraa.com

of the Executive Branch of the government and reports directly to Parliament. The Board has powers to produce statistics, provide statistical services and promote statistical research, including the preparation and publication of the census. While the Act is still relatively new, it should lead to more data sharing between the now centralized statistical agency and other government agencies.

The Australian Bureau of Statistics (ABS) is a centralized agency established by the Census and Statistics Act of 1905 as amended. Its underlying approach is to use social science statistics to measure the well being of the nation. Based on guidelines proposed by the Organization for Economic Co-operation and Development (OECD) that wellbeing could be effectively measured using key indicators, such as good health, sufficient income, and rewarding work, the ABS measures health, family and community, education and training, work, economic resources, housing, crime and justice, and culture and leisure.(ABS, 2008).

Studying the record sharing approaches of these and other international agencies could help guide necessary changes in the U.S. statistical system. Although it is not likely that the U.S. system will be consolidated into a centralized statistical agency, there are still functions that might be centralized, such as record linkage. Such a study would look at the benefits and costs of adopting successful practices from other countries in our ore decentralized system and contribute to the body of knowledge about approaches to safeguarding privacy, new technologies, and best practices that could be applicable in the U.S.

www.manaraa.com

2. How might inconsistent laws and regulations be changed or superseded in order to advance the ability of agencies to share data?

Although the legal barriers to sharing administrative records can be and in fact, on occasion have been overcome, they are still an impediment to growing the body of public policy research that potentially could be undertaken. Some of the existing obstacles arise from strict interpretations of laws. Others obstacles stem from unwillingness on the part of organizations or individuals within those organizations to develop regulations that would allow easier access to data. In these instances, the law acts as a shield to stand behind in order to justify behaviors motivated by other factors, such as protecting turf.

What is called for is an objective, in-depth look at the panoply of legislation and implementing regulations governing privacy, confidentiality, and record sharing. That is not to say that these have not been identified. For those working in the administrative records field, and for many individuals at agencies involved in record sharing, the statutes and regulations themselves are well known. Missing, however, is an analysis of how changes in existing laws might enable more data integration without compromising the essential missions of the programs and agencies involved.

The IRS – Census Bureau case study illustrated that a major impediment to speedier approval of projects involving FTI was the Treasury/IRS interpretation of Title 26. That strict interpretation is designed to protect the privacy of taxpayers, but it also serves the dual purpose of protecting the interests of IRS and Treasury. It would be hard to argue that taxpayers are either more or less at risk because the IRS is reviewing proposed RDC research projects to assure that the correct percentage of the research is aimed at improving the work of the Census Bureau. Along the same lines, one might

www.manaraa.com

question why it was originally declared not in the interest of taxpayers to have the IRS share W2 data with the Census Bureau in order to move the LEHD project forward, yet three years after the project began, it was declared to be in the interest of taxpayers. Why did it take 10 years to negotiate an MOU between the Census Bureau and SSA to share the NUMIDENT file? It is doubtful that during that period something occurred that radically changed the effect on people holding SSNs.

Many people involved in the agencies that provide administrative records are concerned that there will be unintended consequences of sharing data with statistical agencies. Even with CIPSEA protections, there are unknowns. And because the public perception of how the government is safeguarding its information is so important, agencies do not want to make missteps. Even those who support administrative records research may be in favor of a conservative approach that is designed to avoid controversy.

Research that is designed to separate legitimate programmatic needs from human motivations is needed to identify how legislative mandates might be harmonized in the statistical data sharing arena. What is really required to achieve the goal of safeguarding information that is being used to conduct statistical research? What should the next generation of CIPSEA legislation contain? How can the government data owners and custodians make sure that data are being used for appropriate research purposes without causing costly and unnecessary delays in projects that may provide information that would substantially inform public policy and improve program administration? Which laws would need to be retained in order to assure that linked records are not used for law enforcement or other nonstatistical research purposes?

www.manaraa.com

The government is moving ahead rapidly to mandate the use of electronic health records among a broad community of users. Vast amounts of personal health data are currently traveling electronically between providers, hospital systems, insurers, and a multitude of government agencies including the Veterans Administration, the Department of Defense, and SSA. More sharing of electronic health records is mandated for the future. HIPAA is considered adequate protection for these records. Additional research could compare and contrast the legislative underpinnings and consequences of the medical community sharing individual health records on the one hand, and on the other hand, the Census Bureau, CMS, and health researchers sharing data files that have been cleaned and all identifiers removed in order to conduct policy research.

Resources for administrative records research are limited. By diverting these resources towards lengthy negotiations and approval processes, designing and implementing the actual research projects is delayed. A major contribution could be made by sorting out the inconsistencies in legislation and focusing attention on what is actually required to operate programs, conduct research, and protect the confidentiality of the data.

Other Research Questions

Other issues raised by this study also merit additional attention and examination. One area is the role of state agencies in data sharing. For example, state unemployment data provided directly by states are a critical component of the LEHD project. However, agreements had to be negotiated individually with each state to obtain data, a time consuming process that has taken over 8 years and is not yet complete. States are also the source of data provided to CMS on Medicaid participants. Currently the states are not

www.manaraa.com

being held accountable for the quality of data that is put into the system. Efforts should be devoted to designing mechanisms to put in place that would provide incentives to states to be more vigilant about data quality. Improved timeliness of data supplied by states is also needed. In some instances, state data that are available through CMS are over four years old. Examining the role of states feeding information into the federal statistical system is an aspect that should not be overlooked.

Another area meriting attention is how agencies providing data might benefit more from sharing their records. Currently, Title 13 prohibits much microdata from being shared with other agencies. This is true even when agencies are paying the Census Bureau to conduct surveys on their behalf. While Title 13 protections are extremely important to the Census Bureau and are considered necessary in order to gain respondent cooperation, there may be more detailed data that the Census Bureau could provide to agencies contributing to integrated data sets that would still protect the confidentiality provisions of Title 13. For example, are there ways that CMS could be informed when birthdays, gender, SSN or other data are incorrect in its files? How would Census sharing information with the Statistics of Income Division in IRS affect public perception?

