Exploring Improved Access to Individual Income Tax Data for Evidence-Building

Amy O'Hara, Stephanie Straus, Ron Borzekowski

December 2023

In recent years, important and headline-grabbing findings have emerged from research using individual income tax data for statistical purposes. Demand for these data, accessible under the tax administration authority of the Internal Revenue Code and through programs such as the Joint Statistical Research Projects in the IRS Statistics of Income (SOI) Division, continues to grow. However, such approaches constrain access to these data and impose substantial resource costs and risks on the public organizations providing the data.

This paper describes an alternative approach, using a trusted service provider and trusted PETs, to allow access to, and use of, tax information for critical applied research while recognizing the need to protect privacy. The project explores the feasibility of a privacy preserving secure query system (SQS) linking end-users of the data, a data intermediary and a data provider. In the early stages of development, the end-users will be state or local governments, or nonprofit institutions; the intermediary will be at an academic institution; and all processing will be done within and by the IRS. For example, a local government will use tools developed in this project to produce a dataset with personal identifiers and any tabulation groups necessary for their analysis. They will send those files to the intermediary for secure transfer to the IRS. Using approaches developed in this project, the IRS will match these data to individual income tax information and produce privacy protected standardized output tables. The automated system should run efficiently enough to allow greater production of evidence at much lower cost to the data providing entities.

There are several key components of the system being developed in this project. First, the system imposes a strict schema and various quality controls on input data, along with the technology to enforce these. While the data can come from varied sources (and therefore address a broad array of questions), the automated system requires strict controls on the input file (e.g., format, completeness). Second, a common data model supports accurate and ultimately privacy protected record linkages. Third, the system produces limited output data using a combination of privacy preserving methods through automated disclosure review protocols. Enabling these steps, the intermediary manages the legal, security, and administrative aspects of the system through agreements with the IRS and end-users.

The system's privacy protecting technologies and automated disclosure reviews ensure compliance with IRS statutory and regulatory requirements. The IRS can use their data under existing tax administration authority under Internal Revenue Code (IRC) 6103(n) and special statistical studies (IRC 6108(b)). This project considers outputs from a secure query system as compilations of tax information under IRC 6108(b). The project focuses on individual tax data derived from information returns (Forms W-2, 1099) as well as individual returns (1040). Future efforts could possibly utilize business returns, for example to measure outcomes for the Inflation Reduction Act. However, in all cases, the outputs of the system must meet the agency's disclosure guidelines and, notably, cannot reveal fact-of-filing or any specific federal taxpayer information. This may limit the outputs in terms of geography or other information that could be disclosive. Current laws penalize IRS employees if Federal Tax Information is published and this results in rigid and conservative approaches to output privacy protection. The project will explore

potential changes to laws that would shift and share penalties for improperly using tax information to identify taxpayers.

Successful development and deployment of the query system should support both the IRS and the user community. Most notably, an automated system would offer substantial reductions in the administrative burden at the IRS. Currently, projects at the IRS require tailored one-off data use agreements for each project; customized delivery of input files that then require data cleaning and preparation within the IRS often by contractors overseen by IRS employees; and manual disclosure review. Such a process is burdensome and cannot scale, leaving only a limited number of researchers with access to this valuable resource.

The automated system would also limit the risks of disclosure while furthering the mission of the IRS. Fewer individuals accessing the data and standardized outputs based on automated review should both limit opportunities for any data leakage. Further, should the system develop toward a process where linkage is done with hashed identifiers, privacy protections will be even further enhanced. Considering this new tier of access should help the IRS further comply with the Evidence Act which pushes toward a presumption of accessibility. This ambitious use of PETs could inform the National Secure Data Service and what services it may offer.

While this approach to automation and privacy protection limits the types and depth of available analyses, end-user outreach to date indicates strong support to the system as described in the paper. A number of states and non-profit institutions indicate that they are willing to agree to the terms, can provide data necessary to make this approach viable and would benefit from the, admittedly, limited standardized output from the system.

SOI has been exploring and implementing privacy enhancing technologies (i.e. synthetic data for non-filers and public use file, differential privacy for the College Scorecard) to increase access to tax data. The secure query system adds a new tier of access, enabling statistical analysis of valuable, restricted federal administrative data to support evidence-building needs.