6

IRS Data, Data Users, and Data Sharing

Nick Greenia and Mark Mazur
Internal Revenue Service[1]

The importance of tax data to the federal statistical system, in both identifiable and anonymized form, derives from the fact that they are a national asset, a virtual treasure of information. Tax data are rich in both individual and organizational financial details that are useful in a wide variety of situations. First, these data underpin the administration of the federal tax system, which in turn provides the resources for many federal obligations. Second, tax data have almost as important a role as inputs to critical statistical systems that inform analysts and policy makers both inside and outside government. A critical issue for both the tax system and the federal statistical system is the determination of when a compelling need exists for identifiable tax data (often known as federal tax information or FTI) in lieu of aggregate anonymized data. The balance between the sometimes opposing interests of these systems is the focus of this chapter.

TAX DATA COLLECTED

The Internal Revenue Service (IRS) collects data for a variety of entities—covering over 130 million individuals and over 20 million businesses, tax exempt organizations, and governmental entities. The scope of tax return data, often including complete balance sheets and financial

[1]The views expressed in this paper are those of the authors and may not represent the official positions of the Internal Revenue Service or the Treasury Department.

statements, is vast and contains information on everything from net business profits to charitable contributions made by individuals. Moreover, the regularity of the data provided—annually, quarterly, and even monthly for some returns—and the fact that much of the data are captured electronically and for the universe of filers, makes FTI a potent resource for research and analysis.

The subject of business data can be a broad one, covering corporations, partnerships, and sole proprietorships and both employers and nonemployers. Intuition can be a poor guide for the types of data collected and available. For example, employer data are collected for part-time and full-time sole proprietorships, associated with individual tax returns, corporations, and partnerships, as well as entities that are not typically thought of as businesses, such as nonprofit organizations. Employment data themselves can be compiled at the employer level through the employment tax returns filed by businesses (for example the Form 941 series long used by the Census Bureau). They can also be compiled at the employee level and associated with the related employers through Social Security number/employer identification number (SSN/EIN) crosswalks (for example, using the SSNs and EINs captured from Form W-2, used to report annual wage and salary payments).

Typically, the IRS tracks business data at the EIN or enterprise level, but not at the establishment or place of business level unless they are one and the same. This practice differs from that used by most federal statistical agencies. Tax data accuracy is helped by the IRS compliance programs, including legal disincentives for noncompliance. Nevertheless, given the scope and frequency of the data processed by the IRS, the agency cannot ensure the accuracy of all items or the complete (100 percent) coverage of entities. That is, FTI faces limitations similar to those of data sets maintained by statistical agencies, so the tax data system per se should not be viewed as the panacea for statistical program deficiencies. As experience has shown, there will always be gaps and inconsistencies, even in relatively high-quality data sets.

PURPOSE OF DATA COLLECTION

Fundamentally, FTI is collected for use in administering the tax system, including tax policy analysis in the administration (the Department of the Treasury's Office of Tax Analysis) and Congress (the Joint Committee on Taxation). The IRS considers the successful administration of the tax system as highly dependent on voluntary compliance by millions and millions of taxpayers. In turn, voluntary compliance is seen as reliant on the protection—including the perceived protection—of taxpayer data confidentiality. Taxpayers share personal information with the IRS and are

assured that their personal data will be handled with the utmost care. The IRS believes that the tax administration purpose of FTI is paramount, and other uses of tax data, including statistical uses, must not interfere with that purpose.

The statistical usage of FTI is authorized by the statute (Section 6103(j) of the Internal Revenue Code, IRC) and associated Treasury regulations, which detail specific items and clarify the purposes for which access by parties outside the IRS may be granted. The tax code implicitly recognizes that statistical and administrative uses share common ground, in that both missions are dependent on high-quality data. In summary, there are two major goals for FTI. First, the data's confidentiality should be protected, so that voluntary compliance and the workings of the tax system are not harmed. Second, the data should be used effectively and efficiently for authorized purposes. It should be clear from these two goals that the role of the IRS with respect to tax data is less one of ownership than stewardship.

