A proposal to improve our understanding of innovation using the

FEDERAL TAX DATA: A CASE OF UNDERUTILIZATION FOR ECONOMIC ANALYSIS

Proposed by
Nicholas Greenia
Statistics of Income Division,
U.S. Internal Revenue Service
Nicholas.H.Greenia@irs.gov

Potential advancement in understanding innovation
Full utilization of tax data could shed new light on employee benefits, business ownership structure, businesses evolution, effective tax rates and professional employer organizations, just as examples.

SUMMARY

It is 95 years (1913) since the U.S. Constitution's 16th Amendment authorized Congress to enact an income tax and over 50 years (1954) since the modern tax code was instituted. The tax system now is explicitly used to affect the economic behavior of workers, businesses, and nonprofit organizations. As a result, an entire "tax industry" has developed, including well over 100,000 employees and contractors of the Internal Revenue Service (IRS); Treasury Department analytical offices, such as the Office of Tax Policy; analysts in the legislative branch, such as the Joint Committee on Taxation and the Congressional Budget Office, as well as many professionals in the legal and accounting services who make their livings by keeping pace with tax law. This explicit economic engineering provides a substantial opportunity to use tax data for economic analysis. However, restricted access and incomplete data set architecture have limited this use, thus constraining not only the information available to decision-makers but also the transparency of the engineering process.

This paper examines the utility of tax data for economic analysis. It discusses two improvements that could be made to increase the utilization. One of these is structuring the processing design to connect major related features of the tax picture. The second is promoting access to tax microdata. The paper concludes with recommended next steps for the researcher and statistical communities.

THE UTILITY OF TAX DATA

The U.S. Internal Revenue Service collects population data annually[1] for a variety of entities: more than 130 million individuals, including sole proprietors; more than 50 million businesses;[2] more than 800,000 tax-exempt organizations; and related entities, such as approximately one million employee benefit arrangements, mostly retirement plans. 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.[3] In addition, the tax system is able to capture these business and employee populations regularly, ranging from annually and quarterly to even monthly for some returns. Partial 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

1 The numbers presented here are for Tax Year 2005.

2 Includes more than 40 million sole proprietorships, represented by Schedule C and Schedule F filings.

3 Because of both penalties for noncompliance and tax benefits – credits, adjustments, deductions -- accruing to businesses, it is in their interest to be captured by the tax system.

KAUFFMAN Symposium on Entrepreneurship and Innovation Data 2008

to one of several master files, including the Individual Master File (IMF, containing data for Form 1040 and related filings), the Business Master File (BMF, containing data for many business returns and also nonprofit organizations), and the Employee Plans Master File (EPMF, containing data for employee plan filings, especially retirement plans, such as the Form 5500 series). Business births are captured by the system. The beginning of a business employer entity, but not necessarily every new business, starts with the filing of a Form SS-4 for assignment of an Employer Identification Number (EIN) by the IRS in order to establish its account on the BMF. In a sense, the BMF can be viewed as the business register of the tax system, and BMF population extracts provide the core of the U.S. Census Bureau's own business register with annual infusions of selected data variables for the tax system's business employer population. In fact, the U.S. Census Bureau considers tax data to be the lifeblood of their business register.

The type of business also is captured: the SS-4 requires the business to tell the IRS whether it is beginning as a sole proprietorship, partnership, corporation or personal service corporation; the state or foreign country in which it is incorporated, and whether it is applying because it is a new entity, has hired employees, has purchased an ongoing business, or has changed the type of organization (specifying the type). For sole proprietorships that require EINs, generally employers, the form also asks for the owner's name and Social Security Number (SSN), providing an instant crosswalk from the BMF to the IMF, where the sole proprietorship's Schedule C would post as part of the related Form 1040 filed by the sole proprietor. In addition, this information is requested for the principal officer or general partner. The SS-4 also begins classifying a firm in terms of industrial activity by requesting a verbal description of its principal activity and principal line of business, information which is used later by the Social Security Administration (SSA) to assign the firm's first [at least for this EIN] code from the North American Industry Classification System (NAICS).

