Developing Adoptable Disclosure Protection Techniques: Lessons Learned from a U.S. Experience

Nicholas H. Greenia

Statistics of Income Division
U.S. Internal Revenue Service
P.O. Box 2608
Washington, DC 20013 USA

The views expressed in this paper are the author's and not necessarily those of the U.S. Internal Revenue Service.

Abstract. The development of new disclosure protection techniques is useful only insofar as those techniques are adopted by statistical agencies. In order for technical experts in disclosure limitation to be successful, they are likely to need to interact with the appropriate statistical offices. This paper discusses just such a successful interaction in the United States. It describes the foundation that three major U.S. agencies -- the Census Bureau, the Social Security Administration, and the Internal Revenue Service – laid in order to develop more useful statistical products. These included a proposed synthetic data public-use file based on the confidential microdata from all three agencies. Since then other governmental organizations, such as the U.S. Congressional Budget Office, have become involved with this inter-agency effort, which seeks to provide researchers and other users in the broader statistical community with a data utility often possible previously only with access to the confidential microdata. The confidentiality implications for all three agencies – and the potential for more -- of a successful conclusion to this work would be enormously beneficial to data users, data producers, and data respondents. This paper describes the importance of developing the necessary framework, which includes an understanding between statistical office decision makers and the technical experts, before beginning such an endeavor. It provides a description of how this effort even became possible, and uses the history of events and related lessons to describe essentials that might be useful for other national statistical offices facing similar constraints and goals.

1 Introduction

The development of new disclosure protection techniques is useful only insofar as those techniques are adopted by statistical agencies. For technical experts in disclosure limitation to be successful, they are likely to need to interact with the

appropriate statistical offices. This paper discusses just such a successful interaction in the United States.

Since 2001 inter-agency efforts have been underway on a synthetic data approach to produce a public-use file (PUF), which would combine selected statistical and administrative data from three U.S. agencies: the Census Bureau's Survey of Income and Program Participation (SIPP), retirement and disability benefits data from the Social Security Administration (SSA), and limited earnings data from tax records filed with the Internal Revenue Service (IRS). Based on progress so far, the outlook for this work is promising. The confidentiality and research benefits of this approach, if successful, could be substantial, but details of that technical discussion are left for other papers.

It is important to note, however, that technological advances in disclosure protection are necessary, but not sufficient, conditions for the adoption of new techniques. This paper focuses primarily on describing the evolution of the legal, institutional, and bureaucratic environment that was the critical precursor of the interagency effort. Out of the story come lessons that may help other national statistical offices cope with similar challenges.

This story is largely a confluence of separate but related events:

- The development of an institutional interagency trust, after a serious test of the fundamental relationship;
- The recognition by the Census Bureau of the deteriorating tradeoff between data quality and data protection in the release of previous SIPP public use files, which was influential in deciding to pursue the synthetic data PUF approach; and
- The development of a new program (Longitudinal Employer-Household Dynamics) that brought in the technical know-how that permitted the integration of statistical and administrative data within the new program, and the creation of the aforementioned SIPP/SSA/IRS PUF.

This paper focuses primarily on the first of these, but also notes the relevance of the other events.

2 Background

Statistical agencies have become increasingly aware that two relatively new challenges may seriously affect their ability to release data into the public domain, whether in tabular or public-use file format. Increasing capabilities of computing power and advances in mathematical/statistical techniques have led to the increase in technical re-identification capacity. This challenge is matched by a practical increase in this capacity due to the proliferation of datasets in the public and private/commercial domain. In spite of these challenges, the need for publicly collected confidential data to inform decisions in both government and the private sector is not expected to abate.

