The Release of IRS Data: Challenges and New Approaches

Nick Greenia
IRS Statistics of Income Division
July 8, 2002

This paper was presented at the 2002 IRS Research Conference.
The views expressed in this paper are the author's and not necessarily those of the Internal Revenue Service.

I. Introduction

Statistical agencies have become increasingly aware that two new challenges may seriously affect their ability to release data into the public domain. Declining costs 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.

The Internal Revenue Service (IRS) faces additional challenges. Tax data have always been particularly susceptible to re-identification, both because of their relatively widespread distribution and because of their sensitive content. In addition, because publicly and privately available datasets are often directly based upon entities also in the tax system there is more potential to match to tax data and re-identify taxpayers. But IRS also faces new challenges. In particular, the increase in the number of tax returns that are electronically transmitted or filed in magnetic format has increased the quantity and quality of data held internally, and thus the potential for an even faster release of new data products. In addition, the decentralization of IRS's statistical research functions raises new potential for complementary disclosure if different areas release new products without closely coordinating such releases with each other.

Finding solutions to these challenges will be a major issue as IRS continues to accelerate its transformation to a producer and provider of data. This paper provides some background on the current situation, explores some of the challenges in more detail, and suggests some possible approaches to the problem.

II. Background

While there are issues surrounding disclosure of confidential data that are common to all federal statistical agencies,[1] IRS also has its own idiosyncratic issues. All confidential tax data, also known as Federal Tax Information (FTI), are treated homogeneously when it comes to disclosure; FTI has a wide variety of uses, including statistical; and the law governing the release of FTI is particularly stringent. Each of these then affects the way in which FTI can be accessed and released.

The homogeneous treatment of FTI is a result of the Internal Revenue Code (IRC), which does not allow IRS to make distinctions among data elements in FTI - even as to data age. That is, there is no statute of limitations as there is for

[1] An excellent review of the core issues facing statistical agencies is provided in the CDAC brochure "Confidentiality and Data Issues Among Federal Tax Agencies" (http://www.fcsm.gov/cdac/brochur10.pdf) and a good survey of the technical issues in a recent book "Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies" by P. Doyle, J. Lane, J. Theeuwes, and L. Zayatz, Elsevier, 2001.

confidential microdata at statistical agencies such as the US Census Bureau (72 years for demographic data, 30 years for business data). In addition, the IRC does not distinguish among different types of data or taxpayers, so that the Social Security Number of 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[2] - is treated and protected in perpetuity as equally confidential. This task of protecting confidentiality, given the amount of data for which IRS becomes responsible over time, is expensive and technically challenging.

The uses of tax data go well beyond that of tax administration. The nation has long recognized their value for not only the formulation of tax policy and other administrative uses (such as Social Security) but also statistical purposes. Tax data, mostly in statistical non-identifiable form, are broadly used to assist the nation's decision makers both within the federal government such as Congress and the Administration and outside the federal government such as businesses, policy think tanks, academic research groups, and state entities. This widespread use is due to a number of characteristics unique to the tax data system:

1. The size of the tax reporting population—over 20 million businesses, tax exempt organizations, and government entities and over 125 million individual taxpayers;
2. The scope of return data—complete balance sheets and financial statements as well as a plethora of information on everything from mortgage interest and property taxes to charitable contributions; and
3. The regularity of the data provided—quarterly, annually, and even monthly; and a compliance program to "assist" with the validity and timeliness of the data reported.

These facets make the federal tax system not only invaluable for the function it performs in collecting the funds to administer the government's many programs and services, but also a national treasure of vital information about the nation.

As might be expected, given the size and usefulness of FTI, the law governing access is extremely restrictive both with respect to access and to use. Thus, for example, even though the release of FTI to the Department of Agriculture (USDA) is authorized under IRC section 6103(j)(5), the FTI can be used only for purposes of conducting the Census of Agriculture. Traditionally, the association

[2] A crash 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.

of any FTI, including fact of filing or non-filing[3], with an identifiable taxpayer constitutes disclosure, which can be authorized or unauthorized. For example, authorized disclosures for statistical purposes are covered in section 6103(j) of the IRC, but unauthorized disclosures are confidentiality violations and subject to both fine and imprisonment as outlined in sections 7213, 7213A, and 7431.