The various needs and perspectives among agencies that can benefit from participation in record linkage should be studied further. Because these perspectives are often in conflict, they exert a negative effect on the success of using integrated data sets to improve public policy. What strategies can be identified and implemented that will enable additional public policy research involving linked administrative records and survey data, successfully overcoming the human resistance in affected agencies?

www.manaraa.com

Implications for Public Policy and Administration

This section discusses two related trends affecting public policy and administration research that arose out of this study. First, technological capability in the private and public sectors will continue to create new opportunities for both enhancing research opportunities and for putting data containing personal information of individuals at risk of being disclosed to unauthorized parties through data linkages. Second, as pressure increases for the government to take a broader, more active role in regulating or managing activities such as health care, the financial sector, economic recovery, and disaster recovery, there will be a need for more and better data to measure the effectiveness of government actions, evaluate the outcomes of government programs, and research future policy and program administration approaches. These two trends represent ongoing challenges that will either be addressed piecemeal or in a cohesive fashion.

What are some of the technological developments that will affect record linkages? One example is the extensive work that is being done in the health care arena to assure that various electronic health record systems are interoperable or can communicate with each other. Technology has progressed beyond simply the ability to link records. Rather the focus is now on what intelligence can be gathered from the linked data. In the case of health records, the emphasis is on the concept of "meaningful use". Under the Health Information Technology (HITECH) Act, hospitals and healthcare providers must demonstrate meaningful use in order to qualify for CMS incentives that pay reimbursement for putting electronic health record systems in place. An advisory committee has been established by the Office of the National Coordinator for Health

www.manaraa.com

Information Technology to determine the definition of meaningful use. The advisory committee (which includes representative of SSA and CMS) recently came up with its first attempt at this definition which includes, among many other things, that by 2011, hospitals must have the capability to exchange key clinical information among providers of care, that lab-test results must be incorporated into electronic health records as structured data, and that patients must be provided with an e-copy of their health information (Health, 2009). As large databases are amassed that contain a wealth of information about individual patient outcomes, pressure will increase to conduct research to improve patient safety, help with diagnostics, and manage risk in other healthcare related areas. There will be a need to look to the federal government to help develop policies on record linkage and access to data for research.

Further, agencies such as SSA, VA, and DOD are taking a leading role in developing systems that allow active military personnel and veterans to have portable medical records that can follow them through the health system, particularly if they are disabled and need continuing care, are eligible for SSA disability benefits, and change their geographic location. These technological developments will certainly have an effect on other record linkage activities within the government for research purposes as well as program administration.

Other technological developments that will have an effect on public policy are the use of synthetic data developed from original data, which retains the meaning and relationships in original data, but are intended to protect the confidentiality of the respondents; advances in the use of disclosure avoidance techniques such as variable suppression, top- and bottom-coding, re-categorization, noise infusion, swapping, and

www.manaraa.com

geographic aggregation (Weinberg, Aboud, Rowland, Steel, & Zayatz, 2007). These techniques, developed in response to increased computing capability and enhanced ability to link open source records through internet access, reduce the amount of data available to researchers. On top of this, there is increasing pressure to make business microdata more readily available to researchers. The tensions between the needs of researchers and the need to protect data confidentiality will continue and will need to be addressed in the public policy arena.

The second set of emerging issues is related to the competing needs of researchers and privacy protection. As the need to measure and evaluate the effectiveness of government programs increases, so will the demand for high quality data. As the case studies showed, data quality can be greatly improved through properly designed studies that take advantage of record linkages. The proven success of these studies will fuel demand for more such projects. Continuing to address the start up of these studies on a case-by-case basis is inefficient and could be detrimental to effective program administration, as well as slowing the policy responses developed in response to critical national issues. Can the country really afford to spend 10 years negotiating a single MOU? How long can the nation wait to improve the key economic indicators because of the need to accommodate IRS and Treasury requirements?

These questions need to be addressed in a more systematic way than is being done currently. Although the case studies illustrate that many people throughout government are involved in administrative records research, there is still no overarching framework that is sufficient to address all the outstanding issues. OMB continues to serve in a coordination role rather than in a role with any power or leverage to change the behavior

www.manaraa.com

of agencies. Agencies are not compelled to share records except in the few instances required by law. As individuals struggle to initiate and maintain critically important research, the system as a whole is indifferent. More focus and attention to addressing the problems in the system is desperately needed if available data are to be utilized to conduct the analysis and research required to accurately measure the state of the nation.

The discussion in this section highlights just a few of the emerging issues surrounding administrative record sharing and linkage among federal agencies. The two case studies of the IRS and the Census Bureau and CMS and the Census Bureau have provided important insights into the challenges to conducting important research with linked microdata and integrated data sets while still protecting the privacy and confidentiality of the individuals who are providing the data. The case studies also lay the groundwork for continuing research to advance the ability of the federal agencies to gain access to high quality data that will inform future public policies.

www.manaraa.com

REFERENCES

A.J. Reiss, J. (1980). Victim proneness by type of crime in repeat victimization. In S. E. Fienberg & J. A.J. Reiss (Eds.), *Indicators of Crime and Criminal Justice: Quantitative Studies* (pp. 41-53). Washington, D.C.: U.S. Government Printing Office.

Abowd, J. M., Lane, J., & Haltiwanger, J. (2004). Integrated Longitudinal Employee-employer Data for the United States. *American Economic Review*.

ABS. (2008). 2008, from http://www.abs.gov.au/ausstats/abs@.nsf/Latestproducts/8AA31CAB719513A6CA2571B7000724DB?opendocument

American Statistical Association. (1977). Report of Ad Hoc Committee on Privacy and Confidentiality. *The American Statistician, 31*(2), 59-78.

Anderson, M. (1988). *The American Census: A Social History*. New Haven: Yale University Press.

Anderson, M., & Seltzer, W. (2004, August 2004). *The Challenges of ":Taxation, Investigation, and Regulation:" Statistical Confidentiality and U.S. Federal Statistics, 1910-1965.* Paper presented at the Census Bureau Symposium, Woodrow Wilson International Center for Scholars.

Barabba, V. (1975). The Right of Privacy and the Need to Know. In U.S. Census Bureau (Ed.), *The Census Bureau: A Numerator and Denominator for Measuring Change*. Washington, DC: Government Printing Office.

Barron, W. (2000). *Criteria for the Review and Approval of Census Projects that Use Federal Tax Information* (Memorandum). Washington, DC: US Census Bureau.

Barth, A., Datta, A., Mitchell, J. C., & Nissenbaum, H. (2006). *Privacy and Contectual Integrity: Framework and Applications.* Paper presented at the Proceedings of the 27th IEEE Symposium on Security and Privacy.

Bohme, F. G., & Pemberton, D. N. (1991, August 18-22, 1991). *Privacy and confidentiality in the US Censuses -- A History.* Paper presented at the Annual Meeting of the American Statistical Association, Atlanta, GA.

