WP. 21
ENGLISH ONLY

UNITED NATIONS STATISTICAL COMMISSION and ECONOMIC COMMISSION FOR EUROPE CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION STATISTICAL OFFICE OF THE EUROPEAN COMMUNITIES (EUROSTAT)

Joint UNECE/Eurostat work session on statistical data confidentiality
(Bilbao, Spain, 2-4 December 2009)

Topic (iii): Research data centres and virtual labs

A NEW EXTERNAL RESEARCH PROGRAM AT THE U.S. TAX AGENCY'S STATISTICAL OFFICE: BACKGROUND, CHALLENGES, POSSIBILITIES WITH FEDERAL TAX DATA

Supporting Paper

Prepared by Nick Greenia, Internal Revenue Service, United States of America

A New External Research Program at the U.S. Tax Agency's Statistical Office: Background, Challenges, Possibilities with Federal Tax Data

Nick Greenia[i]
U.S. Internal Revenue Service
Statistics of Income Division, Director's Office

Abstract: Since the U.S. Constitution's 16th Amendment authorized Congress in 1913 to enact an income tax, the federal tax system's profound influence upon the economic behavior of workers, businesses, and non-profit organizations has become indisputable. As a result, economic analysis based upon tax microdata has developed into a critical tool for tax policy purposes, mostly as the preserve of the Treasury Department's Office of Tax Analysis (OTA) and Congress' Joint Committee on Taxation (JCT). Although additional economic analysis has also been conducted at the few agencies authorized to receive tax data, such as the Bureau of Economic Analysis, external researcher access has been largely restricted to tabular data and a limited number of public use files. In recent years, tax microdata access has been expanded for external researchers via the U.S. Census Bureau's Center for Economic Studies (CES), yet the number of variables available is small, and since 2000 the number of CES projects using tax data is not much more than a hundred. In order to better fulfill its data dissemination obligation as a statistical office and to increase its own human analytical capital, in 2008 the Internal Revenue Service's (IRS) Statistics of Income Division (SOI) instituted an external researcher program seeking high quality research proposals that required access to confidential tax microdata. The tax data available for such projects are much richer than those available to any user other than OTA and JCT, and extend to approximately 10 years of population data for individuals and businesses, as well as more years of the even richer data content of SOI sample files containing the universe of high income entities. This paper reviews the program's background, including SOI's portfolio approach to data dissemination, and describes the program's purpose, as well as challenges to its continuation. It concludes with thoughts on possible options for the future.

Background

Utility of Tax Data

Confidential Federal Tax Information (FTI) is a national treasure, but one that is underutilized. The U.S. Internal Revenue Service (IRS) annually collects population data[ii] for a variety of entities: over 130 million individuals, including sole proprietors; over 50 million businesses[iii]; over 800,000 tax exempt organizations; and related entities, such as approximately a million employee benefit plans[iv]. Detailed balance sheets and income statements provide the means for tracking and analyzing the economic stocks and flows of businesses, even the smallest entities that some analysts might consider as hobbies and not "true" businesses. (Nevertheless, these smallest entities might prove ideal for analyzing start-up entrepreneurs, particularly for studying the life cycle of innovation; e.g., from garage startup to publicly securitized corporation.) In addition, detailed financial data including retirement contributions and withdrawals, and demographic data on age, marital status, alimony, blindness and other disability status, veteran

and military combat status, education and major medical expenditures make individual tax returns equally valuable as an information resource. In short, it is apparent that tax return information could be useful for answering a whole host of important socio-economic questions. The richness of the data stems in part from their universality. This enables important measures, such as employment totals, to be compiled at the employer level through the employment tax returns filed by businesses (for example the Form 941 series long used by the U.S. Census Bureau). The richness stems also from the ability to link returns across economic entities. Thus, employment totals can also be compiled at the worker level – both employees and contract workers -- and associated with the related employers through SSN/EIN crosswalks (for example, using the Social Security Number (SSN) and Employer Identification Number (EIN) captured from the employee's Form W-2 used to report annual wage and salary payments and taxes withheld). Similarly, because of the dependent SSN's available on individual tax returns, it is possible to identify demographic families. In conjunction with the corporate families described above, it is possible to link the demographic population of households – workers and a large segment of their family dependents -- with the employer population, including corporate families of parent and subsidiary firms.