As a consequence of these factors, the IRS provides disclosure-proofed data in one of two forms: tabular data or public use files. However, even these disclosure-proofed data, also known as "sanitized" data, must be produced for statutorily authorized purposes. Most often, such publicly released data have been produced by IRS's Statistics of Income Division (SOI) under either section 6108(a) or 6108(b) of the IRC. Section 6108(a) generally covers the statistical products the Secretary of the Treasury is actually required to publish based upon tax returns filed and 6108(b) covers statistical products produced as a result of outside requests from users without access to the FTI necessary for such a product. Thus, two IRC sections seem to predominate in the release of FTI for statistical purposes; namely, 6103(j) for external access and 6108 for internal access. However, all data released into the public domain under either section must also meet an anonymity standard. Section 6103(j)(4) and section 6108(c) state an absolute standard: that the data released cannot "be associated with, or otherwise identify, directly or indirectly, a particular taxpayer." Information released must be anonymous so that taxpayers cannot be identified whether the product is a table provided on the Internet or a public use file. Note that the anonymity standard does not specify the methodology for accomplishing this goal.

III. Interpreting the Law

The general standard of anonymity provided by IRC 6108(c) is translated into specific form for statistical data required for release by internal access (i.e., by SOI) for either 6108(b) or 6108(a) purposes as the administrative rule of 3, which requires each cell of tabular data to be based upon at least 3 returns. For data below the state level, this becomes a rule of 10. These administrative rules of 3 and 10 cover direct and indirect or so-called derivative disclosures, so that if financial data are produced for both corporate returns and only those with net income, disclosure must also be prevented for the implied table of corporate returns without net income.

The statute is silent on whether one cell of disclosed data is more problematic than another; i.e., it is implicit that disclosing the net receipts of General Motors is just as important to protect as the filing status of an individual. The anonymity standard is simply indiscriminate and absolute in requiring that all data be released in anonymous form. The anonymity requirement also applies to data released by outside users authorized to receive FTI under 6103(j), but while the general standard applies, the actual disclosure protection methodology is not

[3] IRS is not authorized to disclose whether a taxpayer did or did not file a return.

specified. The requirement is simply that whatever methodology is used be as good as that employed by IRS; namely, the rules of 3 and 10.

The question which must accompany any methodology attempting to meet the anonymity standard is: from what sort of intrusion must we protect the data? Must it be absolutely impossible to re-identify a taxpayer using any means available, or is there some less rigid methodological standard? The answer traditionally is that the data must be protected from potential intruders who, using "reasonable means", might attempt to make such a re-identification. Reasonable means includes 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 we can probably imagine in 2002, is that the concept of reasonable means is a technology-relative concept and may be a moving target too elusive to be relevant.

IV. The Challenges

Tax data clearly have an important role in enabling the nation to administer its tax system, formulate its tax policy, and otherwise inform decision makers in many other sectors of government. In short, tax data plays an essential part in ensuring good government through informed decision making. However, the protection of citizens' confidentiality is also a part of good government. Optimizing the utility of tax data for users while at the same time protecting the confidentiality of taxpayers' data requires recognizing, and addressing, the key challenges facing IRS.

i) Wide variety of users

Informed decisions require information based upon the best data available, whether the data are derived from administrative record populations or samples from the federal tax system, records obtained through statistical agency surveys and censuses, or both. Increasingly, the specific nature of the questions asked has led to the creation of specifically tailored datasets, released to a large variety of recipients. The following is an incomplete list of FTI recipients:

- IRS internal functions such as Collection and Audit;
- IRS Statistics of Income Division;
- IRS Research functions;
- Child support agencies;
- State tax agencies;
- Federal statistical agencies (Census Bureau, Bureau of Economic Analysis (BEA), National Agricultural Statistics Service (NASS));
- Congressional entities (Joint Committee on Taxation (JCT), Congressional Budget Office (CBO), General Accounting Office (GAO));
- Social Security Administration (SSA);

- Treasury Department.

