A Secure Query System to Improve Access to Individual Income Tax Data

Amy O'Hara, Stephanie Straus, Ron Borzekowski,
Paul Arnsberger, and Barry Johnson

Submitted to Data Privacy volume edited by Ruobin Gong, V. Joseph Hotz, and Ian M. Schmutte

NBER

08/15/2024

Introduction

In recent years important and headline-grabbing findings have emerged from research using individual income tax data for statistical purposes. Demand for these microdata, accessible under the tax administration authority of the Internal Revenue Code and through the IRS Statistics of Income (SOI) Division's Joint Statistical Research Program, continues to grow. This paper describes a new approach to address demand from government agencies and nonprofit institutions for such statistics.

The project explores the feasibility of a privacy preserving secure query system (SQS) linking end-users of the data, a data intermediary, and SOI. In the early stages of development, end-users may be state or local government agencies or nonprofit institutions (e.g., non-degree programs at community colleges); the intermediary is Georgetown University; and all processing will be done within and by SOI staff. The objective is for an SQS client, such as a state department of social services, to prepare and submit a dataset with personal identifiers for SOI to match to individual income tax records, in order to produce tables of predefined output statistics. This efficient and automated process should allow greater production of evidence at much lower cost and burden for clients and SOI.

Building on SOI Tiered Access

The Internal Revenue Service (IRS) Statistics of Income (SOI) Division is the Federal Statistical Agency within the Treasury Department responsible for transforming administrative tax records into statistical products. SOI has long been a leader within the Federal Statistical System in responsibly releasing administrative data for research use. Technological advances in data capture, retention, and processing allow the IRS to manage ever-increasing amounts of administrative data that have great potential for use in policymaking. However, access is constrained due to legal barriers, logistical challenges, resource constraints, and privacy concerns. Most notably, aggregate statistics can only be published provided that no federal tax information (FTI) is revealed.

To meet these challenges, SOI has been a pioneer in tiered access, providing varied data access and products while mitigating risks and maximizing utility for end users. The SQS proposed in this paper is the newest access mode for government agencies or nonprofit institutions seeking statistics from federal tax information (FTI), while relying on SOI staff for all secure FTI processing. Below, this paper describes how this privacy preserving SQS aligns with SOI's past and current deployments of tiered

access models, building on their existing projects while also reducing some of these projects' barriers to access.

Current Models for Providing Access

Individual income tax data cover the vast majority of Americans. Tax returns contain limited demographic information (i.e., marital status and number of claimed dependents), and geographic information from mailing addresses (Slemrod, 2016; Vartivarian et al., 2007; Larrimore et al., 2017). Tax data are collected with a great deal of specificity across different income types (available in instructions, worksheets, and forms). All federal tax filers use these forms and worksheets, making this documented and generally uniform data source valuable for comparisons across geographies or over time.

SOI manages tiers of access to tax data, from the most restrictive, closed data to fully open data – and they have been doing so for over sixty years. This paper first describes SOI's current models for providing data access, weighing their benefits and challenges and assessing their reach across external users. The paper then describes the SQS, and how it provides approved requestors with greater access to administrative tax statistics, while retaining privacy protections.

Open Data SOI Products

SOI generates regular statistical series, data books, and special studies using FTI. SOI has changed these products and the data they make available over time, with some expansions in response to user requests and retractions due to privacy concerns. Key examples of SOI open data products are the annual ZIP Code level tax return data, US population migration data, and county income data products. These datasets are aggregated, containing no individual tax return information, and are therefore publishable outside the agency.

Joint Statistical Research Program (JSRP)

SOI established a Joint Statistical Research Program (JSRP) in 2012, permitting selected researchers access to federal tax microdata through a competitive proposal review process. Described in ACDEB report as "a small researcher access program which has yielded groundbreaking studies, as a partial solution toward expanding evidence-building capacity," the JSRP program proves useful for those whose projects are selected (ACDEB, 2022). And to date, the JSRP has increased use of SOI's data to support

tax administration and advancing knowledge of how taxes affect people, businesses, and the economy (Slemrod, 2016).