Information also is available on the ongoing activities of businesses. The BMF contains the Form 1120 series, representing corporations; the Form 1120S and Form 1065, representing pass-through entities; the Form 940 and 941 series of employment tax (FUTA and FICA,[4] respectively) returns; and the publicly available Form 990 series,[5] representing nonprofit and charitable organizations. Business tax return data, including complete balance sheets and financial statements – economic stocks and flows – contain information on everything from net business profits to travel and advertising expenses, as well as NAICS-based industry codes, mailing address and physical location. Entities include corporations, partnerships, and sole proprietorships, employers and non-employers. Data are collected not only for these typical businesses but also for entities not typically thought of as businesses, such as governmental entities and nonprofit organizations, which are increasingly important economically due not only to the activities they perform but also their role as employers, whether workers are compensated employees or volunteers. In addition, for corporations filing on a consolidated basis, it is possible to link the subsidiary corporations with the parent, identifying corporate families and even the ultimate parent of one or more corporate families.

The tax system also captures the components of effective tax rates by recording not only the initial return's tax liability but also post-filing transactions, such as amended returns, NOL and credit carrybacks, and IRS audit actions. Because the BMF generally is designed to retain a tax module (the posted return data and related transactions for a given tax year) within an EIN account from three years after the latest tax transaction, carrybacks can keep some accounts active on the BMF for much more than three years. For example, losses due to product liability can be carried back ten years. (This treatment also has applied to bad loans.) In such cases, the IRS retrieves previously removed modules by tax year to provide a time continuum from the destination year affected by the carryback through the loss-tax year that generated such an action, effectively restoring seven years of previously jettisoned data. In combination with the ricochet effect[6], adjustment transactions can vastly extend the shelf life of data retained on the BMF for decades in some cases. Thus, for many of the most interesting and complex industries and size classes, often the predominant companies in corporate America, this continuous churning creates a dynamic and long-term record on the BMF that may provide a story of electron-level economic activity for the core of American business.

4 Federal Unemployment Tax Act, Federal Insurance Contributions Act.

5 Not all of the Form 990 series data are publicly available; e.g., Form 990-T.

6 Carrybacks must be taken in order of priority so that, say, an NOL CBK could free up a previously taken credit for further three_year carryback, etc.

In addition, the tax system collects information on individuals. The IMF contains the Form 1040 series of individual tax returns, data on employees, as well as more than 40 million Schedule C and Schedule F records, nonagricultural and agricultural sole proprietorships, respectively. In addition, the Employee Plans Master File (EPMF) contains records for employee benefit plans, especially retirement plans, such as the publicly available Form 5500 series.[7]

This wealth of information means that tax return information could be useful for answering a host of important economic questions, in addition to those directly related to tax policy and particularly those that cannot be answered with aggregate or publicly available data.[8] An obvious example is the study of innovation and entrepreneurship, where the apparent over-estimation (40 million) of sole proprietorships could prove a boon for first-identifying, start-up firms. Tax data at the firm level would be ideal as both the sampling frame for such a targeted study and as a source of supplemental financial data, incl uding for validation and imputation purposes.

The universality of the information means that tax return information can be used to construct economic measures, such as employment and payroll compiled at the employer level through the employment tax returns filed by businesses. (Examples include the Form 941 series long used by the U.S. Census Bureau, as well as the Form 940 series not used.) Employment totals also can be compiled at the employee level and associated with the related employers through crosswalks using the Social Security Number (SSN) and EIN captured from Form W-2 used to report annual wages, salaries, deferred compensation, and taxes withheld.[9] The Form 851 (affiliations schedule), filed with the return for consolidated corporations, enables assembling a corporate "family" through parent and subsidiary EIN links. In conjunction with the W-2's EIN/SSN crosswalk, this means that corporate families can be linked to demographic families, using the Form 1040's dependent SSN's. Just as useful are other measures, such as firm entries and exits, which are critical data for any economic analysis of business formation, as well as job creation and destruction.