DATA USERS

The foremost use of tax data is administering the tax system and includes such functions as taxpayer account processing, audit and other compliance activity, research, and the compilation of statistics. In addition, FTI is provided, through the federal-state program, to state tax agencies in order to assist with states' tax administration needs. In fact, states account for the lion's share of FTI record disclosures to outsiders—in 2004, over 3 billion of the total 4.5 billion reported disclosures. However, the uses of tax data go well beyond that of tax administration, as the nation has long recognized their value not only for the formulation of tax policy and other program uses (such as Social Security) but also for statistical purposes. For the former purpose, tax data are used extensively by Treasury's Office of Tax Analysis and the congressional tax-writing committees—the Joint Committee on Taxation, the Senate Finance Committee, and the House Ways and Means Committee. Other congressional uses include oversight work undertaken for a tax-writing committee, for example, by the U.S. Government Accountability Office.

For a handful of federal entities listed in the tax code, selected identifiable tax data—by no means all items—are provided for statistical purposes. These consist of the Bureau of Economic Analysis (BEA), the Census Bureau, the Department of Agriculture's National Agricultural Statistical Service (USDA-NASS), and the Congressional Budget Office (CBO). The Census Bureau accounts for most of the statistical-purpose FTI record disclosures: over one billion in 2004. Tax data are also broadly used in statistical nonidentifiable form (usually tabulations) to assist

other entities, such as businesses, policy think tanks, federal agencies not authorized to receive identifiable data, and academic researchers.

AUTHORIZATION PROCESS FOR ACCESS

Every access to FTI, even within the IRS, must be authorized by statute, meaning that legislation has been codified as part of Title 26 of the United States Code. Furthermore, the statute requires that only the minimum amount of authorized FTI be provided for accomplishing a given authorized task. These constraints apply to all users, including the IRS. For example, virtually all access to FTI within the IRS is authorized for the purpose of tax administration, which is multifaceted, under Section 6103(h)(1) of the IRC.

As might be expected, given the sensitivity of FTI, the law governing access to confidential or identifiable tax data, especially for statistical purposes, is restrictive with respect to both access and use. Thirty years after the tax code was overhauled with the 1976 Tax Reform Act, the number of entities with statistical access to FTI can still be counted on one hand: the Census Bureau, BEA, CBO, and USDA-NASS.[2] Based on the statutory record to date, it seems clear that Congress has regarded any expansion in access to FTI for statistical purposes as deserving of cautious and comprehensive consideration. Unsurprisingly, the rate of change has been glacial, primarily involving USDA-NASS, when the Census of Agriculture was transferred to that agency from the Census Bureau in the late 1990s, and CBO soon after, with its statutory addition for the purpose of long-term modeling of Social Security and Medicare. Even in these instances, however, the historical precedent provided some reassurance regarding the entry of these two new members to the FTI club. Working as special sworn status individuals in the Census Bureau, NASS had conducted much of the previous agriculture censuses. Similarly, CBO had long-standing experience in handling FTI as an agent for the Joint Tax Committee under Section 6103(f)(4) of the IRC. Thus, neither was a novice with regard to either FTI or the associated culture of confidentiality that FTI access requires.

Adding statistical users or increasing access for current users of FTI means Section 6103(j) must be amended; that is, a new law must be passed. Thus, the first requirement of any data-sharing proposal entailing access

[2]The Federal Trade Commission's inclusion in this statute is vestigial, as its Quarterly Financial Report function was transferred to the Census Bureau in the mid-1980s. Although Treasury is also listed in the statute, virtually all of its FTI receipts are authorized by Section 6103(h)(1) as being related to tax administration.