Finally, the richness stems from data quality. The tax compliance program makes it possible to ensure a certain degree of data quality through legal requirements for not only the timely filing of returns, but also their accurate completion[v]. Thus, the tax system is able to capture these business and demographic populations regularly – annually, quarterly, and even monthly for some returns. Part year returns capture partial accounting periods and final returns indicate the death of a business as that entity, whether due to bankruptcy, acquisition, or family decision to cease operations. The record level microdata are posted to one of several master files, including the Individual Master File (containing data for Form 1040 and related filings), the Business Master File (containing data for many business returns and also non-profit organizations), and the Employee Plans Master File (containing data for employee plan filings, especially retirement plans). In sum, the variety of tax return filers, the financial and entity detail provided on their returns, the regularity of filings, the universe of coverage, and their data linkability make the tax system a potent resource for research and analysis.
In addition to population data, sample files are produced by SOI for individual and business data[vi] and contain much more detailed record content than the population master file data. On the other hand, the master file accounts contain post-return transactions such as amended returns and carrybacks of unused net operating losses and credits, making possible the calculation of ultimate effective tax rates. Together the SOI samples and Master File populations present rich data resources for economic analysis of businesses and workers.

Challenges on Use

Legal Constraints, Confidentiality Protection

All access to tax data begins with statutory authorization; even for access by IRS employees. For example, staff at SOI are authorized by statute[vii] to access tax data to produce statistics of income both authorized and mandated by another statute[viii]. This legalistic focus has long been recognized as the basis for tax data's confidentiality protection, but it has proven a challenge for

other uses that tax data serve, including their important role in economic analysis. IRS is obsessed with protecting taxpayer confidentiality primarily for one reason. It is viewed as the cornerstone of voluntary compliance, which is considered the backbone of the tax system itself. Anything which threatens, or is perceived to threaten, the protection of taxpayer confidentiality – including statistical use -- is unacceptable and will not be tolerated.

Protecting the confidentiality of tax data is challenging[ix] and expensive, for two reasons: there is no statute of limitations and the Tax Code treats all FTI the same with respect to confidentiality protection. That is, to IRS a business' name or address is as deserving of confidentiality protection as are income items for a large corporation's or individual's tax return, and all must be protected in perpetuity.

The protection approach taken by IRS is two-pronged. Part of the protection is physical in nature: statistical agency recipients of FTI must undergo regular on-site safeguards[x] reviews that include examinations of not only physical and computer security systems, but also scrutiny of past uses. These reviews confirm the recipients' understanding and implementation of the many requirements covering physical and computer security, data need and use, and appropriate documentation.

Part of the protection is legal. The access of FTI must be only for purposes authorized by statute, possibly supplemented with regulations and, infrequently, policy agreements. Proposed new uses of tax data may be scrutinized not only as part of the official inter-agency request correspondence process, but under a more formal review process established by IRS and the recipient agency. This "need and use" review is another tool used by the agencies to ensure that due diligence is conducted, including documentation, for such accesses of tax data.
Given these constraints, the resource consequences of safeguarding taxpayer confidentiality over time are non-trivial. These constraints are exacerbated by the potential for complementary disclosure, or the re-identification of taxpayer data using indirect means; e.g., using data in other publicly released data to identify FTI related to a particular taxpayer.

Another part of the protection strategy is to minimize access: statutory policy on tax data authorizes provision of *the minimal* amount of tax data for an authorized purpose. This leads to an historical tension with statistical agencies, such as Census, since their mandate on administrative records is to *maximize* such usage. The tension may lead to friction unless a mutual agreement on process, protocols, and access parameters addresses the needs of each agency in the provision and usage of tax data. Sometimes this agreement may result only after a catalytic crisis, followed by some period of "turbulence" and bargaining towards an equilibrium position[xi].

In sum, the tax system seeks 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.

Since padlocking the treasure of tax data is viewed as neither a desirable nor a viable outcome, some disclosure risk is accepted as part of the necessary balance of protection and access. But there is clearly an 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 the perception of a *plausible access quantity* limit. Even with unlimited funding and perfect safeguards, 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.