Tax data also can be enhanced by linking not only the different economic entities but also the different forms of legal ownership for the same entity as it matures. Because the IRS must be able to track the same tax entity through different forms of legal ownership, including mergers or acquisitions, transaction records are generated for changes from one EIN to another. Thus, it might be possible to follow the life cycle of a business through its evolution from start-up to publicly traded corporation, and, as in the case of someone such as Bill Gates, even to philanthropic nonprofit.

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. Nevertheless, even with a compliance program, the IRS database, like other databases, is imperfect in response quality. This inconvenient fact is supported by the most recent estimate of the tax gap – the difference between what is owed and what is paid – at $345 billion.[10]

PRIORITY IMPROVEMENT 1

A major improvement would be to restructure the design of tax processing in such a way that major related features of economic activity could be connected. Although the various master files at the IRS capture the population of records, they do not capture the entire content of each record. The arrival of electronic filing, combined with dramatic reductions in storage costs, mean there are new opportunities for processing data for analytical purposes. In what follows, we highlight these possibilities for such important elements as employee benefits, business ownership structure, business evolution, effective tax rates and professional employer organizations.

7 Not all of the return data are publicly available; e.g., Schedule SSA.

8 See Lane, Julia, "Optimizing the Use of Microdata: An Overview of the Issues", Journal of Official Statistics, pp. 299—317.

9 The W-2 and other information documents (such as Form 1099 that captures payments of interest, dividends, etc., for both payer and payee) repose on yet another IRS master file, the Information Returns Master File (IRMF).

10 For Tax Year 2001 -- see "IRS Updates Tax Gap Estimates," IR-2006-028, available at www.irs.gov/newsroom.

IMPLICATION OF IMPROVEMENT 1

Form 5500 Series, Employee Benefit Plans

With the advent of the Employee Retirement Income Security Act (ERISA) in 1974, a new form, the 5500 series, was required to be filed for the employee benefit plan[11] by plan sponsors (usually the employer) and/or the plan administrator (sometimes also the employer) to satisfy requirements of the four agencies charged with administering ERISA: the Department of Labor (DOL), the Pension Benefit Guaranty Corporation (PBGC), the Social Security Administration, and the IRS. The DOL and IRS have the lion's share of responsibility for processing the 5500 series of returns (now largely contracted out to a private sector vendor), which, except for a few sections, is publicly available and captured on a separate IRS master file system, the Employee Plans Master File (EPMF). Like records posting to the BMF, each 5500 series record also posts on its EIN, but because an employee plan need not have the same EIN as the plan sponsor – indeed, a plan might have multiple sponsors – linking the financial information on the 5500 series to the sponsoring company taking an employee plan-related tax deduction; e.g., under section 401(a) of the tax code, is not a simple matter of matching a common EIN. The financial information provided on the 5500 series is considerable, essentially complete economic flows and stocks – income and asset statements – for the plan. This information and other information on the plan, such as numbers of participants and employees by vesting status, could be even more useful for economic analysis if matched to the plan sponsor's tax returns, and then to related employee information, such as the Form 941 series of employment returns and the W-2 withholding form.

Presently, the EIN linkage of plan to sponsor is accomplished only on a case-by-case basis by the IRS examination function, but a systemic match, like what is done with the W-2 and Form 1040 series, is not possible. One solution to this problem might be what is done with the posting to the BMF of subsidiary corporations when only one Form 1120 is filed by the parent for a consolidated corporation, along with the Form 851 listing the subsidiaries, including their EINs. Namely, the validated parent EIN is included on the transaction satisfying the subsidiary's filing requirement, enabling construction of the corporate family for a particular consolidated filing. A similar idea might enable linkage of the EPMF to the BMF (including nonprofits represented by the Form 990 series). That is, the posted 5500 series record data might include validated plan sponsor EINs (validated against the BMF) and the plan sponsor's tax return design. For example, a Form 1120 might be amended to include the EINs of the employee plans it sponsored (validated against the EPMF).