to tax data is that the agencies proposed for sharing data must all be in the tax statute or Section 6103(j). This is a necessary but not a sufficient condition, as the agencies must also share statutory authorization to receive the same types of data—for example, corporate total income and individual investment income. In addition, applicable regulations, which require formal approval by the Treasury Department's assistant secretary for tax policy, may also be needed to authorize access and use of the same tax items for the agencies statutorily enabled to receive FTI. Treasury regulations may not only list the specific item content an agency is authorized to receive but also stipulate a more focused purpose. Regulations can also be amended to remove items that are no longer needed by an authorized recipient. In fact, need and, in particular, the requirement of providing only the minimum amount of data needed to accomplish a compelling agency task, is a bedrock principle used for determining not only the necessity of a new statute, but also a regulation amendment. Historically, both Congress and the Treasury Department have required that a compelling data-driven case be made for amending either statute or regulation, although clearly, it is more difficult to amend the statute.

Occasionally, policy agreements crafted by the Treasury Department or the IRS and one of the statistical agencies may be used to supplement the statute and regulations. For example, the IRS-Census Bureau Criteria for the Review and Approval of Census Bureau Projects that Use Federal Tax Information (sometimes called the Criteria Agreement) has been used to delineate and clarify the process under which FTI may be accessed for new Census Bureau purposes, especially authorized research purposes at the Census Bureau as part of their Research Data Center arrangement.

CHALLENGES IN PROTECTING CONFIDENTIALITY

Protecting the confidentiality of tax data is challenging for the IRS, especially because there is no statute of limitations and because the tax code treats all FTI the same with respect to confidentiality protection. That is, to the IRS, a business name or address is as deserving of confidentiality protection as income items for a large corporation's or individual's tax return, and all must be protected in perpetuity, even after they have been anonymized as statistical tabulations for public release. Given these constraints, the resource consequences of safeguarding taxpayer confidentiality over time are nontrivial. These constraints are exacerbated by the potential for complementary disclosure, or the reidentification of taxpayer data using indirect means, for example, using data in other publicly released data to identify FTI related to a particular taxpayer. Given the ever-increasing public releases of tax and other data, the task of protecting FTI is daunting, especially over time.

CONSTRAINTS ON USING DATA

As indicated earlier, the access of FTI must be only for purposes authorized by statute, possibly supplemented with regulations and, infrequently, policy agreements. In addition, authorized recipients are subject to regular safeguards reviews in order to confirm their understanding and implementation of the many requirements covering physical and computer security, data need and use, and appropriate documentation. Other requirements include separate systems for processing or accessing FTI and background checks on individuals accessing it within facilities certified for such purposes. All these requirements are intended to preserve the confidentiality of FTI, whether maintained in its original form or commingled with data from other sources. In addition, the penalties and fines for unauthorized disclosures or inspections (also known as browsing) can be severe and are detailed in the tax code.

All of these constraints are largely driven by concerns for taxpayer confidentiality, and, in general, they seek to control or regulate the use of tax data by conceptually limiting, physically confining, and tracking such access in order to provide a documented audit trail that will withstand outside or third-party scrutiny. Implicitly, both the IRS and Congress recognize that this approach does not guarantee complete confidentiality, as the only means for such an assurance would be not to release any data at all. However, padlocking the treasure of tax data is viewed as neither a desirable nor a viable outcome, so some disclosure risk is accepted as part of the necessary balance of protection and access. The challenge is to identify acceptable risk, and the approach utilized to date is taking steps that prevent reidentification of tax data through "reasonable means." The interpretation of reasonable means includes the use of reasonably available computer technology, mathematical/statistical techniques, and a working knowledge of the related subject matter. The reasonable means standard attempts to avoid system meltdown in the use of FTI. "Reasonable means" is a technology-relative concept and thus, it may be a moving target. Nevertheless, it represents an attempt at balancing the two goals for tax data: their protection and their effective use.

It seems clear that there is probably some overall limit on tax data access, even if that limit is not precisely known. The need for this limit can be attributable to both resource costs of protection and what might be termed as the perception of a plausible quantity limit on access. To see why such a limit makes sense, consider that even large amounts of safeguarding resources cannot enable unlimited access to FTI. The reason is credibility. It is simply not credible that unlimited access would ever pass a perceptions test on confidentiality protection, especially for third-party scrutiny. That is, such an outcome would not seem plausible, as it would seem to turn the very concept of confidentiality on its head.