Burgess, R. G. (1984). *In the Field: An Introduction to Field Research*. London: Allen and Unwin.

Carroll, J. D., & Kerr, C. R. (1976). The APSA confidentiality in social science research project: a final report. *PS (The American Political Science Association, 9*(4), 416-419.

CDC, C. f. D. C. (2008, August 2008). *National Health Interview Survey (NHIS)*, from http://www.cdc.gov/nchs/data/nhis/earlyrelease/wireless200812.htm

Cecil, J. S., & Griffin, E. (1982). *The Role of Legal Policies in Data Sharing*. Paper presented at the Panel on Data Access for Research Purposes, Washington, D.C.

Census, B. (1997). *Guidelines for the Operation of the Census Bureau's Research Data Centers*. Retrieved June 26, 2009, from http://www.ces.census.gov/index.php/ces/cmsdownloads

www.manaraa.com

Census, B. (2000). Criteria for the Review and Approval of Census Projects that USe Federal Tax Information. In IRS (Ed.) (pp. Interagency Agreement). Washington, DC.
Census, B. (2002a). *Policy for Controlling Non-employee Accessto Title 13 Data Policy*. Washington DC.
Census, B. (2002b). *Policy on Negotiating Collaborative Arrangements with Agencies for the Acquisition of Administrative Record Data to Support title 13 Projects*. Washington, DC.
Census, B. (2005). *U.S. Census Bureau Policy Regarding Custom Data Tabulations*. Retrieved December 2008, from http://www.census.gov/privacy/DS-021_Custom_Tabs_signed_102005.pdf
Census, B. (2009a). *Business Register*. Retrieved July 3, 2009, 2009, from http://www.census.gov/econ/overview/mu0600.html
Census, B. (2009b). *Description of the National Logitudinal Mortality Study*. Retrieved July 13, 2009, 2009, from http://www.census.gov/nlms/projectDescription.html
Census, B. (2009c). *Small Area Health Insurance Estimates*. Retrieved July 2009, 2009, from http://www.census.gov//did/www/sahie/index.html
Census Bureau. (2006). *U.S. Census Bureau Privacy Principles*. Retrieved December 12, 2007, from http://www.census.gov/privacy/Privacy_Principles_signed_updated_040606_.doc
Clemetson, L. (2004, July 20, 2004). Homeland Security Given Data on Arab-Americans. *New York Times*.
CMS. (2002). Privacy Act of 1974; Report of Modified or Altered system (Vol. 67, pp. 6714-6715). Washington, DC: Federal Register.
CMS. (2007). *Active Projects Report: Research and Demonstrations in Health Care Financing*. Washington, D.C.: Department of Helath and Human Services; Centers for Medicare and Medicaid Services.
CMS. (2009). *CMS Mission Statement*. Retrieved july, 2009, from http://www.cms.hhs.gov/MissionVisionGoals/
Craig, L. (1997, April 15, 1997). *S. 522 - Taxpayer Browsing Protection Act*. Retrieved July 3, 2009, from http://rpc.senate.gov/releases/1997/BROWSING.CW.htm
Creswell, J. W., & Miller, D. L. (2000). Determining validity in qualitiative inquiry. *Theory into Practice, 39*(3), 124-130.
Davern, M. (2006). Why Are Survey Counts of Medicaid Enrollees Lower Than Administrative enrollment Counts? Seattle: State Health Access Data Assistance Center (SHADAC).
Davern, M. (2007). *Phase I Research Results: Overview of National Medicare and Medicaid Files*. Washington, DC: SHADAC.
de Leeuw, E., & de Heer, W. (2002). Trends in Household Survey Nonresponse: A Longitudinal and International Comparison. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 41-54). New York: Wiley.
Domingo-Ferrer, J., & Torra, V. (2001). A Quantitative Comparison of Disclosure Control Methods for Micro-data. In P. Doyle, J. Lane, J. Theeuwes & L. Zayatz (Eds.), *Confidentiality, Disclosure and Data Access* (pp. 111-133). Amsterdam: North-Holland.

www.manaraa.com

Doyle, P., Lane, J., Theeuwes, J., & Zayatz, L. (2001). *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*. Amsterdam: North-Holland.

DSD, D. S. D. (2007). *Survey Abstracts*. Washington, D.C.: U.S. Census Bureau.

Duncan, G. T., Jabine, T. B., & Wolf, V. A. d. (1993a). *Private Lives and Public Policies*. Washington, D.C.: Committee on National Statistics; National Research Council.

Duncan, G. T., Jabine, T. B., & Wolf, V. A. d. (1993b). *Private Lives and Public Policies: Confidentiality and Accessibility of Government Statistics*. Washington, DC: National Research Council.

Eckstein, H. (1975). Case Studies and Theory in Political Science. In F. Greenstein & N. Polsby (Eds.), *Handbook of Political Science* (Vol. 7, pp. 79-138). Reading, Mass: Addison-Wesley.

Economic Policy Council. (1991). *Report of the Working Group on the Quality of Economic Statistics*. Washington, D.C.: Economic Policy Council.

Eddy, W. F., Fienberg, S. E., & Griffin, D. L. (1981). Estimating victimization prevalence in a rotting panel survey. *Bulletin of the International Statistical Institute, 43*(2), 719-731.

EPIC. (2007). *EPIC Online Guide to Privacy Resources*. Retrieved December 30, 2007, 2007, from http://epic.org/privacy/privacy_resources_faq.html

EU. (2006). Directive 2006/24/EC of the European Parliament and of the Council of 15 March on the rentention of data generated or processed in connection with the provision of pulicly available electronic communications services or of public communications networks and amending Directive 2002/58/EC. On *Official Journal L.*

FCSM. (2005). *Report on Statistical Disclosure Limitation Methodology*. Washington, DC: US Office of Management and Budget.

Fienberg, S. E., Martin, M. E., & Straf, M. L. (Eds.). (1985). *Sharing Research Data*. Washington, D.C.: National Academy Press.

Fixler, D., & Landefeld, J. S. (2006). The Importance of Data Sharing to Consistent Macroeconomic Statistics. In C. a. c. m. Kuebler (Ed.), *Improving Business Statistics Through Interagency Data Sharing: Summary of a Workshop*. Washington, DC: National Academy of Sciences.

Friedman, L. S. (2002). *the Microeconomics of Public Policy Analysis*. Princeton: Princeton University Press.

GAO. (1992). *Tax systems Modernization:concerns Over Security and Privacy Elements of the Systems Architecture* (No.