Business Ownership Hierarchy

Although it is possible to construct a corporate family using information posted from a consolidated corporate return filing, it is not possible to determine which corporation owns which in such an arrangement. This information can be important for determining the decision-making unit in a variety of analyses, and it might be obtainable with a minor change to both the Affiliations Schedule (Form 851) and its electronic processing. Presently, the form only enables such an ownership hierarchy for a maximum of ten corporations before a new Form 851 is required, again numbered 1-10. Instead, with electronic filing, the hierarchy identified could be a continuum, not discrete sets of 10, of subsidiaries with the transactions satisfying the subsidiary's filing requirement, including the number of immediate owner. In conjunction with the ultimate owner information available from the corporate Schedule K, this might provide another important tool for economic analysis.

Tracking Business Evolution: Mergers & Acquisitions

The IRS has the rudiments for tracking changes in legal form of ownership (e.g., from partnership to corporation) but it is less than systematic and mostly untested due to a lack of regular and rigorous analytical use. In addition, the fact that the IRS posts sole proprietorships to the IMF with an individual's Form 1040, but other businesses and nonprofits to the BMF, causes a break in the continuum for tracking a sole proprietorship to other forms of legal ownership when the sole proprietorship does not have an EIN

11 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.

(most sole proprietorships). As in most, if not all, of the sectors described so far for this section on processing and design inadequacy, there is a tax administration purpose that complements the analytical need. That is, the IRS needs to be able to keep up with changes in entity[12] or identity information for a given taxpayer in order to mail tax forms and assess the taxpayer's compliance with the tax laws based on the documentation provided by the taxpayer. The latter includes avoiding unnecessary or false positive compliance actions against the taxpayer. Mistakes are always expensive, and tax administration, such as assessing delinquent behavior when outreach or education might have resulted in tax-compliant behavior, is no exception.

Currently, the most promising possibility for tracking a change in business type is available on the BMF when an account changes its EIN (often entailing the filing of a new SS-4) and/or entity type and the "new" or acquired business's EIN account records the old EIN. Such a transaction might be due to business expansion, perhaps resulting in the incorporation of a former sole proprietorship or the acquisition of one business by another and the subsuming of its account information into the new owner's account. Enabling such time tracking could provide important information for studying innovation and entrepreneurship, especially when it is important to establish the birth of a business or business activity and track its evolution.

One option for resolution of the sole proprietorship problem might involve a slight amendment to the SS-4 so that it requires the SSN for not only a sole proprietor requesting an EIN but also for a requestor that was formerly a sole proprietor, providing the IMF/BMF crosswalk needed. More differentiation on entity changes also would help distinguish acquisitions from business entity changes that result from business expansion.

Effective Tax Rates

Of course, tax data also are the only real way of comprehensively understanding business responsiveness to taxes because effective tax rates only can be calculated using post-return filing information or tax adjustments, available from the filing of return amendments; carrybacks[13] of an unused credit portion, net capital loss (NCL), or net operating loss (NOL); and IRS-initiated examination efforts of a taxpayer-initiated transaction, including the original tax return but also the delinquent omission of a return filing. Because the BMF is designed to retain a tax module (tax return and associated transactions for a given tax year) within an EIN account for three years after the latest tax transaction, post-filing transactions, especially carrybacks, can keep some modules active on the BMF for much longer than three years. For example, losses due to product liability can be carried back ten years. (This treatment also has applied to bad loans.) For such a case, two phenomena are worth mentioning. First, not only is the destination tax module retrieved – if it had expired and been purged from the BMF – but its retention clock is reset for three more years. Second, for such cases, the IRS retrieves previously removed tax modules between the loss-year module or tax year originating the carryback and the destination tax year, providing a time continuum for at least another three years (the BMF retention rule). A loss due to product liability might be carried back the full ten years, effectively restoring all seven years of data scheduled for purging under the routine BMF retention schedule. In combination with the ricochet effect[14], these adjustment transactions, in some cases, can vastly extend the shelf life of data retained on the BMF for decades. The ricochet effect is caused by the requirement that carrybacks must be taken in a hierarchical order of priority. In general, the NOL and NCL come first, followed by credits, in the order they appear on the tax return. For example, an NOL carryback could free up a previously taken credit for further three-year carryback, and so on, resulting in the ultimately released credit being carried forward for use on a future tax return, possibly resulting in some of that year's credits being carried back, triggering a similar fall of dominoes. This constant churning – the release of previous carrybacks for further carryback – means some firms have a continuum of tax modules for much longer than three years. If any of these carryback transactions draw examination interest, the three-year retention period can be reactivated several times, further extending a module's duration on the BMF.