Pro's and Con's of Analysis by External Researchers

An increasingly sensitive subject in recent years has involved the role of tax data in economic analysis or research, not merely for producing the disclosure-processed, redacted products historically produced by statistical agencies.

One reason for concern about expanding access to microdata for external research is the exclusive mandate on tax policy analysis – involving access to FTI -- for proposed legislation historically afforded the Joint Committee on Taxation (JCT) in Congress and the Office of Tax Analysis (OTA) in Treasury. An understandable issue for both groups is that they may be "blindsided" on some controversial item by the conclusions of outside researchers with access to tax microdata, particularly if they have access to too little tax data -- or are not sufficiently conversant in their analysis -- to be completely informed about the tax policy issue under consideration. Another issue with expanded access by outside researchers is that unless the justification is compelling, there might be the dangerous perception of such access enabling a fishing expedition. Indeed, the term "data mining", dear to so many researchers, has such a connotation among certain circles within the tax world at both Treasury and IRS. Because of this connotation, it raises alarms because it invokes the prospect of data browsing, so recently such a problem for IRS that special legislation was enacted to make it an explicit crime with serious consequences.

However, given that statistical tax data are produced with publicly provided funds and given that publicly available data are often too imprecise – due to their heavy redaction after disclosure processing – to answer important economic questions[xii], another option would seem necessary. In effect, this might be seen as an additional systemic check for purposes of further "democratizing" the decision process by improving its transparency. Several factors might argue for at least some additional access by external researchers. First, the analytical questions being answered by JCT and OTA are largely driven by the political discussion in Congress and the Administration. Other questions – not being actively pursued -- might also be relevant to the public debate, and this role might be played by outside researchers, on a carefully controlled basis, and for a *bona fide* national economic imperative such as the study of innovation and entrepreneurship mandated by the America COMPETES Act. Second, analytical resources are so heavily burdened at JCT and OTA that there is sometimes less than optimal opportunity to ensure the comprehensive accuracy of analytical results. Third, both JCT and OTA may call upon outside researchers to assist them with their analysis, but such requests are strictly at their discretion. Perhaps a different type of "third party" scrutiny would be provided by more outside researcher analysis, especially if it might be viewed as helpful, not divisive or destructive, to the ultimate decision-making process. Fourth, the days of public use files, including the Tax Model produced by SOI, may be numbered due to the increasing difficulty of protecting confidentiality while also delivering analytical utility. That is, even the publicly available data options may be declining for analysts. Fifth, because of the limitations of publicly available data combined with the ever increasing appetite for data and information, it is certain that surveys will be conducted

by outside institutions using lesser quality datasets for sampling frames and supplementary data – especially if they are all that is available. This might have the unwanted effect of a perverse version of Gresham's Law, in which bad data analysis drowns out or neutralizes analysis done with good data. Finally, the increasing capacity of technology itself, including more powerful computers and techniques for both research analysis and confidentiality protection, argues for at least some expansion in tax data access when it can actually reduce risk historically viewed as inevitable with expanded data access. In sum, the case for some expanded but controlled access would seem to be compelling, especially given the alternatives – including the economic cost to decision makers and the nation – of not expanding it.

Portfolio Approach[xiii]

Over the past few years SOI has pursued an approach to data dissemination that explores a menu of modalities contingent upon resource and confidentiality constraints, as well as an external researcher's actual data access needs. Traditional release products include aggregate disclosure-processed tables and a public use microdata file (PUF) based upon individual tax returns and known as the tax model. However, the suppression of cells in tabulations and the use of top-coding and blurring in the PUF have made them less than useful for many researchers. Moreover, for the PUF, the proliferation of data sets in the public sector, as well as previously confidential variables, such as property taxes, make the protection of taxpayer confidentiality, especially via complementary disclosures, increasingly difficult if not impossible – especially for data with no statute of limitations on the protection of confidentiality. In short, the days of traditional PUF's may be numbered due to confidentiality concerns alone.