GAO/IMTEC-92-63). Washington, DC.

GAO. (1993). *IRS Information Systems: Weaknesses Increase Risk of Fraud and Impair Reliability of Management Information* (No. GAO/AIMD-93-34). Washington, DC.

GAO. (1994). *Tax Administration: IRS Can Strengthen Its Efforts to See That Taxpayers Are Treated Properly* (No. GGD-95-14). Washington, D.C.

George, A. L., & Bennett, A. (2004). *Case Studies and Theory Development in the Social Sciences*. Cambridge: MIT Press.

www.manaraa.com

Gilheany, S. (2000, December 2, 2004). *The Decline of Magnetic Disk Storage Cost ofer the Next 25 Years*, from http://www.berghell.com/whitepapers/Storage%20Costs.pdf

Gostin, L. O. (2003, July 1, 2003). *Health Information: Privacy and Confidentiality.* Paper presented at the National Health Information Infrastructure 2003: Developing a National Action Agenda for NHII.

Gostin, L. O., & James G. Hodge, J. (2002). Personal Privacy and Common Goods: A Framework for Balancing Under the National Health Information Privacy Rule. *Minnesota Law Review, 86*, 1439.

Groves, R. M., & Couper, M. P. (1998). *Nonresponse in Household Interview Surveys*. New York: Wiley.

Health, C. I. (2009, July 17, 2009). *Meaningful Use Proposal Revealed*. Retrieved July 20, 2009, 2009, from http://www.healthcare-informatics.com/ME2/dirmod.asp?sid=&nm=&type=news&mod=News&mid=9A02E3B96F2A415ABC72CB5F516B4C10&tier=3&nid=27B7FE2AFE05411497EC4979A78E5161

Hsu, S. S. (2007, May 5, 2007). TSA Hard Drive With Employee data is Reported Stolen. *The Washington Post,* p. 9.

IRS. (1992). *Review of Controls OVer IDRS Security* (No. Internal Audit Reference No 030103). Washington, DC.

IRS. (2001). *IRS Strategic Plan Fiscal Year 2000-2005* (No. Publication 3744 2-2001). Washington, D.C.

IRS. (2006). *2006 IRS Research Conference*, from http://www.irs.gov/taxstats/productsandpubs/article/0,,id=151642,00.html

IRS. (2007a). *Privacy Impact Assessments*, from http://www.irs.gov/privacy/article/0,,id=122989,00.html

IRS. (2007b). *Publication 1075: Tax Information Security Guidelines for Federal, State and Local Agencies and Entities*. Washington, DC.

IRS. (2009). *The IRS Mission Statement*. Retrieved July, 2009, from http://www.irs.gov/irs/article/0,,id=98141,00.html

Lane, J. (2005). *Optimizing the Use of Micro-data: an Overview of the Issues.* Paper presented at the American Statistical Association Annual Meeting, Minneapolis, MN.

Lane, J. (2009). *Linking Administrative and Survey Data*. Washington, DC: National Science Foundation.

Lane, J., & Stephens, B. (2006). *Integrated Employer-Employee Data: New Resources for Regional Data Analysis*. Suitland: U.S. Census Bureau.

Lerner, M. (2002, Dec. 15, 2002). Private Records, Public Benefit. *Minneapolis Star-Tribune,* p. A1:A7.

Liiphart, A. (1971). Comparative Politics and the Comparative Method. *American Political Science Review, 65*(3), 682-693.

Mason, J. (2002). *Qualitative Interviewing* (Second ed.). Thousand Oaks: SAGE Publications.

McLaughlin, J. A., & Jordan, G. B. (2004). Using Logic Models. In J. S. Wholey, H. P. Hatry & K. E. Newcomer (Eds.), *Handbook of Practical Program Evaluation* (Second ed., pp. 7-32). San Francisco: Jossey-Bass.

www.manaraa.com

Nakashima, E. (2007, April 21, 2007). U.S. Exposed Personal Data. *The Washington Post,* p. 5.
Nissenbaum, H. (2004). Privacy as Contextual Integrity. *Washington Law review, 17*, 101-139.
NLIHC. (2006). Memo To Members: National Low Income Housig Coalition (NLIHC).
NRC, N. R. C. (2004). *The 2000 Census: Counting Under Adversity*. Washington, DC: The National Academies Press.
Obenski, S. (2006). Methods and Applications of Administrative Records Research. Washington, DC: US Census Bureau.
Obenski, S., & Jones, C. (2007). *Larger Statistical and Policy Implications of Moving Toward Interagency Collaborations and Integrated Data*. Washington, DC: US Census Bureau.
Obenski, S., & Prevost, R. (2004). *A Policy Application: Using Administrative Records to Supplement Census Bureau Programs*. Washington, D.C.: U.S. Census Bureau.
Office of Federal Statistical Policy and Standards. (1978). *A Framework for Planning U.S. Federal Statistics for the 1980's*. Washington, D.C: U.S. Department of Commerce.
OMB. (2006). Proposed Implementation Guidance for title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA). *Federal Register, 71*(199), 60772-60773.
OMB, U. S. (1997). Order Providing for the Confidentiality of Statistical Information (Vol. 62, pp. 35044-35049). Washington, D.C.: Federal Register.
OMB, U. S. (2006). *OMB Implementation Guidance for Title V of the E-Government Act, confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA)*, from www.omb.gov
Panel on Data Access for Research Purposes. (2005). Expnding Access to Research Data: reconciling Risks and Opportunities. In. Washington, D.C.: National Academies Press.
Patton, M. Q. (1990). *Qualitative Evaluation and Research Methods*. Newbury Park, CA: SAGE Publications.
Paulson, K. A. (2002). *State of the First Amendment 2002* (No. 02-F07). Arlington VA: First Amendment Center.
Portia Project. (2004). *Home Page*. Retrieved December, 2008, from http://crypto.stanford.edu/portia
Potok, N. (2002, October 9, 2002). *Building a Data Stewardship Framework at the U.S. Census Bureau.* Paper presented at the Interagency Council on Statistical Policy, Washington, D.C.
Potok, N., Bailey, R., Sherman, W., Harter, R., Yang, M., Chapline, J., et al. (2005). *The 2003 Survey of Small Business Finances Methodology Report* (No. OSS 3). Chicago: NORC.
Prevost, R. (2001). *Managing Statistical Information Systems with Commingled Survey and Administrative Data: A U.S. Census Bureau Example.* Paper presented at the Joint ECE/Eurostat Meeting on the Management of Statisitical Information Technology, Geneva, Switzerland.

www.manaraa.com

Privacy Protection Study Commission. (1977a). *Personal Privacy in an Information Society* (No. 052–003–00395–3). Washington, D.C.: U.S. Government Printing Office.