SOMETIMES CONFLICTING MANDATES

Statistical agencies such as the Census Bureau are mandated to use existing data systems (especially administrative records) to the maximum extent possible. Combined with the IRS statutory mandate to provide FTI only to the minimum extent needed for authorized purposes, these mandates create a tension that drives a need to negotiate the appropriate amount of FTI accessible for a given statistical task. This is not to say that the relationship between tax agency and nontax agency needs to be very adversarial. In effect, access to FTI should be treated as a scarce resource. Accordingly, the opposing mandates create an initial starting point that requires interagency cooperation in order to find a welfare-improving outcome, in which both parties find it in their interest to move to this new point. Thus, while tension from the conflicting mandates may be viewed initially as a problem, it is probably necessary to ensure the protection of taxpayer confidentiality and the provision of FTI only to the extent necessary for compelling statistical needs. Without such tension, there would probably be some bias—either too much access or too little. With this constraint, the IRS and the authorized statistical agencies are compelled to bargain hard toward an equilibrium that upholds their respective mandates, and that both sides are willing to defend. Ultimately, any interagency agreement must be documented in order to be clearly implemented and to successfully withstand outside scrutiny. Thus, the conflict in mandates provides a type of pricing mechanism for achieving a balance between supply and demand for FTI access. Forces likely to continue exerting pressure on this mechanism would include declining survey response rates, statistical processing costs, response burden, and, of course, the need to maintain voluntary tax compliance by protecting taxpayer confidentiality.

THE CENSUS BUREAU-IRS CRITERIA AGREEMENT

A 1999-2000 IRS safeguards review of the Census Bureau raised concerns over access by its research data centers (RDCs) to FTI, especially from the perspective of statutorily authorized purpose. As a result, the Census Bureau and IRS agreed to the terms of the coauthored Criteria for the Review and Approval of the Census Bureau Projects that Use Federal Tax Information, effective September 19, 2000. The Criteria Agreement outlined protocols and other requirements governing access to FTI for new uses by the Census Bureau, especially for RDC projects. The IRS review role, assigned to the Statistics of Income Division, consists of approving or concurring on the predominant Title 13, Chapter 5, purpose of proposed projects and ensuring that the minimal FTI needed would be accessed for a given proposed usage. Scientific merit remained the prov-

ince of the Census Bureau and the researcher community, and, for purposes of the Criteria Agreement, outside researchers were treated as Census Bureau employees (under the special sworn status designation). The Criteria Agreement can be seen as a good outcome for the opposing agency mandates governing access to FTI for statistical purposes. As a policy agreement, it established an explicit interagency standard for authorized purpose that supplemented long-standing statutes and regulations in adapting to changing user needs. A cornerstone of the agreement was its emphasis on proposal review documentation, including explicit dual agency approvals on both the project proposals and post-project certifications. In addition, the review process it fostered implicitly enlisted the research community's active participation by forcing it to develop and maintain review capital and adding to the interagency appreciation of confidentiality needed to make this process viable. Such an outcome recognizes the limited review resources available in both the Census Bureau and IRS and was essential in order to promote a viable flow of projects. The process also made all three participants—the Census Bureau, the IRS, and the researcher community—aware of the need to work together in order to make the process demonstrably credible for purposes of potential third-party scrutiny.

DATA-SHARING PROPOSALS

For three decades, the federal statistical community has attempted to overcome certain deficiencies—particularly list frame coverage across agencies—of the decentralized data collection system by submitting a number of proposed statistical data-sharing bills to Congress. These usually died in committee after being introduced, and many required an amendment to the tax code, due to the importance of tax data in these proposals, and because FTI for business is inextricably commingled with non-FTI on the Census Bureau's business register. As a result, a number of the data-sharing proposals were accompanied by companion tax bills, known as "J bills," due to their proposed amendment of Section 6103(j). In the 1990s, two of these proposals addressed both demographic and business data and encompassed all 10 major statistical agencies. Neither was enacted.