Synthetic data may hold some promise for continuing the production of public use files – and without the baggage of either the confidentiality concerns or the analytical constraints. In fact, it may be very helpful in creating other genres of PUF; e.g., a business PUF. Nevertheless, its adoption presents a number of problems, most notably, opportunity cost. In addition, SOI has no direct experience with this methodology, although SOI's parent organization has collaborated in a small way with the U.S. Census Bureau on its efforts (Abowd et al.). Because of the relative lack of experience and expertise in SOI, the learning curve on any meaningful development of synthetic data could be expensive. Nevertheless, it is under consideration for the future, especially because of the positive prospects for saving and even improving the current PUB, as well as developing others; e.g., a business data PUF.

Buffered remote access is another option in the SOI data dissemination portfolio. This modality consists of a program being submitted directly or indirectly by an external researcher without direct access to the actual gold standard data, including viewing of the results, which must be disclosure-processed by the data custodian before being provided to the researcher. The progress of this sort of "black box" access has been hampered in the US particularly by confidentiality concerns, including the chances of one or more external researchers "gaming" the system to produce outputs that might be triangulated to divulge identifiable data through complementary disclosures.

The gold standard data access modality is another option that SOI has increasingly adopted in two ways. First, it is partnering with other federal offices on their needs for access to FTI that entail a compelling national imperative consistent with SOI's statutory mandate on statistics. For example, the National Science Foundation (NSF) has agreed to be SOI's contractor on its

innovation survey of small businesses. FTI will serve as the sampling frame for the survey and also to validate and supplement survey responses. A limited number of NSF staff (and NSF-contracted researchers) will be granted access to FTI initially for purposes such as sample design and later for analysis, under carefully controlled conditions. Second is the SOI External Research Program itself.

The External Research Program (ERP)

Like many other federal statistical offices, SOI used the Intergovernmental Personnel Act (IPA) to enable a limited number of academic researchers to access FTI for certain research purposes, often resulting in publication of one or more articles in professional publications and related conference presentations. Although the researchers were not monetarily compensated, some expense reimbursement was provided by SOI; e.g., to travel to Washington, DC, for discussion or data access purposes.

This venue was abruptly terminated in the 1990's when SOI was advised by IRS Counsel that such arrangements were viewed as unacceptable for accessing tax data. Reasons included the lack of a tax law statute authorizing such access and the lack of a competitive process to allocate such access. In addition, there was the danger that such a process might be perceived as somehow conferring an ethically problematic *quid pro quo* without the legitimacy provided by a monetary contract, under which customer and contractor roles were seen as more clearly delineated.

A confluence of events served to inspire the idea of the current external research program. First, in preparing its ten year Strategic Plan in 2006, SOI realized that, as one of the fourteen major federal statistical offices, it needed to focus more on data dissemination as a critical part of its future responsibilities. Second, although SOI's primary microdata consumers are OTA and JCT, the SOI Advisory Panel of outside experts had helped reinforce the notion that SOI's ultimate customer is the American public, including its external researcher contingent. Given the millions of dollars spent on the collection of tax data, continuing to produce only heavily redacted aggregate tables for public release was seen as shortchanging not only the analytical research community but also the data respondents; namely, the taxpayers, who had endured the burden of completing and submitting detailed return documentation which could serve more than a narrow tax administration purpose. In sum, the return in aggregate expurgated data tables was not commensurate with the investment either for taxpayers or for policy makers, both public and private. SOI also realized that in order to avoid stagnating as only a data producer, it needed an infusion of ideas from the outside research community on compelling purposes for which to make available such rich datasets. In addition, SOI and its parent organization had begun to develop its portfolio approach to data dissemination, in which external researcher access to confidential microdata access was viewed as a critical component.