Resnick, D., & Obenski, S. (2006). *Methods and File Acquisitions Supporting the Expanded Use of Administrative Records.* Paper presented at the American Statistical Association Prceedings.

Romero, M. (2003). *Open Access and the Case for Public Good: The Scientists' Perspective*, July/August 2003, from www.infotoday.com/online/jul03/romero.shtml

Rubenstein, S. (2008, February 2009). *Grassley Probes Medical Ghostwriting by Wyeth*. Retrieved December 12, 2008, from http://blogs.wsj.com/health/2008/12/12/grassley-probes-medical-ghostwriting-by-wyeth/

Seltzer, W., & Anderson, M. (2007, March 29-31, 2007). *Census Confidentiality under the Second War Powers Act (1942-1947).* Paper presented at the Population Association of America Annual Meeting, New York, NY.

Setness, P. A. (2003). When Privacy and the Public Good Collide: Does the collection of health data for research harm individual patients? *Postgraduate Medicine online, 113*(5).

Singer, E. (2002). The Use of Incentives to Reduce Nonresponse in household Surveys. In R. M. Groves, D. A. Dillman, J. L. Eltinge & R. J. A. Little (Eds.), *Survey Nonresponse* (pp. 163-178). New York: Wiley.

Statistics Canada. (2000, July 26, 2006). *Record linkage Policy*. Retrieved January 2, 2008, from http://www.statcan.ca/english/recrdlink/policy4-1.htm

Statistics, U. O. o. N. (2000). *Framework for National Statistics*, from www.statistics.gov.uk/about/national_statistics/documentation.asp

Steele, P. M. (2005). *Disclosure Risk Assessment for Micro-data*. Suitland: U.S. Census Bureau.

Steeves, V. (2004, October 20,2004). *Will Changes in Data Health Privacy Legislation Kill Research as We Know It?* Paper presented at the 2004 Annual LaBelle Lectureship, Centre for Health Economics and Policy Analysis, McMaster University.

Sweeney, L. (1997). *Computational disclosure control for medical microdata: The datafly system.* Paper presented at the Record Linkages Techniques 1997: Proceedings of an International Workshop and Exposition, Washington, D.C.

Thomson, C. (2002, August 9, 2002). *Balancing the Needs of Researchers and the Individual's Right to Privacy Under the New Privacy Laws.* Paper presented at the Round Table 14 National Scholarly Communications Forum, Australia.

TIGTA. (2008). *Treasury Inspector General for Tax Administration Investigation Hiughlights*. Retrieved January 18, 2008, from http://www.ustreas.gov/tigta/oi_highlights.shtml#3

Tucker, C., J. , Brick, M., & Meekins, B. (2007). Household Telephone Service and Usage Patterns in the United States in 2004: Implications for Telephone Samples. *Public Opinion Quarterly, 71*(1), 3-22.

USMIHS. (2001). *Patients, Privacy, and Federal Healthcare: What Policy Makers Need to Know About the HIPAA Privacy Rule*, 2003, from http://www.usminstitute.org

www.manaraa.com

Washington Post. (2007, April 8, 2007). Stealing From the IRS. *The Washington Post,* p. 6.
Washington Post, T. (2007, Monday, June 4, 2007). *Washington Post,* p. 13.
WebCPA. (2007). *IRs Security Still Lax*. Retrieved January 18, 2008, from http://www.webcpa.com/article.cfm?articleid=25009
Weber, T. M. (2005, July 2005). *Values in a National Information Infrastructure: A Case Study of the U.S. Census Bureau.* Paper presented at the 14th International conference of the Society of Philosophy and Technology, Delft, The Netherlands.
Weinberg, D. H., Abowd, J. M., Rowland, S. K., Steel, P. M., & Zayatz, L. (2007). *Access Methods for United States Microdata*.Unpublished manuscript, Washington DC.
Winkler, W. E. (1998). Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Research in Official Statistics, 1*, 87-104.
Yin, R. K. (2003). *Case Study Research: Design and Methods* (third ed. Vol. 5). Thousand Oaks: SAGE Publications.

www.manaraa.com

Appendix I: Census Bureau Surveys

Reimbursable Demographic Surveys

SURVEY	BRIEF DESCRIPTION	AUTHORITY
HUD		
American Housing Surveys (AHS)	To provide a current and continuous series of data on selected housing and demographic characteristics.	The Department of Housing and Urban Development (HUD) sponsors the survey under the authority of Title 12, United States Code, Sections 1701z-1, 1701z-2(g), and 1701z-10a. The Census Bureau performs the work under Title 13.
Survey of Market Absorption (SOMA)	To measure the rate at which different types of new rental apartments and new condominium apartments are absorbed, usually by being rented or sold over the course of the first 12 months following completion of a building.	HUD sponsors the survey under the authority of Title 12, United States Code, Sections 1701z-1 and 2. the Census Bureau performs the work under Title 13.
New York City		
New York City Housing Vacancy Survey (NYCHVS	To determine the vacancy rate for New York City's rental stock. New York City also uses the data to measure the quality and quantity of housing and demographic characteristics of the city's residents.	The New York City Department of Housing Preservation and Development (NYCHPD) sponsors the survey. Local authorization of the survey is pursuant to the Local Emergency Housing Rent Control Act, Sections 26-414 and 26-415 of the Administrative Code of the City. The Census Bureau performs the work under Title 13.
BLS		
American Time Use Survey (ATUS)	To provide nationally representative estimates of the amount of time that Americans spend in various activities. The Bureau of Labor Statistics (BLS) uses the data to measure the value of unpaid, productive work, such as housework and child care, and nonproductive activities, like waiting in line and commuting.	The Bureau of Labor Statistics (BLS) sponsors the survey under the authority of Title 29. The Census Bureau performs the work under Title 13.
Consumer Expenditure (CE) Survey	To provide a current and continuous series of data on consumer expenditures and other related characteristics which are used to determine the need to revise the Consumer Price Index (CPI), update the weights used to calculate the index, and for use in family expenditure studies and other analyses.	BLS sponsors the survey under the authority of Title 29. The Census Bureau conducts the work under Title 13.
Current Population Survey (CPS)	To provide estimates of employment, unemployment, and other characteristics of the general labor	The Census Bureau and the BLS jointly sponsor the survey under the authorities of Title 13, United States Code, Section