12 The entity section of a master file account contains information pertinent to locating and classifying a taxpayer, such as name, mailing address, physical location, industry code, and taxpayer identification number (SSN or EIN).

13 A carryback filed within one year of the loss-year return's end of accounting period requires usage of Form 1139 (Corporation Application for Tentative Refund). Form 1120-X (amended return) is required for carrybacks taken up to three years after the loss-year return's end of accounting period. Generally, a credit or NOL must be carried back (the period has generally varied at two or three years) , before it can be carried forward (this period also has generally varied from five to seven years) for use on future tax returns. An exception is the NOL, which can be carried directly forward if an irrevocable waiver is established, bypassing the carryback requirement.

14 Carrybacks must be taken in order of priority so that, for example, an NOL CBK could free up a previously taken credit for further three-year carryback, etc.

Over time, sophisticated tax avoidance strategies maximizing carryback tax laws can be employed by savvy firms, but the transactions need to be large to reap the investments in legal and accounting capital that may be necessary to optimize this use. Thus, for many of the most interesting and complex industries and size classes — often the predominant companies in corporate America — this continuous churning creates a dynamic and long-term record on the BMF that may provide a story of electron-level economic activity for the core of American business.

Unfortunately, although the BMF captures the net tax liability effect of a tax adjustment, say, a carryback, it does not capture details, such as whether it was a credit and which type of credit – essentially, the detail on Form 1139. The reason is that almost all post-filing BMF transactions are limited to 80 characters – a feature of legacy processing, including the net tax liability amount effect, necessarily limiting the amount of conveyable information. However, the information is captured in the IRS adjustments and examination systems, which are responsible for processing the various tax adjustment transactions generally differentiated by whether or not they are IRS examination-initiated. In conjunction with this knowledge and the proliferation of electronically filed returns, especially by corporations, it seems reasonable to think that the merging of data from the BMF and the adjustments area (especially, the Form 1139 detail) could be done, given the compelling analytical motivation that would seem to exist.

Professional Employer Organizations

Increasingly, businesses are contracting out personnel and business services that include the filing of certain tax returns, especially the employment tax returns represented by the Form 940 and Form 941 series. The problem for economic analysts – and for the IRS – is that the EIN used on the client's return is increasingly that of the contractor or professional employer organization (PEO). In an era of electronic filing, technology, again, seems like an apt solution for this problem with a minor modification to the employment return that would require validated capture of the client's EIN information, even if the return is posted to the PEO's BMF account.

PRIORITY IMPROVEMENT 2

In addition to data architecture, the second major reason for the underutilization of tax data is access, meaning the concern for protecting confidentiality. The rule of law has been inconsistent over time. For example, tax records were once made available to the public. Even though access may have been determined by societal needs at the time, in 2008 it is clear that all access to tax data begins with statute. Even access by employees at the IRS is governed by statute. For example, staff members at the Statistics of Income Division (SOI) are authorized by statute[15] to access tax data to produce statistics of income both authorized and mandated by another statute.[16] This legalistic focus long has been recognized as the basis for tax data's confidentiality protection, but it has proven challenging for other uses that tax data serve, including their important role in economic analysis.