The U.S. tax administration agency, the Internal Revenue Service (IRS), faces additional challenges in its role as an important administrative data provider for the Federal statistical system. Tax data have always been particularly susceptible to re-identification, both because of their relatively widespread distribution in public form and because of their sensitive content. In addition, because publicly and privately available datasets are often directly based on entities also in the tax system, there is more potential to match to tax data and re-identify taxpayers. Moreover, IRS views the protection of taxpayer confidentiality as an essential component of successful voluntary tax compliance, upon which the tax system relies. Because of the several U.S. statistical agencies authorized to receive confidential tax data, IRS must not only preserve tax data confidentiality within its own administrative system, but also oversee the safeguarding of tax data in the systems of the recipient statistical agencies. In a related vein, IRS must ensure that the numerous products produced by each statistical agency cannot be statistically "cross-matched" and thereby enable complementary disclosure of identifiable information.

Because of these additional challenges, IRS must insist that its safeguarding standards be met by a recipient statistical agency, regardless of the agency's standards for data it collects directly. This requirement of compliance with administrative data provider standards also influenced the authorization process for statistical use of tax data by Census, as will be shown later, but this requirement may differ for other countries. For example, the United Kingdom's Office of National Statistics stipulates that, "the same confidentiality standards will apply to data derived from administrative sources as apply to those collected…for statistical purposes".[1] Nevertheless, the unmistakable conclusion is that it is becoming increasingly difficult to release even aggregate tabular data into the public domain, and public-use files (often of most use to researchers without access to the original source data) pose special challenges that are exacerbated over time in the public domain. Although closer coordination of all releases is advisable, new methods of confidentiality protection may afford the most hope for data users, data providers, and ultimately, the respondents themselves.

While issues surrounding the disclosure of confidential data are common to all Federal statistical agencies, IRS also has its own idiosyncratic issues.[2] Confidential tax data, also known as Federal Tax Information (FTI), have several uses, including specifically authorized statistical purposes. The homogeneous treatment of FTI results from restrictions in the tax statute, the Internal Revenue Code (IRC), which do not allow IRS to distinguish among FTI data elements--even as to age. That is, there is no statute of limitations as there is for confidential microdata at statistical agencies such as the U.S. Census Bureau. In addition, the tax statute does not distinguish among different types of data or taxpayers, so that the Social Security Number of

---

[1] P. 6, Working Paper No. 11, *Contexts for the Development of a Data Access and Confidentiality Protocol for UK National Statistics*, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003.

[2] Confidential data are any identifiable data whose public release is unauthorized. The removal of identifier information, such as name, address, and identification numbers, is a necessary but insufficient condition to render such data anonymous or unidentifiable.

John Q. Citizen in Anywhere, USA, would receive the same protection as that of Bill Gates which, in turn, would be protected as much as all the financial information on any business tax return which Microsoft Corporation might file. Accordingly, all FTI--whether entity or tax module information[3]--must be treated and protected in perpetuity as equally sensitive and confidential. This task of protecting confidentiality, given the ever-increasing amount of data for which IRS becomes responsible over time, is expensive and technically challenging.

The tax law's anonymity standard is indiscriminate and absolute in requiring that all tax data, whether business or individual, be released in anonymous form. The anonymity requirement for data publicly released by IRS also applies to statistical agencies authorized to receive FTI. However, although the general standard applies, the actual disclosure protection methodology is not specified. The requirement is simply that whatever methodology is used be either identical to that employed by IRS or else an equivalent approved by IRS.

The practical question confronting any methodology attempting to meet the absolute anonymity standard is: From what sort of intrusion must the data be protected? Must it be absolutely impossible to re-identify a taxpayer using any means available, or is there some less rigid methodological standard? Traditionally, the answer has been that tax data must be protected from potential intruders who, using "reasonable means," might attempt to make such a re-identification. Reasonable means include the use of reasonably available computer technology, mathematical/statistical techniques, and a working knowledge of the subject matter to which the data apply. The reasonable means standard is a good effort to keep the entire system from shutting down and being replaced by a policy of no data release at all--probably the only way to guarantee no re-identification. The problem, as can probably be imagined in 2004, is that the concept of reasonable means is a technology-relative concept and may be a moving target too elusive to be relevant for the absolute standard of anonymity. As a result, in a time of increasingly tight budgets protecting the confidentiality of tax data is becoming a task virtually impossible to execute successfully.