However, given the sensitivity of tax return data and the potential damage from a breach, the program's data access controls are stringent. The number of research teams granted access is modest, with 25 projects selected in 2023 out of 114 applications. Researchers on approved projects must undergo a thorough background clearance and are bound by the same rules and penalties as IRS employees. Each approved project is assigned to SOI subject matter experts who monitor all aspects of the project to ensure compliance with data access and security standards as well as project scope. This adds burden to SOI employees who must provide monitoring, technical assistance, and disclosure avoidance review to the selected researchers who take on Intergovernmental Personnel Act (IPA) assignments during their four-year projects.

SOI Public Use File (PUF)

Based on a subsample of individual income tax returns, the SOI Public Use Files (PUF) have been a critical resource for economic and tax policy researchers, enabling studies of the effects of tax policy changes, the distribution of tax burdens, and economic incentives (Bowen et al., 2020). The PUF contains anonymized and altered individual income tax return records to protect against the risk of disclosure. Researchers using the PUF agree to terms of use (e.g. pay fee, do not attempt to re-identify any taxpayers, do not share their copy of the PUF). However, the growing availability of external data sources that could be linked to the PUF, as well as increasing technological capabilities to re-identify individuals in the sample, forced the IRS to distort the data in increasingly aggressive ways, making the data less useful for analysis (Bowen et al., 2020).

SOI Synthetic Public Use File (PUF)

In response to growing threats over use of actual tax records in the PUF, SOI partnered with the Urban Institute and Tax Policy Center to develop a synthetic PUF using statistical methods to generate fake data that mirror the underlying IRS data (Burman et al., 2024). There are a number of challenges when creating and releasing a synthetic PUF. The structure of tax returns complicates the production of synthetic data, since virtually all variables are related through accounting relationships and nonlinear tax computations across the forms and schedules. Synthesizing a specific variable, such as total income, necessarily involves synthesizing its components, such as wages and salaries, but relationships between

all variables must be preserved for the synthetic PUF to be useful (Vartivarian et al., 2007). Despite these difficulties, the team at the Urban Institute and SOI persevered. They have produced fully synthetic supplemental PUFs for tax years 2012-2015, based on information returns such as forms W-2 for reporting wages and Forms 1099 that report pensions, interest, dividends, etc., for individuals whose income is likely below the tax filing threshold. They have also produced a fully synthetic beta Tax Year 2015 Form 1040 file that is being tested by current PUF users.

College Scorecard

SOI supports projects that need custom tabulations of FTI for tax administration purposes and for special studies. One example is the College Scorecard, a joint project with the US Department of Education (ED) that produces aggregate earnings data for cohorts after graduation. Statistics are produced for students who received federal financial aid, by U.S. postsecondary school, credential level, and program of study. Data are displayed in a web tool "designed to help students make informed decisions about their education options after high school, bringing together information on college costs, graduation rates, student loan debt, post-college earnings, loan repayment rates, and more" (Kaouk et al., 2021). To release this fine-grained information, SOI developed a disclosure avoidance protocol using differential privacy that reduces re-identification risks for students and programs. The protocol relies on SOI and ED determining the acceptable tradeoff between privacy and accuracy.

As described above, SOI's efforts on open data resources, including the JSRP, PUF and synthetic PUF, and College Scorecard meet many user needs for tax and income information. However, these products fail to address the measurement needs of program administrators, evaluators, and policymakers who often seek information about specific cohorts (where the PUF or synPUF may not have sufficient sample) or need information that is more timely (where the JSRP is too costly a process).

Secure Query System (SQS)

Georgetown is assisting SOI with design options for the SQS, relying on outreach conducted with Yale University to federal agencies, state agencies, local governments, and nonprofit organizations. In describing their needs for metrics about tax filing behavior and income, these organizations cited benchmarking, evidence-based budgeting, informing workforce development programs, increasing economic mobility, and understanding labor flows across state borders as reasons for seeking SQS outputs. State officials were eager to explain these varied needs, and to specifically explain gaps in their

current measures. For example, state education and workforce officials need to know whether their learners are earners. Most are using their state Unemployment Insurance wage data but lack visibility for out-of-state earners and non-employer earnings. They are looking for aggregate statistics that show the margin of missing earnings, and they are looking for indicators about the extent of out-of-state earnings. State health and human services officials lack visibility into tax filing and credit claiming behavior; they cannot measure whether outreach efforts are inducing more state residents to file and claim refundable credits. State justice officials lack information on training programs prior to prisoner re-entry. State economic development officials lack insights about retention of in-state college graduates and career pathways. These officials also seek information on industries in which learners or beneficiaries work and the impact of program changes and interventions. Some are focused entirely on the individual, and others are interested in tax units.