www.manaraa.com

	force, of the population as a whole, and of various subgroups of the population. Monthly labor force data for the country are used by the Bureau of Labor Statistics (BLS) to determine the distribution of funds under the Job Training Partnership Act.	182, and Title 29, United States Code, Sections 1-9.
Current Population Survey (CPS) Supplements	There are a number of supplemental surveys conducted during specific months for a variety of agencies not listed here	
RENT Survey (aka Rent and Property Tax Survey or the CPI Housing New Construction Survey.	To provide the BLS with a sample of addresses from listings of multi-unit addresses from the permit area listing (PAL) frame. The BLS conducts interviews at the addresses for the housing component of the Consumer Price Index (CPI).	The BLS sponsors the survey under the authority of Title 29, United States Code, Section 2. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
Telephone Point-of-Purchase Survey (TPOPS)	To obtain the names and locations of retail, wholesale, and service establishments (outlets) at which consumers purchase specified goods and services (commodities). The BLS uses the data to select and update outlets included in their Consumer Price Index (CPI) pricing surveys.	The BLS sponsors the survey under the authority of Title29, United States Code, Section 2. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
NCES		
Schools and Staffing Survey (SASS)	To collect the information necessary for a complete picture of American elementary and secondary education. The linkage of the SASS components enables researchers to examine the relationships among these elements of the education system.	The NCES, Institute of Education Sciences, sponsors the survey under the authority of Public Law 107-279, Title 1, Part E, Sections 151(b) and 153(a) of the Education Sciences Reform Act of 2002. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
Library Media Center (LMC) Questionnaire/Schools and Staffing Survey	To collect the information necessary for a complete picture of American elementary and secondary school libraries. The survey is a component of the SASS allowing for linkage of the library data to the school and district for analysis.	The National Center for Education Statistics (NCES), Institute of Education Sciences, sponsors the survey under the authority of Public Law 107-279, Title 1, Part E, Sections 151(b) and 153(a) of the Education Sciences Reform Act of 2002. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
Private School Survey (PSS)	To develop and maintain a comprehensive universe file of private schools in the United States and to obtain data from these schools that are comparable to the state level data obtained by the NCES for the public school sector.	The NCES, Institute of Education Sciences, sponsors the survey under the authority of Public Law 107-279, Title 1, Part E, Sections 151(b) and 153(a) of the Education Sciences Reform Act of 2002. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
School Crime Supplement (SCS)	To provide information on school-related victimizations on a national level.	The NCES sponsors this survey under the authority of Title 42, United States Code, Section 3732 of the Justice

www.manaraa.com

		Systems Improvement Act of 1979 authorizing the collection of statistics on victimization.. The Census Bureau performs the work under Title 13.
Schools Survey on Crime and Safety (SSOCS)	To provide estimates of school crime, discipline, disorder, programs, and policies. The SSOCS questionnaire asks principals to report on a variety of topics related to crime and safety.	The NCES, Institute of Education Sciences, sponsors the survey under the authority of Title I, Part E, Sections 151(b) and 153(a) of Public Law 107-279, the Education Sciences Reform Act of 2002. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
Teacher Follow-Up Survey (TFS)	To determine the teacher attrition rates in public and private schools and to obtain data on the characteristics of teachers who leave the profession and those who stay.	The NCES, Institute of Education Sciences, sponsors the survey under the authority of Public Law 107-279, Title 1, Part E, Sections 151(b) and 153(a) of the Education Sciences Reform Act of 2002. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
NSF		
National Survey of College Graduates (NSCG)	A longitudinal survey to provide data on the number and characteristics of experienced individuals with education and/or employment in science or engineering (S&E) living in the United States.	The NSF sponsors the survey under the authority of the National Science Foundation Act of 1950, as amended (Title 42). For the sample members whose source is either the 1990 or 2000 census, the survey is conducted under Title 13. For all other sample members, the survey is conducted under Title 15.
National Survey of Recent College Graduates (NSRCG)	To provide data on the size and characteristics of new entrants to the science and engineering (S&E) workforce by surveying recent bachelor's and master's degree recipients. Together with the Survey of Doctoral Recipients and the National Survey of College Graduates, the National Science Foundation (NSF) includes data from this survey in the Scientists and Engineers Statistical Data System (SESTAT) as mandated by the Congress.	The NSF sponsors this survey under title 42. The Census Bureau performs the work under Title 15, United States Code.
BJS		
National Crime Victimization Survey (NCVS)	To provide personal victimization and property crime rates from a general population sample. Data are gathered on types and incidence of crime; monetary losses and physical injuries due to crime; characteristics of the victims; and, where appropriate, characteristics of the offender.	Bureau of Justice Statistics (BJS) sponsors the survey under the authority of Title 42, United States Code, Section 3732. The Census Bureau performs the work under Title 13.
Supplemental Victimization Survey (SVS)	To provide information about the nature and consequences of a series of unwanted contacts or harassing	The Office of Violence Against Women (OVW) sponsors the survey under authority of Title 42, United States Code,

www.manaraa.com

	behavior directed toward respondents that frightened, concerned, angered, or annoyed them.	Section 3732 of the Justice Systems Improvement Act of 1979. The Census Bureau performs the work under Title 13.
National Prisoner Statistics (NPS) Program	To provide information on adults incarcerated in state and federal correctional institutions, including their characteristics, movements, and history.	The BJS sponsors the survey under the authority of Title 42, United States Code, Section 3732. The Census Bureau conducts the survey under Title 15, United States Code, Section 1525.
Survey of Inmates of Local Jails (SILJ)	To provide detailed information on the criminal histories of jail inmates, their recent offenses and sentences, their socioeconomic and family backgrounds, their use of drugs and alcohol, and their activities and the health care they receive while confined. The survey also provides information on victims of violent offenders.	The BJS sponsors the survey under the authority of Title 42, United States Code, Section 3732. The Census Bureau conducts the surveys under Title 15, United States Code, Section 1525.
NCHS		
National Health Interview Survey (NHIS)	To provide information on a continuing basis about the prevalence and distribution of illness, its effects in terms of disability and chronic impairments, and the kind of health services people receive.	The National Center for Health Statistics (NCHS) is sponsoring the survey under the authority of Title 42, United States Code, Section 242k. The Census Bureau is performing the work under Title 15, United States Code, Section 1525.
National Hospital Discharge Survey (NHDS)	To provide demographic and medical data on discharged patients and other hospital information on a national basis annually.	The National Center for Health Statistics (NCHS) sponsors this survey under the authority of Title 42, United States Code, Section 242k. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
National Hospital Ambulatory Medical Care Survey (NHAMCS)	To provide information about the health problems of ambulatory patients and the treatment given to them in hospital emergency rooms and outpatient departments. Information from the NHAMCS is used to supplement existing ambulatory care data obtained from the office-based survey, the National Ambulatory Medical Care Survey (NAMCS).	The NCHS sponsors the survey under the authority of Title 42, United States Code, Section 242k. The U.S. Census Bureau is performing the work under Title 15, United States Code, Section 1525.
National Survey of Ambulatory Surgery (NSAS)	To gather and disseminate nationwide data about ambulatory surgery performed in hospitals and freestanding ambulatory surgery centers in the U.S. Survey data are abstracted from sampled medical records of ambulatory surgery visits.	The NCHS sponsors this survey under the authority of Title 42, United States Code, Section 242k. The Census Bureau performs the work under Title 15, United States Code, Section 1525.
FWS		