---

[3] An abbreviated course in IRS master files might summarize data maintained on these systems (whether individual or business master file) as being one of two types: entity information or tax module information. Entity information refers to information used to identify and locate a taxpayer such as Taxpayer Identification Number (Social Security Number—SSN, Employer Identification Number—EIN), Name, Address, and perhaps Industry Classification Code (NAICS or SIC-based) for a business. Everything else is tax module information.

3 Developing Interagency Trust

3.1 A Breakdown in the Relationship

In 1999 IRS began its mandated triennial safeguards review of a principal U.S. statistical agency, the Census Bureau. Although the U.S. statistical system is more decentralized than that of many European Union countries, Census receives the preponderance of confidential tax data for statistical purposes as a result of the statutory authorization conferred by section 6103(j)(1)(A) of Title 26 of the United States Code (USC). The implementing Income Tax Regulations specify both the actual items authorized for access and their access purpose or Title 13, Chapter 5, USC.

The mandated IRS safeguards review of Census (and other recipient agencies of confidential tax data) is a result of the same section, 6103, which authorizes such access in the first place. As a result of the 1999 IRS safeguards review, deficiencies in the oversight process were uncovered by IRS, some of which reflected poorly on both Census and IRS. For example, Census used tax data for some projects which had not received explicit IRS approvals, but IRS had made explicitly clear neither the need for such approvals nor the process for effecting them in a coordinated fashion.

As it became clear that neither Census nor IRS could resolve the resulting crisis, intervention at high levels of government became necessary. Eventually, the U.S. Office of Management and Budget (OMB), which has broad oversight responsibilities for Federal statistical agencies, helped broker an understanding between the two agencies based upon three essential points:

(1) Census must comply with IRS safeguard standards in order to protect the confidentiality of tax data,

(2) informed decisions by policy makers inside and outside government require the best possible data available, and

(3) tax data are so important to these information decision systems that their exclusion is not a viable option.

Thus, the conclusion of this process was that IRS, as an administrative data provider, and Census, as an administrative data user, would have to find a way to make their relationship work in order to satisfy the several stakeholders involved; that is, an inter-agency “trainwreck” or shutdown was viewed as unacceptable and would not be tolerated.

As a result, IRS and Census recognized that the increasingly murky and implicit boundaries within which their relationship had been struggling were inadequate as guidance. Further, a relationship was needed which would not only work but which would better accommodate the increasingly complex needs of the many end users. Essentially, the relationship needed to be not only re-evaluated but also recalibrated, especially to accommodate a new form of confidential data access created by Census for outside researchers meeting new Census study needs: the Research Data Center

(RDC) consortium operated by its Center for Economic Studies. Like statistical agencies in other countries,[4] Census had realized the need to explore other venues for purposes of improving its statistical knowledge base and data quality, but only as a result of the IRS safeguards review did this realization include the need to integrate its RDC's into the overall process encompassing its other longstanding functions.

To meet especially the need on new statistical research uses of FTI, a clear and detailed understanding that met the mandates of both agencies needed to be documented. Accordingly, an IRS-Census policy agreement, *Criteria for the Review and Approval of Census Projects that Use Federal Tax Information*, better known as the Criteria Agreement, was mutually devised and eventually signed into effect by both agencies in September 2000. At the core of this agreement, available at www.ces.census.gov, was the understanding that any data use or access had to be authorized by an explicit approval process involving both the data provider, IRS, and the data user, Census, and that, especially for outside researcher access, the predominant purpose of such access had to be the benefit of Census under its own statutory mandate; namely, Title 13, Chapter 5, United States Code.