These discussions informed Georgetown's SQS process design and choice of planned output statistics. The initial set of SQS-1040 output statistics to be produced, using individual income tax data linked to the client's input file, include: percent of records that filed 1040, percent that claimed each the Earned Income Tax Credit (EITC) and Child Tax Credit (CTC), average EITC and CTC amounts, as well as mean and quartile statistics of Adjusted Gross Income, broken out by filing status. The initial set of SQS-Info output statistics, produced using combined Form W-2 and Form 1099-NEC data that matched to the client input file , includes mean and quartile statistics for wages, non-employer compensation earnings, and total earnings (W-2 and 1099-NEC combined), as well as descriptive statistics about the prevalence of non-employer compensation forms linked to the cohort.

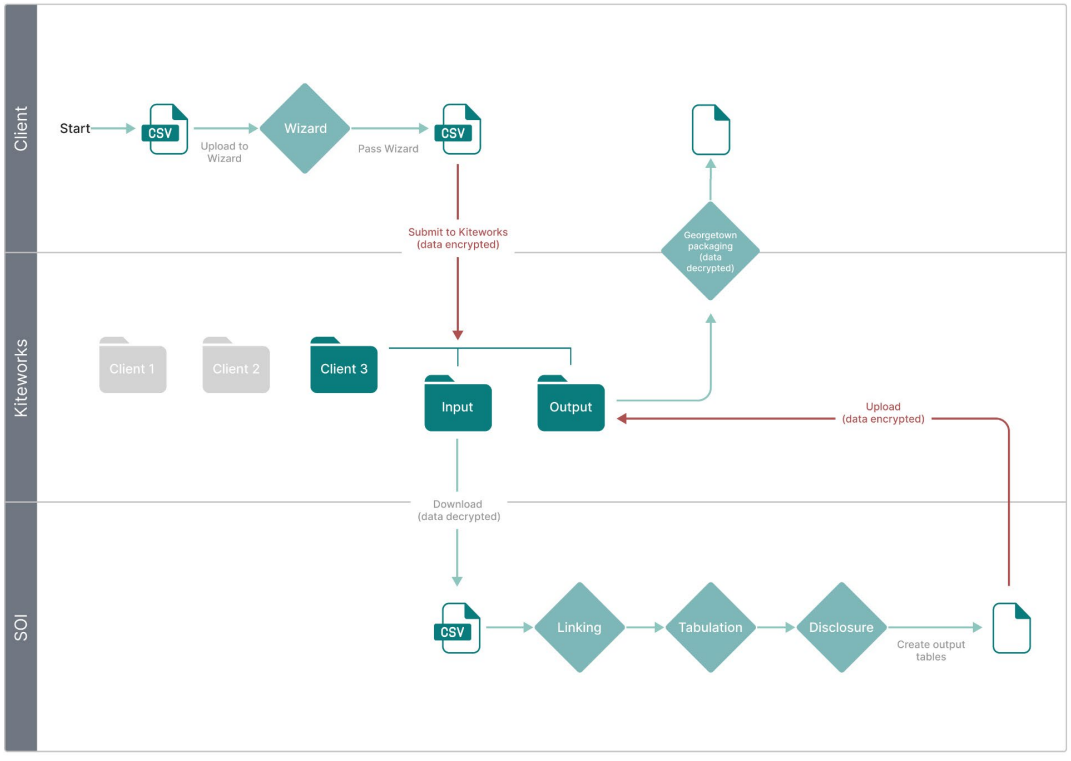
SQS Process Overview

To establish the SQS, SOI and Georgetown University, the temporary intermediary, are developing options for client data submissions, data linkages inside IRS, statistical analyses and tabulations, and disclosure avoidance methods. In parallel, Georgetown is working with Yale and potential clients to confirm that the pre-defined SQS outputs will meet their measurement needs. As illustrated in Figure 1, the proposed workflow includes:

- Clients validate, using a tool Georgetown is developing, their input data to verify the presence of sufficient identifying information for linkage and adequate cell sizes for output statistics.