www.manaraa.com

National Survey of Fishing, Hunting, and Wildlife-Associated Recreation 2001 (FHWAR)	To provide current data on fishing, hunting, and wildlife-related activities of a nonconsumptive nature, such as feeding, observing, and photographing wildlife.	The Fish and Wildlife Service (FWS) of the Department of Interior sponsors this survey under the authority of the Fish and Wildlife Coordination Act of 1956 and the Federal Aid in Sport, Fish, and Wildlife Restoration Acts (Title 16). The Census Bureau performs the work under Title 13.
NIH		
Wave 2 of the 2001 National Epidemiologic Survey on Alcohol and Related Conditions (NESARC)	A longitudinal study for the National Institute on Alcohol Abuse and Alcoholism (NIAAA) on alcohol use, experiences with alcohol and related conditions, as well as the demographics and family history used in analyzing health data.	The NIAAA, an agency of the National Institutes of Health, is the sponsor. The Census Bureau performed the work under Title 13.
National Long-Term Care Survey (NLTCS)	To obtain data on the elderly's ability to perform daily acts of living, the limitations that prevent or impair their ability, the amount and type of care required, and their socioeconomic characteristics, such as age, income, and marital status.	The Center for Demographic Studies (CDS), Duke University, sponsors the survey under a grant from the National Institute on Aging (NIA) under authority of Title 42, United States Code, Section 285-e-1. The Census Bureau performs the work under Title 15.
CNCS		
2007 Youth Volunteering and Civic Engagement Survey (YVCES)	Information from this survey will be used as the basis to promote the growth of teen participation in the community.	The Corporation for National and Community Service is the sponsor and the work is performed under Title 45. The Census Bureau performs the work under Title 13.

www.manaraa.com

Economic Surveys

Survey of Business Owners (SBO)	**Measures the demographic characteristics of business owners across the economy of the United States**	**Information collected under Title 13**
Commodity Flow Survey (CFS)	**Measures the movement of goods in the United States**	**Joint project of several federal agencies: the Census Bureau, the Department of Commerce, the Bureau of Transportation Statistics, and the Department of Transportation**
Business Expenditures Survey (BES)	**Estimates expenditures, depreciable assets, and operating costs for wholesale, retail, and some service companies**	**Bureau of Economic Analysis uses BES data to benchmark several economic indicators, including its national income and product accounts. Other federal agencies also use BES data for cost and expenditures data**
Building Permits Survey	**Designated principal economic indicator and the only source of current and consistent small-area data on newly-authorized construction**	
Annual Monthly Retail Sales Survey	**Provides an early indication of sales by retail companies with one or more establishments that sell merchandise and associated services to final consumers**	
Monthly Wholesale Trade Survey	**Collects monthly estimates of sales and inventories data from companies primarily engaged in merchant wholesale trade**	

www.manaraa.com

Appendix II: The Census Bureau's Privacy Principles

From the Census Bureau website: U.S. Census Bureau website:
http://www.census.gov/privacy/Privacy_Principles_signed_updated_040606_.doc

U.S. CENSUS BUREAU PRIVACY PRINCIPLES

1. **Principle of Mission Necessity**: ***The U.S. Census Bureau will only collect information that is necessary for meeting the Census Bureau's mission and legal requirements.***

   Subprinciple 1 - The Census Bureau will only collect or acquire information about individuals and businesses that is necessary to meet its legal responsibility and fulfill its mission to provide timely, relevant, and quality data about the people and economy of the United States.

   Subprinciple 2 - The Census Bureau will only engage in projects requiring data protected under Title 13, United States Code, if there is a clear benefit to Census Bureau programs.

   Subprinciple 3 - The Census Bureau will only collect or acquire information on a reimbursable basis, or in exchange for products or services, if such collection or acquisition would be seen as being consistent with the Census Bureau's reputation of providing relevant statistical data for public policy and maintaining the public's trust.

   Subprinciple 4 - The Census Bureau will ensure that it uses the data it obtains or collects only for statistical purposes and will advise the public of these limited uses.

2. **Principle of Openness:** ***The Census Bureau will be open about its programs, policies and practices to collect and protect identifiable data used to produce statistical information.***

   Subprinciple 1 - The Census Bureau will make it easy to access information about what we collect and why, and provide opportunities for public comment prior to collecting new information.

   Subprinciple 2 – When we collect information, respondents will be informed about the purpose, authority, reporting obligation, legal protections, and uses.

   Subprinciple 3 - When we acquire and link identifiable records from other organizations as part of creating statistical products, we will be open about our activities and inform those supplying the records of proposed uses in order to confirm that they are permitted.

www.manaraa.com

Subprinciple 4 - Once we have assured the confidentiality of the data, the Census Bureau does not attempt to control the uses or users of its products. Further, we release the identity of all requesters of custom data products and make those same products publicly available.

3. **Principle of Respectful Treatment of Respondents:** ***The Census Bureau will be considerate of respondents' time and desire for privacy and will respect their rights as research participants.***

   Subprinciple 1 - When we design our data collections, the Census Bureau will employ efficiencies to minimize respondents' time and effort.

   Subprinciple 2 - The Census Bureau will engage only in legal, ethical and professionally accepted data collection practices.

   Subprinciple 3 - The Census Bureau will request sensitive information from children and other sensitive populations only when it has determined that doing so will provide a clear benefit to the public good and will not violate federal protections of human research participants.