This number of different FTI recipients combined with the number of different ways the same or similar FTI records may be used to answer similar questions with both published and unpublished reports gives some idea of the challenge faced by IRS in protecting a given taxpayer's confidentiality. The task is even more daunting when we remember that the anonymity standard of sections 6103(j)(4) and 6108(c) is absolute, the standard for methodology is reasonable means, and it is up to each FTI recipient to determine a methodology that satisfies these criteria.

At least two questions now seem reasonable to raise:

1. How sure are we that confidentiality is protected in the data already publicly available?
2. How do we release data in the future, given what is already out there?

If the answer to the first question is less than certain, the answer to the second is probably not much different.

ii) Commercial datasets based on tax identifiers

The private sector creates data systems for commercial purposes (Dunn & Bradstreet, Compustat, and others). These datasets are publicly available at the micro record level and include such items as name, address, EIN, number of employees, and a plethora of financial information at the company level. In addition, there is substantial information about the sampling frame used for statistical data collection and about statistical disclosure limitation (SDL) techniques. The potential for individuals to use these data and their knowledge of SDL techniques in conjunction with datasets produced from FTI, threatens the absolute protection requirement of sections 6103(j)(4) and 6108(c), especially when softened by the corollary of reasonable means.

iii) Electronic Filing

For several years Congress and the Administration have exhorted IRS to modernize not only its computer systems, but also its system of tax data collection. The old system in which taxpayers file hardcopy tax returns whose information is then laboriously and expensively keypunched into master file computer records is supposed to be replaced eventually by a system in which returns are filed and processed from start to finish in electronic format. The benefits could be many, including fewer human errors introduced during data processing, and possible processing cost reductions. From an information perspective, electronic filing (known within IRS as E-File) provides the potential for a veritable cornucopia of data. Data quality is likely to be improved, since the data themselves do not have to be transferred from the taxpayer's record to the computer record involving the traditional and separate stages of transcription, testing, and error correction. Thus, the entire E-File record's data would be in computer record form, and, given sufficient computer resources, a complete population of records could theoretically be tabulated and analyzed in the aggregate almost immediately.

The existence of these E-File data increases the potential for the generation of more information. A report, for example, on the number of taxpayers with mortgage deductions classified by adjusted gross income and nine-digit ZIP Code, is less likely to rely on sample information, with the associated input of mathematicians and statisticians, since universe information will readily be available. The likely concomitant production of many different sets of reports raises a number of associated confidentiality issues – particularly the importance of centralized coordination and control of data dissemination.

V. Recommendations

Every year IRS is required to provide Congress' Joint Committee on Taxation (JCT) an accounting of authorized FTI disclosures made during the year. These counts are categorized by recipient; e.g., the Florida Department of Revenue, Census Bureau, GAO, etc. The input reports from IRS to JCT also provide some explanatory information regarding the type of information disclosed; e.g., size of employment and total compensation reported on employment tax returns. Congress and the nation then get a very big picture of where confidential tax data are going and to whom, not to mention a notion of the enormous data stewardship job that IRS has to perform. For calendar year 2001 over 3 billion accountable FTI disclosures were reported.

Given the challenges outlined previously in this paper, one recommendation might be to initiate an expanded accounting system, based upon that developed by JCT, for the same authorized recipients of the reports and for the other products produced from FTI that might find their way into the public domain. Such an expanded accounting system might consist of the following elements:

i) A coordinated system of control for all FTI users that not only records such events, but also takes past releases into account before allowing a future data release.
ii) A regular expert assessment of disclosure risk in these products, both cardinal and ordinal in nature. That is, we need to know—again, on a systematic basis--not only how many taxpayers might be re-identified but also the information which can be associated with such re-identifications and thus, the significance of such unauthorized disclosures, perhaps even scored as to potential importance. For example, knowing that IBM filed a tax return might not be as significant as knowing the complete income and expense statement detail on IBM's tax return. In other words, we may need to analyze the disclosure risk and start making quantifiable, or at least comparable, distinctions represented by these different risks.
iii) A regular examination of alternatives to the current product system of exact tabulations and public use files explained by precise sample and methodological