- When potential clients pass these validation checks, they enter a standard agreement with Georgetown as the intermediary[1] and prepare their data extract for submission, conforming to a common data model.
- Georgetown facilitates transfer of encrypted client input files to SOI without seeing the file contents.
- SOI retrieves the client data and runs an automated matching, data transformation, and tabulation process. The automated matching occurs between the client data and individual tax data derived from information returns (Forms W-2, 1099) as well as individual income tax returns (1040).
- SOI populates the pre-defined output tables and applies privacy protections. The resulting aggregate statistics (no longer Federal Tax Information) are transferred back to the clients through Georgetown.
- Once Georgetown confirms receipt of the output tables, SOI destroys the input files. Georgetown distributes the output tables back to the clients, and documents SQS inquiries, submissions, and completions for SOI.

Figure 1: SQS Data Flow



[1] Note: A separate legal agreement was also developed between Georgetown (the intermediary) and SOI in order to govern the SQS design process.

Having Georgetown act as intermediary in this process removes SOI from the burden of negotiating one-off legal agreements, handling data transfers and receipts that do not conform to agreed standards, recoding input files to align with SOI schema, and designing disclosure review protocols on varying output statistics. SQS is an opportunity to standardize all aspects of the external data access approach, from the first request of statistics by the outside agency or organization, down to the disclosure avoidance protocols before the statistics are released. By rethinking the access process and building it to scale, SOI will achieve greater equity, transparency, efficiency, and risk reduction.

How the SQS Differs from Existing Models

Within the SQS, shifting some responsibilities and constraining both inputs and outputs allows for automation, scaling, and efficient use of tax data for these analytic and policy goals. The bulk of data preparation and alignment work is carried out by the potential clients in state and local governments and nonprofits, and the design and administrative burden are handled by Georgetown as the initial intermediary. Instead of SOI staff negotiating separate data sharing agreements with organizations wishing to have their data matched to produce tax and income statistics, SQS has a single template for clients to request SQS outputs. Risk is further minimized with SQS, as only SOI employees use FTI and limited predefined outputs are produced, after passing a standardized and automated disclosure review process.

Effective Privacy Protection

The outreach to potential SQS clients—state and local government agencies, academic institutions, non-profit service providers, and evaluators—informed how they view the trade-offs of various disclosure avoidance methods. Georgetown and Yale provided to clients examples of the different levels of granularity and types of output statistics they could receive, and how to avoid receiving tables full of suppressed values. The potential SQS clients confirmed their eagerness to obtain privacy-protected output statistics, and their willingness to participate in conversations about privacy trade-offs. They asked about metrics to quantify benefits and risks when comparing disclosure avoidance methods.

SOI must abide by its laws and guidelines, including those in IRS Publication 1075, that, “Statistical reports may only be released in a form that cannot be associated with, or otherwise identify, directly or indirectly, a particular taxpayer” (IRS, 2021). Georgetown explained these constraints and found that the potential SQS clients understood the need for coarsening, suppression, or noise injection to obscure

observations in output statistics. While they would like more information on how noise infusion would affect output utility, the potential clients supported a reduction in precision of estimates in order to gain strong privacy protections. They confirmed their willingness to receive (and the usefulness of) the aggregate tax and earnings information with values rounded to the nearest hundreds place, with the understanding that further coarsening may be needed (to 500s or 1000s, depending on the statistic). They also understood that percentages may be released as ranges, that quartiles may not include the bookends of value ranges, and that only certain mean income statistics would come with standard deviations. Finally, they understood that cell size limitations may trigger SOI suppression rules in cases where counts do not allow publication.

These tradeoffs between utility and privacy illustrate the implications of using privacy protected data for estimation and inference when conducting applied research. Potential clients provided input on statistics of interest; Georgetown confirmed feasibility with SOI, and shared possible rounding and suppression approaches with potential clients. When the need arises, differentially private statistics will be developed and socialized with potential clients.