The most recent data-sharing legislation, part B of the Confidential Information Protection and Statistical Efficiency Act (CIPSEA), focused on sharing only business data and was restricted to the three major business data statistical agencies: BEA, the Census Bureau, and the Bureau of Labor Statistics (BLS). A companion J bill accompanied CIPSEA when it was introduced in Congress in July 2002. This strategy led to the enactment of partial (nontax) data sharing when CIPSEA was signed into law

in December 2002. However, the accompanying J bill attracted less legislative support, never made it to the floor of the House, and expired with that Congress.

From this experience a number of lessons have emerged, some of which CIPSEA had already absorbed, and some of which, with hindsight, might have led to some differences in both approach and content.

LESSONS LEARNED

Strong Leadership Is Needed

Strong leadership was provided for CIPSEA throughout the 2002 effort by both the Council of Economic Advisers and the Office of Management and Budget. More advocates were probably needed, especially in Congress, in order to advance both (tax and nontax) parts of the proposal once it arrived there. In fact, support by members in both the House and the Senate would seem critical in order to overcome concerns about increased sharing of identifiable tax data. In a similar vein, the support of congressional staff—especially on the tax-writing committees—needs to be enlisted, with a compelling case on why the bill is needed, including not only how government statistical operations would be improved but also how taxpayer confidentiality would remain protected.

Dispell the Myth of Access Being a Zero-Sum Activity

Part of the education effort needed would be well spent focusing on the myth in the tax community that expanded access to FTI is undesirable in general and that access cannot be expanded in one area without a commensurate reduction somewhere else. For example, the notion that increasing the number of agencies accessing business data can only be accomplished at the expense of reducing another agency's existing access to FTI must be countered. One strong argument countering this position is the evidence provided by adding both USDA-NASS and CBO to the statute as authorized recipients of FTI without reduction in access elsewhere and without observable problems in terms of weakened confidentiality. Another counterargument might be the controlled expansion in access enabled by the Census Bureau-IRS Criteria Agreement for the RDCs—now in its sixth year of implementation.

Discrete Steps May Be Better than Bold Leaps

Concerns articulated by some opponents of the 2002 CIPSEA effort include the notion that the J bill's expanded access seemed too broad be-

cause (1) it was modeled on the Census Bureau-IRS Criteria Agreement, which included access by researchers at the Census Bureau RDCs, and (2) the statutory language referenced regulations to be released in the future for purposes of authorizing access to specific items of FTI. The two major purposes of sharing FTI for Part B of CIPSEA have largely been described as establishing a common business list frame for all three agencies (BEA, the Census Bureau, and BLS) and in providing common identifier information that would enable the three agencies to exchange nontax data with each other. The potential for excessive access to tax data under the CIPSEA J bill was a concern, given that expanding item access via regulation would require only Treasury approval, not an act of Congress. This concern may have been heightened by the interest BLS has indicated in sharing limited FTI with its state partners. Moreover, both BEA and BLS might some day want to pursue access arrangements with researchers in a way similar to what the Census Bureau has done with its RDC model, which would increase the number of persons with FTI access. In short, these concerns were raised about the ability to limit, track, or control access to FTI and should be addressed by any future J bill.

One possibility for assuaging such concerns is to stipulate a limited amount of FTI in the statute itself, obviating the need for regulations. The items themselves might be limited to, say, taxpayer identification number (TIN), name, address, industry code, and one or two magnitude variables, such as employment size and income, for the purpose of stratifying a sample. Listing in the statute only the items needed for purposes of addressing the central problem (i.e., mutual list frame coverage, exchange of nontax data) might help emphasize the agencies' good faith effort to request and use only the items justified by the data-sharing rationale, so that the principle of the minimum FTI needed would be met. In addition, such a statutory limitation might help signify that these agencies did not intend to replace surveys with FTI per se, an argument sometimes raised by opponents of expanded access.