IMPLICATION OF IMPROVEMENT 2

New access modalities, including Research Data Centers (RDCs) and remote access sites, offer the potential for increasing researcher access to tax data. However, the expansion of data access must be sanctioned by law. Three existing tax statutes are the most promising for relieving some of the historical impasse on tax data access. One is section 6108(a), which provides the mandate for the Treasury Secretary to produce tax statistics in the first place. Indeed, this section provides the mandate for the IRS's Statistics of Income Division (SOI). A second statute, 6108(b), provides the mechanism for an outsider to make a request for either special statistics or a special statistical study, possibly on a reimbursable basis, subject to confidentiality concerns and other work priorities; i.e., the availability of staff. The SOI is most often the function tasked with fulfilling such requests and a related revenue procedure ensures that it can keep such reimbursable funds for costs incurred, instead of passing them to a general fund as is done with

15 Section 6103(h)(1) authorizes tax data access for tax administration purposes, which include statistical and research components.

16 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.

requests made under the Freedom of Information Act. A third statute, 6103(n), provides a means of assisting the SOI in the event it lacks sufficient resources, including expertise, to conduct on its own work necessary for either 6108(a) or 6108(b) needs by enabling the SOI to engage outside contractors for such purposes. Thus, statute provides for both incoming funds, and the SOI's control of such funds, to engage contractors for compelling work under 6108(a) and 6108(b).

Presently, any entity, including federal statistical agencies, may access confidential tax microdata, or Federal Tax Information (FTI), only if a statute provides such authorization. This requirement is formidable, as it means that legislation containing such authorization has been proposed and passed by Congress and signed into law by the president. For some recipients, such as the U.S. Census Bureau, Treasury regulations also are necessary and may stipulate not only the specific purpose for which the FTI may be used but also the specific tax items the recipient may receive. Although the process is less arduous than that needed to change the statute, a regulations amendment still must undergo scrutiny by both the IRS and the Treasury Department, requiring approval by the Assistant Secretary of Tax Policy, often considered to be the nation's highest tax official after the Treasury Secretary. Both statute and regulation policy require only the minimal amount of tax data be provided to accomplish an authorized task. Even for authorized recipients, there also are official protocols for provision of tax data that include official request letters at the departmental level, although delegation orders provide for some routine correspondence to be done at lower executive levels. Annual reimbursable interagency agreements then may be developed, allowing the IRS to recover the costs of providing the data to recipients.

Protecting the confidentiality of tax data is challenging 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 the IRS a business's name or address is as deserving of confidentiality protection as are income items for a large corporation's or an individual's tax return. And, all must be protected in perpetuity.

The challenge is to identify acceptable risk, and the approach used to date is taking steps that prevent re-identification 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 is a technology-relative concept and, thus, may be a moving target. Nevertheless, it represents an attempt at due diligence in balancing the two goals for tax data: their protection and their effective use.

The protection approach taken by the IRS is two-pronged. Part of the protection is physical in nature: Statistical agency recipients of FTI must undergo regular on-site safeguards reviews that include examinations of physical and computer security systems[17] but also scrutiny of past and present 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. Related requirements include separate systems for processing or accessing FTI and background checks on individuals accessing FTI 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.

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 interagency request correspondence process but under a more formal review process established by the IRS and the recipient agency.[18] This "need and use" review is another tool used by the agencies to ensure 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 not trivial. These constraints are exacerbated by the potential for complementary disclosure or the re-identification of taxpayer data using indirect means, such as using data in other publicly released data to

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

18 For the U.S. Census Bureau, this process is described in Criteria for the Review and Approval of Census Projects that Use Federal Tax Information, effective September 15, 2000, at www.ces.census.gov.

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.

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 a historical tension with statistical agencies, such as the U.S. Census Bureau, whose mandate on administrative records is to maximize such use. 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 use 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.[19]

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 of demonstrable credibility.