Accordingly, since that time, the only mechanism for outside access has been contractual. In 2008, SOI began a program using this approach for outside access, albeit for proposals that SOI must approve through a competitive review process, but whose specifics are determined by the outside researcher. This outside researcher access program, while relatively unconstrained by

the fixed deliverable requirement of most contracts, nevertheless, has four criteria: utility of the proposal for tax administration (that is, its statistical/research component) and the larger statistical community; familiarity with the SOI confidentiality culture and SOI products (mostly, due to limited SOI resources available for educating a researcher) ; contributions to SOI's human capital through partnering with SOI staff on the work or by teaching a course to SOI staff; and the researcher's credentials in terms of academic qualifications and experience.
In its request for proposals, SOI solicits high quality research projects necessitating access to FTI. The objectives of the proposals include, but are not limited to: (1) improving and advancing SOI's analytical and institutional knowledge base, (2) improving data quality or processing methodology, including sample design and statistical disclosure avoidance, (3) reconciling measures of income-related variables using tax and non-tax sources, (4) creating new products, including a business data public use file or linked data files and (5) analyzing economic and tax behavior characteristics such as the propensity to claim various deductions, adjustments and credits at both the individual and business level over time.

Prospective researchers themselves must be US citizens or have permanent resident status and three years residency in the US. In addition to having appropriate professional credentials, especially in economics, they must pass background checks that include fingerprinting. In addition, other logistical tasks after contract award include obtaining the data access permissions for universe level data on the Compliance Data Warehouse and SOI sample files, as well as secure messaging. Although some types of internal IRS transactions; e.g., audit detail, may be excluded from researcher access, no tax form data per se are so far excluded from access. That is, all genres are accessible – whether the data are for individuals, businesses, non-profits, or even linkage involving these entities; e.g., workers and firms. Presently, all access to FTI must occur in the US at an IRS facility, which is usually not a problem unless a researcher is not near a major urban area. Even then, the researcher can be accommodated, as IRS has numerous district offices throughout the country that might be utilized for researcher access, space permitting. Data access may be up to two years, although a six month extension is possible, upon approval.

ERP Experience to Date

Although the awards made in the initial year essentially continued the valuable work of two previous IPA's, the second year yielded three awards for researchers without previous external access to FTI. In particular, Dr. Emmanuel Saez, winner of the 2009 John Bates Clark medal, and his colleagues, Dr. Raj Chetty and Dr. John Friedman, were contracted for his project on "Behavioral Responses to Taxation: A Proposal to Work with IRS Microeconomic Data". They will access selected universe tax data for individuals over the past ten years. In addition, Dr. Tami Gurley-Calvez and colleagues were contracted for her project on "Enhancing Corporate Research Data: Linking IRS and Kauffman Firm Survey Data"; selected universe business data will be accessed. The third contract was awarded to Dr. Robert Carroll, the Treasury Department's previous Deputy Assistant Secretary for Tax Analysis, for his proposal to study Health Savings Accounts, particularly timely in the U.S. as healthcare reform becomes such a priority in the national economic dialogue.

Challenges for the Future

As the program has climbed the learning curve, logistical lessons have been as instructive as those from the proposal review process, given the coordination necessary between SOI and other functions within IRS to enable physical and computer access within acceptable security parameters. Because the post award logistical actions necessarily must occur after the two year clock begins with the contract award, they can reduce the actual data access time by many weeks, so that the 6 month extension may become a *de facto* part of a project's duration. Researcher frustrations at not receiving access upon contract award are most understandable, but so are the limitations of resources available for an external research program at a small statistical office such as SOI. Such frustrations may simply be part of the mutual education process for researchers and SOI, especially in a statistical system that sees virtue in its decentralization for at least three reasons. First, centralized statistical systems may be more susceptible to over-politicization and manipulation of data and analysis. Second, an administrative agency's statistical office is understandably seen as most likely to have the greatest expertise concerning the utility and disutility of its data. Third, by concentrating analytical and quality enhancement resources in the administrative agency's statistical office, external releases -- whether to other agencies or external researchers -- can be optimized for most effectively discharging the functions of data stewardship and confidentiality protection. That is, only the minimal amount of data necessary for an authorized statistical purpose are released, re-emphasized by conducting external research through the administrative agency's statistical office, rather than at a central statistical agency.

A lack of support resources may, however, threaten the program's sustainability, especially in this critical formative stage, as the program is appreciated by some less for its potential and more for the burden it inflicts upon resources called upon to do ever more with less.