Show Some Benefit to the Treasury

When CIPSEA and the J bill were introduced to Congress in July 2002, staff from both the House Ways and Means and the Senate Finance Committees raised questions about how the legislation, especially the data sharing enabled by the J bill, would benefit the Treasury Department. CIPSEA seemed to contemplate a one-way flow of data for statistical purposes, which did not appear to include statistical tax analysis conducted by the Office of Tax Analysis in the Treasury Department and the Statistics of Income Division at the IRS. Thus, the general consensus seemed to

be that Treasury would not directly benefit much by CIPSEA and the accompanying J bill.

One way of addressing this concern might be for Treasury (and other outside analysts) to benefit from the creation and release of more public-use files, including those created with synthetic data. Two problems attend this recommendation. First, the jury is still out on the utility of public-use files created with synthetic data. Second, virtually no public-use files of business data exist due to the difficulty of masking the interesting data features of concentrated industry activity at the same time that these properties are needed for analysis. A more direct way to provide Treasury with analytical benefit might be for the Census Bureau to consider Treasury and the IRS researchers for RDC access on meritorious project proposals, as long as they adhere to the same requirements as other researchers, including predominant Title 13 purpose. Preliminary discussions between the IRS and the Census Bureau so far indicate that this might have value.

One possible objection to this idea pertains to Section 7214(a)(8) of the tax code. This statute requires Treasury and IRS employees with evidence of revenue law violations to report it. The Census Bureau's concern is that this obligation might overshadow the confidentiality oaths required for special sworn status at the Census Bureau. Several factors might help mitigate this concern. For example, the statute's evidentiary standard on what constitutes information that a revenue violation has occurred is high, and it is unlikely to be uncovered during the sort of research and analysis that Treasury or the IRS might propose, especially given the limited data related to actual tax liability available at an RDC. An additional point is that any Treasury or IRS researcher would most likely be intent on statistical research. Enforcement personnel, such as auditors or tax examiners, would not be likely candidates for access, so there would be little emphasis on case-by-case compliance issues. This standard would be consistent with the engagement of some researchers at RDCs who represent commercial enterprises with a variety of clients. That is, the suspension of a non-RDC allegiance for purposes of data access is hardly unprecedented in the Census Bureau's RDC experience.

Interagency Disclosure Coordination Needed

As additional assurance to reviewers of a future J bill, it may be advisable to consider making explicit, perhaps in the narrative accompanying the bill, that the agencies authorized to share FTI would collaborate on statistical disclosure limitation methodologies. It is probably important that such coordination be given clear prominence in the J bill itself, given that the statute (Section 6103(j)(4)) requires that both direct and indirect

means of reidentification be prevented with any public release of data. All FTI, including the building blocks of any list frame, such as name, address, and TIN, will probably remain subject to perpetual protection under provisions of the tax code, as discussed previously. Accordingly, legislators who understand that the statistical community has the issue of taxpayer and respondent confidentiality foremost in mind as it seeks expanded access to confidential data may be more sympathetic to a future J bill. Demonstrating such care and foresight may also assist with any future proposals that might expand data sharing beyond three agencies and encompass more than business data.

CONCLUSION

We view FTI as a national asset that can have great value in many situations faced by the statistical community. However, this asset comes with numerous constraints on its use, in particular, a strong emphasis on taxpayer confidentiality and a requirement that only the minimum amount of FTI be provided to meet authorized uses. These constraints can be productively addressed through good faith bargaining between the IRS and the statistical agencies authorized to receive FTI. We believe this bargaining process is a useful way to strike the right balance between needed access to FTI and concerns for taxpayer confidentiality. Future expansions of the statutory provisions allowing access to FTI are possible, but will take a concerted effort by the affected federal statistical agencies. Learning from past efforts can help increase the chances of success in this endeavor.