In effect, the Criteria Agreement established and refined not only the protocols, but most importantly, the authorization to fully legitimize Census use of confidential tax data. It was implicit in this agreement that exclusively statistical use was a necessary but insufficient condition for authorized access. Instead, an explicit approval by the data provider and user was required which attested to the access authorization under the statutes of both IRS and Census, the IRS implementing regulations, and the Census-IRS Criteria Agreement's specific requirements in order to satisfy the record for a particular programmatic use. This point is worth emphasizing, as it was not enough that data provider and user agreed to the general imprimatur provided by the statutory and regulatory bases for proposed access by the user. Because the Census-RDC model was seen as at the vanguard, if not the frontier, of data access, it was especially important that the record explicitly demonstrate the data provider was convinced of the proposed statistical use's justification. This type of specific dual approval is also necessary for another unique data access model with similar high visibility disclosure risk; namely, the public-use file.

Implicit to this inter-agency relationship is the notion that the record of all actions taken must be able to demonstrate not only authorized intent but credibility--for some pending audience of critics. This inevitable, critical eye is known as third party scrutiny, and it is neither hypothetical nor irrelevant, instead consisting of both explicit and implicit oversight bodies such as the U.S. Congress' General Accounting Office, the U.S. Treasury Inspector General's Office, privacy advocates, the media, and ultimately, the respondents themselves. In preparing for third party scrutiny the record underlying data access should credibly demonstrate that the process has anticipated as many factual questions as possible and that it has also considered

[4] For example, see Working Paper No. 10, *Research Data Centres of Official Statistics*, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003.

perceptions as well. Thus, the process needs to demonstrate consistently that it operates within not only the letter of the agreement but also its intent--so that accountability, authorization of the access granted, and purpose are never in doubt. To address both outside perceptions and the reality of third party scrutiny, Census and IRS agreed on the importance of exceeding the literal requirement of the agreement whenever possible. For this reason both agencies agreed that it would be a rare occasion demanding minimum adherence to predominant purpose as an acceptable criterion; that is, only over 50 percent of the access purpose. Consequently, approval on the margin would not be the rule, but the exception.

Perceptions, in conjunction with concerns about third party scrutiny, played a large role in this need for dual explicit authorization by data provider and user, especially for outside researchers engaged by a national statistical agency such as Census. Again, it was vital that access of the provider's administrative data not be construed as a type of unauthorized usage disassociated from or only loosely associated with the statistical user's mandate and mission, especially when the resulting analytical data had the potential for affecting groups of respondents. Without explicit evidence; that is, the mutual approvals of both the administrative data provider and the statistical user signifying that the specific use was authorized, third party scrutiny might raise troubling questions as to the type of confidentiality protection assured by the administrative data provider, which assumes virtually all risk with its respondent population. This issue goes to the heart of accountability in data stewardship.

One reason for the IRS-Census impasse in 2000 is that there is a fundamental and inexorable tension due to the conflicting nature of their respective mandates. Census is mandated to use administrative data to the maximum extent possible in order to reduce respondent burden and processing costs. IRS is mandated to provide confidential tax information only to the minimum extent necessary. This inherent tension imposes a sort of *de facto* equilibrium in the intersection of the agencies' confidentiality cultures, and only the strongest part of each culture is allowed relevance. It is thereby critical to protecting confidentiality, including perceptions of abuse, as both data provider and user must bargain hard for an acceptable access transaction that satisfies their respective mandates. Critical to such success is a set of clearly defined terms and processes, and the documentation of subsequent actions following such a process. Equally critical is the devotion of sufficient resources to ensure the needed safeguards. Because resources are finite, so must be the amount of access whose safeguarding can be demonstrably credible. Without resource commitment to verifiable standards of protection, the clear implication is that access can approach infinite levels, suggesting both an inability and a lack of commitment to safeguard the data effectively.

3.2 Rebuilding the Relationship: Implementation of the Criteria Agreement

It was clear at the inception of the Criteria Agreement that the many new proposals of the RDCs' outside researchers would be tied to the Census Bureau's future viability,

especially its ability to keep up with the new statistical needs of decision makers. That is, the RDC project proposals were seen as critical to maintaining the statistical heartbeat at Census.