descriptions. We need to be prepared for the possible conclusion after our disclosure risk assessment that it may not be prudent to continue either the level of disclosure risk or the level of significance it represents. This, in turn, means that we will need to consider alternative data access arrangements, depending on analytical need.
iv) A survey of our constituents—the taxpayers themselves—as to how they feel about what we do. We may discover that businesses are less concerned about the protection of some tax items, especially if it means a reduction in reporting burden—whether regulatory, administrative, or statistical. We may even discover that individual taxpayers think we are already sharing their tax data with every federal agency requesting them. But one thing is likely: by gathering such information, we will become better informed to make the decisions necessary to administer the federal tax system, especially its data.

The following is a list of action items suggested for consideration in order to accomplish these objectives.

Catalogue and Categorize FTI

All FTI in the public domain—whether for a special outside request or annual program, tabulations or public use file--need to be documented for inventory purposes, probably categorized by tax year, type of taxpayer (business, individual, etc.), type of FTI item, and releasing entity (IRS, NASS, Census, JCT, SSA, etc.). This would permit the establishment of a system of records that could form the basis for a coordinated system of control to determine what can be released in the future, as well as for assessing current disclosure risk.

Assess Risk of Unauthorized Disclosure

Contract with professional "intruders" to conduct studies assessing the risk of disclosure in publicly available FTI. The risks studied should include, but not be limited to, taxpayer re-identification and FTI that can be associated with such identifications, and should distinguish by type of taxpayer (business, individual, etc.).

Assess Harm and Perceptions concerning Disclosures

Does harm or damage resulting from an unauthorized disclosure differ by type of taxpayer, type of FTI, age of FTI, and if so, how? For example, is the revelation of business proprietary information such as a company's current year sales and profits as damaging as knowing the same company's employment size ten years ago? Conduct surveys of taxpayers and users—both analysts and decision makers—to help with this determination.

Partner with Data Users

Coordinating data protection standards, including self-policing, with data users such as researchers could help ensure the continuous infusion of both techniques and communication systems for information on existing disclosure risks which need to be addressed.

Centrally Coordinate and Control FTI Release

IRS may need to consider the formation of a central FTI Disclosure Review Board, probably coordinated with other FTI recipient agencies in the federal community, especially the federal statistical community. This board would help to determine policy and procedure regarding the access to and release of products containing FTI. This process would probably go beyond what is currently required.

Consider Alternatives to Current Access and Release

After cataloguing FTI, coordinating their future release, and assessing disclosure risk and harm, it may be determined that current procedures and policy are no longer viable or perhaps not the most effective way to meet user need at the same time confidentiality is protected.

Other alternatives could be investigated—notably, masked datasets and products, as well as statistical research centers.

1. Masked Data Sets and Products

It may be that actual data can no longer be released as either aggregate tabulations or public use files, except in very broad classifications. New research by Abowd and Woodcock[4] suggests that core microdata can be transformed in such a manner that it will not only prevent the re-identification of a taxpayer but also preserve the essential data relationships at the micro level which are so important to many analysts producing information for decision makers. Such techniques should be studied with an eye towards using them on a systematic basis.

Under this access arrangement, users might also be required to sign pledges not to attempt to re-identify taxpayers.

2. **Statistical Research Centers**

These inter-agency sites would require the participation of statutorily authorized FTI recipients in a consortium dedicated to providing access to analysts who must have access to actual microdata records (not restricted to FTI) for mandated purposes. A single agency is unlikely to be able to fund and operate such centers. In addition, the control and efficiency offered by an inter-agency approach would help to standardize protection procedures at the same time improving datasets with common items collected across respondents. Thus, in addition to optimizing the data utility afforded users and confidentiality protection for respondents, it is possible that plans for data collection could benefit from such coordination in the future through the reduction of reporting burden and efficiency improvements in data collection systems.

[4] John Abowd and Simon Woodcock "Disclosure limitation in Longitudinal Linked Data," in *Confidentiality, Disclosure, and Data Access Theory and Practical Applications for Statistical Agencies*, North Holland, 2001.