Future Possibilities – Research Consortium, Data Enclave, Remote Access

In 2009, the program's second year, even though the number of research projects increased by only one, the number of researchers with access to FTI increased from two to ten, including two research assistants at Harvard. For such a small program in its infancy, this fivefold increase in researchers resulted in a non-trivial amount of logistical work regarding fingerprinting, background checks, physical and other security passes, software and hardware – including the protocol permissions needed to access the various datasets. This does not include the systems and other support necessary for the lifespans of the proposals. In addition, because of the relatively small size of the contract awards for this program, the IRS Procurement and related functions necessary for contracting and post-award logistical support, may tend to accord this work a lower priority than, say, larger compliance-oriented contracts for the tax agency.
Three enhancements may help to mitigate these problems in the longer term. First, a more significant critical mass might be achieved through pooling the mostly *ad hoc* external researcher needs of other research offices in IRS, and perhaps Treasury's OTA as well. Although SOI statistical files would continue to be used only for statistical projects, and bright line distinctions would also be maintained between SOI external researchers and non-statistical researchers, the compliance component of other external research might have the advantage of helping to elevate the priority of treatment for logistical services such as background checks, etc. In addition,

economies of scale might accrue through this approach. Second, this consortium of researchers might find further efficiencies through outsourcing of the actual data access and maintenance, disclosure review, and other associated "headaches" to a data enclave arrangement, such as that maintained by the National Opinion Research Center (NORC), University of Chicago[xiv]. For example, under the NORC data enclave, researchers would access only the data specified by the provider (IRS/SOI), and under terms articulated by the provider. That is, access might be restricted to on-site between certain hours, no data prints might be allowed, etc. Disclosure review would also be subject to the provider's criteria, and, of course, the data enclave would need to meet all IRS safeguards requirements and even be subject to a safeguards review. A third enhancement, real – not buffered -- remote access to the microdata, might ease some of the access burden for researchers and even help to popularize the program. Although the sensitivity of tax data is well known, this enhancement is not as risky as it may first seem, given that many IRS employees, including at SOI, can now access confidential microdata from their homes via a secure access system known as the Enterprise Remote Access Project (ERAP).

---

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

[ii] The numbers presented are for Tax Year 2005.

[iii] Includes over 40 million sole proprietorships, represented by Schedule C and Schedule F filings.

[iv] Employee benefit plans are mostly retirement plans (defined contribution or defined benefit) but also might be welfare plans, a category that includes dental and legal service plans offered to employees.

[v] Because of both penalties for noncompliance and tax benefits – credits, adjustments, deductions, refunds – accruing to businesses, it is in their interest (especially those seeking access to public capital markets) to be captured by the tax system.

[vi] For a more complete description of SOI and its programs, access http://soi.soi.irs.gov/. A listing of SOI projects and contacts can be found at http://www.irs.gov/taxstats/article/0,,id=169439,00.html.

[vii] Section 6103(h)(1) of the Internal Revenue Code (IRC) authorizes tax data access for tax administration purposes, which include statistical and research components

[viii] Section 6108(a) mandates the Secretary of Treasury to prepare and publish annual statistics with respect to the operation of the internal revenue laws, including various variables and taxpayer classifications.

[ix] See Greenia, Nick, *The Release of IRS Data: Challenges and New Approaches*, IRS Research Conference, 2002.

[x] Safeguards standards are described in the IRS *Publication 1075, Tax Information Security Guidelines for Federal, State and Local Agencies and Entities.*

[xi] See Greenia, Nicholas H., "Developing Adoptable Disclosure Protection Techniques: Lessons Learned from a U.S. Experience", *Privacy in Statistical Databases*, CASC Project Final Conference, PSD 2004, Barcelona, Catalonia, Spain, June 9-11, 2004 Proceedings.

[xii] See Lane, Julia, "Optimizing the Use of Microdata: An Overview of the Issues, Journal of Official Statistics, pp. 299-317.

[xiii] This treatment borrows from Abowd, John and Lane, Julia "The Economics of Data Confidentiality", October 16, 2003, Cornell University and the Urban Institute.

[xiv] For more information on the NORC data enclave, access http://www.norc.org/DataEnclave/. NORC may have a further advantage in its longstanding experience with confidential tax data as the subcontractor to the Federal Reserve Board for helping conduct the Survey of Consumer Finances (SCF). FRB's access to tax data for the SCF is as a contractor to Treasury and SOI.