In fact, most of the FTI access proposals came from Census RDC's, and initially, Census and IRS reviewed these proposals concurrently. This arrangement was soon abandoned for primarily one reason. Although it was inefficient for IRS, the administrative data provider, to spend time reviewing proposals ultimately rejected by Census, it was critical that the fundamental criterion of all tax data access; that is, a proposal's predominant purpose of benefiting Census under Title 13, Chapter 5, be demonstrated in proposals that Census, as data user, first approved. That is, the Census review process was supposed to consider not only scientific merit but also Title 13, Chapter 5, predominant purpose, while IRS review considered only the latter. Once it became clear that Census needed to take responsibility for both aspects of review (although IRS, as data provider, maintained ultimate control as data owner) the human review capital, especially regarding requirements for tax data access, could be transferred upstream from IRS to Census, and then from Census to the researcher community. Thus, the confidentiality culture needed by the data provider to assuage its third party scrutiny concerns was necessarily integrated into the data user's confidentiality culture as well as that of its researcher community. In turn, this culture colonized prospective researchers.

Outside researchers realized they had two critical interests in helping such a system succeed. First, the perpetuation of the Census-IRS arrangement allowed the researcher community access to FTI for authorized purposes, which required undertaking only proposals within scope. Second, by learning the needed culture, researchers could help increase the probability of their own proposals being approved, and even increase the number of proposals which might be possible, by theoretically and *ceteris paribus*, shortening the review process itself.

However, to counter the potential for insincere or even fraudulent researcher behavior, IRS, as administrative data provider, and Census, as data user, also conveyed three fundamental understandings to the researcher community. First, cheating on proposal purpose would inevitably be self-defeating, as it would destroy the process. Thus, implicit, if not explicit, peer-policing among the researcher community was essential to the process succeeding, and was encouraged by both Census and IRS. In fact, both agencies took pains conveying directly to the researcher community that while it might be possible to deceive both agencies' reviews, it would be at a cost fatal to the process. Second, a post-project certification process would be necessary not only to satisfy the potential dangers of third party scrutiny by completing the authorization process, but also to help increase the knowledge capital of the proposal process itself. Third, the entire process was dynamic and was likely to be re-evaluated whenever necessary, to ensure that practice kept up with the multiple needs of decision makers, which included not only adequate data but also confidentiality concerns and related perceptions.

The notion of "Census benefit" may require some amplification, as it might differ from the statistical benefit required by other countries. For example, in the U.K.'s ten principles of protocol, access to confidential data is granted only "where it **will** [emphasis added] result in a significant statistical benefit".[5] This type of arrangement appears to require certainty of tangible success, but it may also include a type of benefit implicitly recognized by the flexibility in the IRS-Census arrangement. That is, to re-assure researchers that a fall from the "high wire" of Title 13, Chapter 5, predominant purpose attempted by ambitious projects would not necessarily be "fatal", IRS and Census agreed that a safety net of sorts would exist for all projects, especially those that failed to meet the criteria in their proposals but made a demonstrably good faith effort to do so. However, the good faith effort of failure needed to be documented, as did that of success, so that the future proposal process could use these outcomes as a learning device for both reviewers and prospective researchers.

4 Recognition of the Deteriorating Tradeoff

In the late 1990's Census became concerned about potential confidentiality problems in a previously released SIPP public-use file. These had been detected through analytical techniques used by a professional intruder whom Census had engaged contractually for just such a purpose. At the January 2002 conference, in which the book, *Confidentiality Disclosure, and Data Access Theory and Practical Applications for Statistical Agencies* was showcased and released by Census, Sweeney (2001) [6] presented some of her methods and how they might be used to re-identify survey respondents. Part of this methodology relied upon the possibility that variables on the public use file might also be individually identifiable in other publicly available datasets. In some respects at least, this event served as a type of catalyst for not only the current synthetic data approach for the SIPP/SSA/IRS public use file, but also for re-examining disclosure risk in the Federal statistical community.