In developing the disclosure avoidance protocols for SQS, Georgetown learned that the most vital aspect was trust of the end users. The potential SQS clients confirmed they are willing to provide their highly sensitive data for linkage to SOI data and accept imprecise aggregate statistics because a value proposition exists. Further, SQS success is measured not only by the number of clients who successfully input their data and receive useful statistics back, but also by minimizing the risk and burden SOI faces as they operate the SQS.

Legal Authority and Policy Environment

IRS Authority

SQS can be conducted under I.R.C. Section 6108(b), stating that: "The Secretary may, upon written request by any party or parties, make special statistical studies and compilations involving return information (as defined in section 6103(b)(2)) and furnish to such party or parties transcripts of any such special statistical study or compilation. A reasonable fee may be prescribed for the cost of the work or services performed for such party or parties." Special Statistical Studies must be conducted by SOI employees; they cannot rely on IPAs or contractors. Other projects, including the College Scorecard, have been conducted under this authority.

Fulfilling the Evidence Act Provisions

The Commission for Evidence-Based Policymaking (CEP) noted in its Final Report that complex legal regulations and internal agency policies "limit the effective, efficient, and transparent use of existing data" (Abraham et al., 2017). CEP identified tiered access as a potential solution that balances greater access and robust privacy protections. Considering this new tier of access via the SQS should help SOI further comply with the Foundations for Evidence-Based Policymaking Act, which pushes toward a presumption of accessibility. The Commission recommended that Federal departments consider the sensitivity of the data, with input from stakeholders, including researchers and privacy advocates, to establish access restrictions based on law, context, and sensitivity.

In 2018, the passage of the Foundations for Evidence-Based Policymaking Act (Public Law 115–435) advanced many of the recommendations of the CEP report. For example, the Evidence Act, as PL 115-435 is known, requires agencies to develop learning agendas, inventory their data assets, name statistical officials, chief data officers, and chief evaluation officers. The Evidence Act also required a committee to research whether and how the U.S. could establish a National Secure Data Service. The proposed, ambitious use of PETs in the SQS could inform the National Secure Data Service and, importantly, what services it may offer. The Advisory Committee on Data for Evidence Building (ACDEB) gathered evidence and proposed recommendations about improving secure and efficient access to government data, through a National Secure Data Service as well as other means.

SQS identifies what is important to clients and mitigates risks of producing those statistics, aligning to the ACDEB recommendations calling for use of risk-utility frameworks, and informing both OMB and NSDS on privacy preserving record linkages (PPRL). Research is underway at Georgetown and through NSDS Demonstration Projects to understand whether clients have sufficient data quality and technical readiness to adopt PPRL. Should the SQS develop toward a process where linkage is done with hashed identifiers instead of plaintext identifiers, privacy protections would be further enhanced. The SQS can demonstrate the value of coordinating with federal, state, territorial, local, and tribal government officials seeking linkage services, and through the strategic use of an intermediary to offset burden on SOI staff.

ACDEB called for models of shared responsibility in Recommendation 1.8, and a shared responsibility model is embedded in SQS client DUAs. Clients must attest that they have authority to share data, demonstrate that their data conform to a common SQS schema, and agree not to attempt re-identification to learn about fact of filing or FTI details. SOI shares responsibility to prevent release of FTI,

acknowledging that not all data are equally sensitive, to guide their application of disclosure avoidance tools to SQS output tables. This relates to another ACDEB recommendation which calls on the NSDS to use realistic risk models. Until the NSDS exists, SOI is asking important questions including: What could a client learn based on the output tables? How could a potential bad actor exploit the system? SOI and Georgetown have planned an expert review of SQS disclosure methods prior to production.

Even as the NSDS and other initiatives work to improve access, the SQS team has identified a real and tangible need for evidence that can be met with a standardized and automated query service. This need will likely exist even if, and when, greater access to underlying data comes online. Each of these will have an important role in the future toolkit for evidence-based policymaking.