The America COMPETES Act of 2007 may provide a useful opportunity for testing the limits of what appears to be a viable and existing mechanism for accessing tax data to further a national economic imperative, such as the study of innovation and entrepreneurship.[20] The Survey of Consumer Finances' (SCFs) long-standing use of tax data would seem to provide the historical precedent for justifying such work under sections 6108(a) and 6103(n) per the above treatment on existing statutory access mechanisms. That is, the Federal Reserve Board (FRB) serves as Treasury's contractor under 6103(n) for producing a product deemed necessary under 6108(a). (In addition, FRB subcontracts some of the work to the National Opinion Research Center (NORC) at the University of Chicago). The IRS data system, as opposed to IRS data at the U.S. Census Bureau, seems particularly suited to such a study given that IRS data would not only provide the sampling frame but also could provide tax data supplementing survey responses, enabling the analysis of effective tax rates along with complete financial stocks and flows on balance sheets and income statements, even employer-employee data linkages. In addition, this arrangement would avoid the resource cost of more amendments to statute or regulations. Keeping the work based at the IRS also could strengthen the human capital of its statistical and research offices, bolstering the decentralized diversity of the U.S. statistical system.[21] Yet another benefit of keeping the study at the IRS is that risk to confidentiality would be reduced as the data would not need to travel outside IRS' jurisdiction, say, to another agency..

A slightly more complex scenario using both 6108(b) and, eventually, 6108(a), might transpire like this. An outside institution with appropriate gravitas, such as the National Science Foundation (NSF) or the National Bureau of Economic Research (NBER), might be tasked, say, as a result of Congressional petition, to produce a specified study of innovation. NSF/NBER could request that a portion of the study, perhaps the sampling, be done by IRS/SOI under 6108(b), as a special study, even under reimbursement. In turn, the SOI might contract out a portion of the work, including analysis, to a private contractor in the research community. If the product were deemed useful and important enough to repeat, it might continue in 6108(b) status or be converted to an SOI mandate under 6108(a). The disadvantage of the latter option is that reimbursable funding would disappear. This problem might be remedied with additional line item funding for SOI.

One reason for concern about expanding access to microdata for analytical research is the exclusive mandate on tax policy analysis, involving access to FTI, for proposed legislation that seems to be afforded only certain groups, such as 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. The legislative process is complicated and stressful enough without such outside factors intervening in or even appearing to disrupt the process. Another issue

.19 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.

20 See "Studying Innovation in Businesses: New Research Possibilities." Greenia, Husband-Fealing, Lane. Presented at the National Bureau of Economic Research's 2008 Summer Institute, Conference on Research in Income and Wealth.

21 Although the increased efficiency of centralized statistical systems is appealing, they carry the risk of more politicization of data for predetermined decisions. One benefit of a strong and cooperative -- repeat strong and cooperative -- decentralized system is the checks and balances conferred by the very nature of decentralization, even competing estimates across different data files.

with expanded access by outside researchers is that unless the justification is a compelling national economic imperative; e.g., equivalent in gravitas to the SCF, 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 negative connotation among certain circles within the tax world at both the Treasury and the IRS. This connotation raises alarms because it invokes the prospect of data browsing; recently such a problem for the IRS that special legislation was enacted to make it an explicit crime with serious consequences. That wound is still fresh and very sensitive.

However, there is some justification for a different view on outside researcher access. Given that statistical tax data are produced with publicly provided funds and given that publicly available data often are too imprecise ,due to their anonymous form after disclosure processing, to answer economic questions so important for all citizens, why should there not be an outlet for such access? In effect, this might be seen as an additional systemic check for purposes of further "democratizing" the decision process by improving its transparency. Although some might argue that the electorate already has spoken with the process currently in place, including statute, the following factors could argue for at least some additional access by outside economists and analysts:

- First, the analytical questions being answered by the JCT and OTA largely are driven by the political discussion in Congress and the administration. It is possible that other questions – not being actively pursued – also might be relevant to the public debate. 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 the JCT and OTA that there is sometimes less than optimal opportunity to ensure the comprehensive accuracy of analytical results.

- Third, both the 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 the 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, and, of course, there is no public use file of business tax data.

- Fifth, given the millions of dollars spent to collect and process data, both administrative and survey, it is increasingly difficult to defend this cost in conjunction with the benefit of publicly available data that are so severely redacted – cell suppression, blurring, etc. – due to disclosure requirements.

- Sixth, 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 data sets 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 conducted with good data. That is, economic analysis, including "analytical replication," will be attempted one way or the other, and using best-quality confidential tax data would arguably seem the most efficient means of fostering the highest quality analytical research needed to provide decision-makers with critical information. Otherwise, an analytical muddle could be the result with increasing frequency.