4. **Principle of Confidentiality:** ***The Census Bureau will ensure that confidentiality protections are included in its procedures to collect, process, and release data.***

   Subprinciple 1 - The Census Bureau will permit authorized users access to, and use of, only that confidential data needed to conduct their work in support of Census Bureau programs.

   Subprinciple 2 - The Census Bureau will use appropriate and comprehensive physical and communications security measures when collecting, storing, and analyzing all legally protected information held by the Census Bureau.

   Subprinciple 3 - The Census Bureau will use comprehensive disclosure avoidance techniques consistent with professionally acceptable standards before releasing data products derived from legally protected information.

   Subprinciple 4 - Agencies supplying legally protected information to the Census Bureau will always be given the opportunity to review and approve either the proposed data releases or the disclosure methodology used to protect the data in order to ensure that the agencies' disclosure-protection requirements are met.

www.manaraa.com

Appendix III: Census Bureau Privacy Principles as Presented to the Public

From the Census Bureau website: U.S. Census Bureau website:
http://www.census.gov/privacy/files/data_protection/002822.html (December 2007)

Our Privacy Principles

We depend on your cooperation and trust, and we promise to protect the confidentiality of your information. The Census Bureau's Privacy Principles remind us of this promise and help ensure the protection of your information throughout all of our activities.

The Privacy Principles are our guidelines. They help us as we design surveys to consider respondents' rights and concerns. Every principle embodies a promise to you, the respondent.

- **Necessity:** Do we need to ask this question? Do we need to collect this information?

  Every time we prepare to ask a question, we determine whether the information is truly necessary. All of the information we collect is used for federal programs.

  - We promise to collect only information necessary for each survey and census.
  - We promise that we will use the information only to produce timely, relevant statistics about the population and the economy of the United States.
- **Openness:** Do you know why we are collecting your information?

  We collect information only for statistical purposes, and it is never used to identify individuals. Before participating, you have the right to know why we are conducting the survey or census, why we are asking specific questions, and the purposes for which the information will be used.

  - We promise to inform you about the purpose and uses for every survey or census we conduct **before** you provide your answers to us.

www.manaraa.com

- **Respectful Treatment of Respondents:** Are our efforts reasonable and did we treat you with respect?

    - We promise to minimize the effort and time it takes for you to participate in the data collection by efficient designs.
    - We promise to use only legal, ethical and professionally accepted practices in collecting data.
    - We promise to ensure that any collection of sensitive information from children and other sensitive populations does not violate federal protections for research participants and is done only when it benefits the public good.
- **Confidentiality:** How do we protect your information?

    In addition to removing personally identifiable information, such as names, telephone numbers, and addresses, from our data files, we use various approaches to protect your personal information; including computer technologies, statistical methodologies, and security procedures.

    Our security measures ensure that only a restricted number of authorized people have access to private information and that access is only granted to conduct our work and for no other purposes. Every person who works with census confidential information collected by the Census Bureau is sworn for life to uphold the law.

    Violating the confidentiality of a respondent is a federal crime with serious penalties, including a federal prison sentence of up to five years, a fine of up to $250,000, or both.

    - We promise that every person with access to your information is sworn for life to protect your confidentiality.
    - We promise that we will use every technology, statistical methodology, and physical security procedure at our disposal to protect your information.

www.manaraa.com

Appendix IV: CMS Privacy Principles

From the CMS website:
http://www.cms.hhs.gov/PrivacyOffice/03_Privacy_BasicPrinciples.asp#TopOfPage

Privacy: Basic Principles

Privacy issues are implicated in a wide range of activities in both our personal and public lives.

Our concept of Privacy includes

- Control of information concerning our personal life
- Freedom from intrusion upon one's seclusion
- Limits on publicity that places one in a false light
- Prevention of identity theft, and the theft of one's name or likeness
- Right to keep personal information confidential

General Privacy Principles for Public and Private Sectors

- Personal information should be acquired, disclosed, and used only in ways that respect and
- individual's privacy.
- Personal information should not be improperly altered or destroyed.
- Personal information should be accurate, timely, complete, and relevant to the purpose for which it is provided and used.

Basic Principles of the Privacy Act of 1974

- Specifically mandates that the government
  - inform people at the time it is collecting information about them, why the information is being collected and how it will be used
  - publish a notice in the *Federal Register* of new or revised system of records about
  - individuals.
  - Publish a notice in the *Federal Register* before conducting a computer matching program.
  - assure the information is accurate, relevant, complete, and up-to-date before disclosing it to others.
  - allow individuals access to records on themselves.
  - allow individuals to find out about disclosures of their records to other agencies and persons.
  - provide individuals with the opportunity to correct inaccuracies in their records.

www.manaraa.com

Appendix V Interview Question Guide

Code Number_____

1. Did you work at the (Census Bureau/IRS) during the records audit that began in 1998?
2. Did you work at the (Census Bureau. CMS) prior to implementation of the HIPAA regulations at HHS?
3. Please describe your job duties and how long you have been performing them.
4. Please describe specifically what work you do with administrative records, including types and sources of records you work with.
5. Please describe what the workflows of administrative records are between your agency and the (Census Bureau, IRS, CMS).
6. What are the laws governing your agency's handling of administrative records?
7. What are your agency's policies regarding the handling of administrative records?
8. Who are the people involved in working with administrative records outside of this agency, including contractors and researchers?
9. Please describe your interactions with these outside parties.
10. Please discuss how enactment of CIPSEA has changed your work processes with regard to sharing of administrative records. (For CMS: HIPAA)
11. Discuss some of the interagency agreements that your agency has with other agencies regarding the sharing and combining of administrative records.
12. Describe how the records are shared with researchers, including the processes for approving projects, maintaining confidentiality, data handling, quality control, etc.

www.manaraa.com

13. To what uses are shared administrative records put? Does this affect how the data are handled?

14. Please describe the training received by the people who handle the records;

15. What is the method of compliance measurement for employees, contractors, researchers, other agencies?

16. Do you believe that there are any significant issues that have arisen as a result of the business process flow and the need to protect privacy and confidentiality? Please elaborate.

17. Are there any areas where you think that the laws, rules, and regulations overlap or conflict?

18. After administrative records are combined with other records and statistical data, which agency "owns" the combined data?

19. What are the benefits of data sharing among agencies?

20. Do you believe that these benefits are being realized? If not, what are the barriers to achieving the intended benefits?

21. Who are the other people in this agency and in the other agencies in this study that you would recommend be interviewed, and why?

22. Would you mind being re-interviewed?

www.manaraa.com