Although the success of the new Census-IRS relationship was largely predicated on a more collegial attitude, it was clear at the outset that this could not be a co-equal partnership, as confidential data flowed only from the administrative data provider, IRS, to the data user, Census, and not vice versa. However, benefits did accrue. Partly as a result of the Sweeney (2001) work, IRS' own Statistics of Income Division decided to subject its public-use file, the tax model file based upon a sample of individual tax return filings, to such an examination and contracted with Sweeney's laboratory at Carnegie Mellon University for a professional intruder assessment of its confidentiality protections. In addition, because IRS approval of the synthetic data SIPP/SSA/IRS public-use file would be required (just as the Census RDC proposals

---

[5] P.7, Working Paper No. 11, *Contexts for the Development of a Data Access and Confidentiality Protocol for UK National Statistics*, Joint ECE/Eurostat Work Session on Statistical Data Confidentiality, Luxembourg, 7-9 April 2003.

[6] Latanya Sweeney, "Information Explosion," in *Confidentiality, Disclosure, and Data Access Theory and Practical Applications for Statistical Agencies*, North Holland, 2001.

required specific approvals) before its public release, IRS was also brought in by Census early in the process as a collaborator, not just a reviewer. If the synthetic data approach is successful at Census, it will help increase the utility to researchers of non-confidential tax data at the same time it reduces the need for access to confidential tax data, possibly even at Census RDC's where the beta testing will occur. Such a win-win outcome would benefit not only the confidentiality protection of administrative tax data but also the utility of researcher analysis for decision makers in both government and the private sector.

5 The Creation of a New Program

In late 2000, as both agencies began to resolve their differences with work on the Criteria Agreement, another Census-IRS crisis was brewing. Namely, a Census request to amend the Income Tax Regulations had been submitted in order to enhance Census estimates of poverty and income for the SIPP program. The detailed earnings items sought were also deemed critical for an emerging Census flagship program, the Longitudinal Employer-Household Dynamics study, which sought, among other goals, to track more closely employment flows in the U.S. economy. Both requests initially encountered opposition, but the justification for each emphasized the minimal need for FTI in these mandated uses. Eventually, the regulations were approved in February 2001, and immediately after work began on the SIPP/SSA/IRS PUF. It is ironic, but not coincidental, that the regulations were approved so soon after the Criteria Agreement's implementation in September 2000. That is, the process which had prepared both agencies for the Criteria Agreement, also galvanized them for purposes of these new proposed uses of FTI by making them focus on the criteria within the agreement as well as the protocols and process which would govern such access. It is also not a coincidence that one of the goals set forth in the Census justifications for the IRS regulations amendment, was the production of a SIPP public-use file, which was to include associated administrative data from SSA and IRS. The utility of this product was clearly seen as not only a predominant Title 13, Chapter 5, benefit for Census, but also a confidentiality boon for administrative tax data in general. However, without the items requested for regulation amendment, both SIPP and the potential robustness of the proposed LEHD program would have been seriously weakened. In fact, had the regulations items not been approved, it is likely that the LEHD program as it is known today would not exist. Had the Criteria Agreement, and even its early implementation not been developed as the SIPP and LEHD requests were prepared and later considered, it is possible, if not probable, that neither would have been approved.

6 Lessons and Recommendations

One consequence of the modern Census-IRS relationship is that the Criteria Agreement process undergone to protect confidentiality also laid the groundwork for

further legitimate access meeting these requirements; for example, the SIPP/SSA/IRS public-use file and the LEHD program described above.

Another lesson is that the record can probably be satisfied for posterity's perceptions of the past by ensuring that clear and sufficient documentation exists to explain those past intentions and actions.

The final lesson learned is that agencies must look outside themselves for the talents and skillsets needed to help them protect confidentiality and meet the needs for which confidential data are collected in the first place. In a time of dwindling budgets and competing priorities, such considerations are no longer options--they are imperatives.

In sum, one of the most important services that government agencies can perform is communicating to decision makers the need to learn the lessons above. If avenues are closed to such pursuits in the future, decision makers need to understand not only that their decisions will be based upon inadequate information--including its quality--but also that the imprimatur for intruding on the privacy of respondents-citizens will not exist. That is, the mandate for data collection will cease, but so will the ability of decision makers to lead and govern.