- 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 actually can 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. This seems so, particularly if the likely outcome is more informed decisions in both the private and public sectors, with increased utility for society as a whole, especially if that can be accomplished without sacrificing confidentiality.

RECENT ACCESS STEPS: SOI'S OUTSIDE RESEARCHER PROGRAM

The IRS' SOI Division once allowed a limited number of academic researchers to access FTI for certain research purposes. Although the researchers were not monetarily compensated, some expense reimbursement was provided by SOI, such as travel to Washington, D.C., for discussion or data access purposes. This approach was terminated abruptly in the 1990s when the SOI was advised by IRS counsel that such use of the Intergovernmental Personnel Act (IPA) was not appropriate. Counsel had two primary concerns: the lack of a statute in tax law to authorize such access and the lack of a competitive allocation process. Counsel also had a concern about the possible perception of a quid pro quo unless the customer and provider roles were delineated clearly in a contract underpinned by monetary compensation for a specific deliverable.

As a result, the recent mechanism for outside access has been very limited, occurring rarely and only through section 6103(n) contracts. However, in 2008 the SOI began another approach for outside access that uses this same statute and retains a competitive bidding process but allows the specifics to be determined by the outside researcher. The advantage of this outside researcher access program is that it is relatively unconstrained by the fixed deliverable requirement of most contracts. Four criteria must be met, however:

1. The proposed work must have utility for tax administration [22] and the larger statistical community
2. The researcher must be familiar with the SOI confidentiality culture and products [23]
3. The researcher must contribute to the development of the SOI's human capital through partnering with SOI staff on the work or by teaching a course to SOI staff
4. The researcher must have strong academic qualifications and experience.

Although only two awards were made in the initial year, the SOI expects the program to be annual and awards to be more numerous. While the program is limited in scope due to resource constraints, it does represent a step forward towards expanding access to FTI for research.

NEXT STEPS

Two fundamental principles would seem to argue for expanding the partnership between economic analysts and administrators of the tax system. First, tax administration needs and the needs of economic analysts are complementary, not mutually exclusive. Second, tax data quality is proportionate to the level of rigorous microdata analytical use, including peer review. The researcher community, in cooperation with the statistical community, needs to be more proactive not only in requesting access to tax microdata but in providing justifications that include benefit to the tax agency and the decision-making community. Simply demanding data for professional self-aggrandizement is unacceptable.

This paper has provided several examples of areas in data set architecture that need improvement for purposes of both tax administration and economic analysis. There are surely others. Of course, the key to improving any data set is the regular use and rigorous analysis of the data to answer complex questions that cannot be answered with aggregate data, especially across time and for entities, such as businesses that not only change their names and EINs but their operating characteristics, too. In the dynamic tax world, something noteworthy always is happening, and only the continuous and complex analysis of tax data can ensure the data's continuous improvement.

In addition, researchers, in cooperation with the statistical community, need to participate in federal agency advisory groups. It is not necessary to be a member of the group in order to participate in or attend these

22 This can be for the statistical or research component of tax administration.

23 This requirement is due to the SOI's limited resources available at SOI to train researchers.

advisory group meetings. This participation also will help ensure that researchers are informed by insiders regarding technical aspects of their desired undertakings, not to mention issues critical to the federal agencies with stewardship over the data needed. In turn, it is incumbent upon the researcher community to note important analytical work that cannot be conducted due to microdata inaccessibility.

Finally, researchers need to participate in existing researcher access programs[24] in order to demonstrate that there is not only a compelling analytical need for such programs but also that the participating agencies are thereby helped. This participation also is needed to help nurture the intellectual capital of the statistical agencies' staffs, not only as processors of the data but as analysts, as well.

24 These include not only the U.S. Census Bureau's Center for Economic Studies, but also the SOI's recently established program that requests proposals for projects conducted as contractual arrangements under section 6103(n) of the tax code. The latter currently are requested annually – usually in the fall – with notices in the Federal Business Opportunities publication,.