Cost and Sustainability

There is a perception that expanded access to FTI requires legislative change and radical increases in IRS funding and infrastructure. SQS demonstrates that academic-government partnerships can accelerate research and development, and that agencies have authority to establish new tiers of access without legislative change. SQS shifts burden from SOI in two ways: SQS relies on clients to review and recode their own data, prepping it for submission to SOI, and it relies on Georgetown (as intermediary) to manage the administrative aspects of the project. Sharing the burden and costs of innovation makes it possible. Still, SQS is not costless for SOI. Depending on the frequency and scale of SQS, SOI will need to plan for cost recovery.

To inform cost models, Georgetown will capture useful metrics from pilot queries, including successes and failures of the input data validation system, legal and agreement issues, and data quality challenges that may hamper or hasten client interest in SQS output tables. Outreach with potential clients indicates high interest in additional individual income SQS queries, including more sophisticated analyses that handle lagged matches (computing outcomes on older cohorts), interventions including data from randomized controlled trials, and specific populations (e.g., learners on credential pathways).

These innovations require funding for SOI staff to run the queries; funding is also needed for research and development, outreach, capacity building in state and local agencies, whether these activities take place within SOI or with assistance from an intermediary. The SQS model can expand within SOI to corporate tax data and can inform efforts at other federal agencies.

Conclusion

The SQS, with the trade-offs inherent in the proposed system, has the potential to increase access to valuable individual income tax data for statistical purposes. Notably, scaling, automation, and efficiency come at the cost of standardization and limited user customization; privacy and security requirements may come at the cost of utility for some user needs. However, outreach to date has indicated that the proposed SQS properly balances these concerns and should therefore offer another useful mode of information provision for SOI as they seek to advance their mission.

References

Abraham, K., Hasking, R., Glied, S., Groves, R., Hahn, R., Hoynes, H., Liebman, J., Meyer, B., Ohm, P., Potok, N., Mosier, K., Shea, R., Sweeney, L., Troske, K., & Wallin, K. (2017). The Promise of Evidence-Based Policymaking Report of the Commission on Evidence-Based Policymaking. https://www2.census.gov/adrm/fesac/2017-12-15/Abraham-CEP-final-report.pdf

Advisory Committee on Data for Evidence Building (2022). Year 2 Report Supplemental Information. https://www.bea.gov/system/files/2022-10/supplemental-acdeb-year-2-report.pdf

Bowen, C. M., Bryant, V., Burman, L., Khitatrakun, S., McClelland, R., Stallworth, P., & Williams, A. R. (2020). A synthetic supplemental public use file of low-income information return data: methodology, utility, and privacy implications. In *Privacy in Statistical Databases: UNESCO Chair in Data Privacy, International Conference, PSD 2020, Tarragona, Spain, September 23–25, 2020, Proceedings* (pp. 257-270). Springer International Publishing.

Burman, L., Johnson, B., MacDonald, G., (others TBD). (2024). Protecting Privacy and Expanding Access in a Modern Administrative Tax Data System, National Tax Journal (forthcoming).

Hauer, M., & Byars, J. (2019). IRS county-to-county migration data, 1990–2010. *Demographic Research*, *40*, 1153-1166.

Internal Revenue Service. (2021, November). Publication 1075: Tax Information Security Guidelines For Federal, State and Local Agencies, Safeguards for Protecting Federal Tax Returns and Return Information https://www.irs.gov/pub/irs-pdf/p1075.pdf

Kaouk, T., Fortelny, G., & Allen, R. (2021). Chief Data Officers Presentation for the Advisory Committee on Data for Evidence Building. https://www.bea.gov/system/files/2021-02/OCDO-ACDEBpresentation-FEB2021.pdf

Larrimore, J., Mortenson, J., & Splinter, D. (2021). Household incomes in tax data: using addresses to move from tax-unit to household income distributions. *Journal of Human Resources*, *56*(2), 600-631.

Slemrod, J. (2016). Caveats to the research use of tax-return administrative data. *National Tax Journal*, *69*(4), 1003-1020.

Vartivarian, S., Czajka, J. L., & Weber, M. (2007, July). Measuring Disclosure Risk and an Examination of the Possibilities of Using Synthetic Data in the Individual Income Tax Return Public Use File. In *American Statistical Association Meetings*.