Statistical Journal of the IAOS 34 (2018) 19–23
DOI 10.3233/SJI-170408
IOS Press

SOI at 100: Traditionally enumerative, now more and more analytic

Fritz Scheuren
Retired SOI director
E-mail: scheuren@aol.com

The U.S. Internal Revenue Service's Statistics of Income program provides comprehensive data and statistics on the federal tax system. Data produced by the Division underpin the analytical work that supports federal budget and tax policy formulation. Statistics and analyses, disseminated on www.irs.gov/taxstats, provide researchers, the media, businesses and ordinary citizens unbiased information to better understand current and proposed tax law changes and their effects on the economy. As one of 13 principal agencies in the decentralized U.S. statistical system, SOI also collaborates to improve the timeliness, utility and cost effectiveness of statistical information produced throughout the federal government.

This special edition of the Journal of the International Association for Official Statistics focusing on the SOI program is primarily intended to showcase recent work and suggest new directions for the Division's future. SOI has recently celebrated its 100th year, so it is especially appropriate to look back briefly at its history, sketch where SOI is now and consider where SOI seems to be heading. SOI is not unique in its evolution or in the challenges it currently faces, and so it is hoped that these pages will serve not just to recognize its 100th anniversary, but more importantly, to suggest ideas that will help other producers of official statistics evolve.

As the papers that follow demonstrate, SOI's processes and SOI products have evolved over time and continue to do so. Below, I will briefly outline SOI's progress from its inception in the early 20th Century to present. Papers that follow will highlight the application of statistical techniques to improve the quality, timeliness and coverage of statistical estimates derived from SOI samples of tax returns. The increased availability of large administrative datasets of tax and personal information afforded by increased electronic filing of personal and business tax information are providing opportunities to reexamine SOI products and processes. Some early fruits of these efforts are presented, as well as some plans for proposed future uses.

On a personal note, as a former Director of Statistics of Income, I am proud to have been associated with SOI for more than 17 of my 31 plus years of Federal service. In fact, I worked at SOI twice: first as a 20-something statistician in the early 1960s, where I produced the first comprehensive SOI estimates of personal wealth using data from federal estate tax returns, and second returning in 1982 as the SOI Director. I continue my close association with SOI as a member of an advisory board that helps ensure that data users' perspectives are represented in the Division's planning efforts. Given my long association, please forgive the sometimes-personal nature of what follows.

Introduction and organizational background

I would like to start then with some pre-SOI history. There had been a federal income tax during the American Civil War in the 1860s, but it was eventually ruled unconstitutional and abolished in 1872. The income tax was re-instituted in 1913, after the enactment of the 16th amendment to the U.S. Constitution, one of a series of progressive constitutional amendments added around that time.[1]

[1] IRS Historical Fact Book: A Chronology: 1646–1992, Department of Treasury: Internal Revenue Service.

Three years later, the Revenue Act of 1916 included the requirement that "[T]he Secretary [of the U.S. Treasury] prepare and publish not less than annually statistics reasonably available with respect to the operations of the internal revenue laws, including classifications of taxpayers and of income, the amounts claimed or allowed as deductions, exemptions, and credits, and any other facts deemed pertinent and valuable," creating the forerunner of what we know as SOI.[2] In fact, the name of the program came out of SOI's statement of legislative purpose.

SOI enumerative statistics

The statistical products of SOI have historically been largely enumerative in nature and based on enumerative sample designs. The word "enumerative" in our use here means that the goal is the same as if a complete census had been done. The original purpose of the SOI program was to provide statistically reliable descriptive summaries on the operation of the then newly re-federalized tax system. The original focus was primarily on the income details reported on individual and business tax returns filed annually; the first report, published in 1918, included statistics on both filing populations for calendar year 1916.

Under the Internal Revenue Code the IRS is required to disclose individual taxpayer data to the Treasury Department and the Congressional Joint Committee on Taxation (JCT) for tax administration proposes. Historically, SOI has closely collaborated with these organizations in determining the content and timing of statistical studies in support of tax policy development and budget formulation. Early on, SOI developed a core set of statistical products for release annually. Initially, the main products produced comprised books of tables. For example, tallies of items from individual returns, with subtotals by district office (or state) were included among the early tables, along with some summary measures of economic activity. Examples of economic activity included total or adjusted gross income and, for corporations, total assets, gross receipts, and net income/loss.

For practical reasons, annual SOI individual income tax return statistics were based on samples that were systematic and stratified, but for all intents and purposes, treated as statistically random stratified samples.[3] Samples were manually selected, using a stratified design based on where the return was filed and size of total income, replaced by adjusted gross income beginning in 1944. Returns were selected systematically-for example, every *nth* return processed by the IRS was selected for the SOI program. Still later, other stratifiers were added and special individual studies (e.g., capital gains) undertaken. Initially, the sampling of the individual returns was done in the local offices.

Other returns, (e.g., business, estate and gift tax returns) were not generally sampled until the 1950s, so early annual statistical products were based on the full population of the returns filed. It is well known that there is a crossover point where the costs of the sample design work can exceed the cost of just processing the entire population. As these filing populations were a lot smaller and more complex, I suspect this was a factor that precluded the sampling of these filing populations.

Ernie Enguist, a former SOI Director, may have been responsible for the decision to begin sampling the corporate returns in 1951 in response to tremendous growth in the number of corporation returns filed annually. The samples were stratified by industry and asset size. It is said that Dr. Enguist worked nights and weekends to smooth out the transition from population-based to sample-based estimates to preserve the time-series by industry and size of total assets.

SOI's core products were occasionally supplemented by special studies focusing on smaller filing populations or subsets of the filing populations, for example sole proprietorships. Over time, the frequency with which these types of studies were requested by Treasury and JCT increased.[4] Sometimes SOI staff were augmented to produce such reports. For example, in the 1930s, there were tabulations produced of corporate returns by the Works Progress Administration (WPA), a temporary New Deal agency. These so-called "Source Books" were continued afterwards in the regu-

[2] Internal Revenue Code §6108.

[3] For a discussion of SOI individual income tax samples, see: Weber, Michael E., Paris, David P. and Sailer, Peter J. (2008) "Statistics from Individual Income Tax Returns: Populations, Samples, and Processing of Individual Income Tax Returns at Statistics of Income". *Proceedings of the 2008 American Statistical Association Annual Meeting*.

[4] Section 6103 of the Internal Revenue Code provides both the Treasury Department and the Congressional Joint Committee on Taxation access to tax data to support tax administration-related analysis. SOI works closely with these organizations to develop and provide carefully curated statistical data for major tax filing populations. These data are also used to produce most of SOI's publicly released products.

lar SOI Corporate program, and they are still produced today.

For much of SOI's history, all returns used to produce statistics were shipped to Washington for data abstraction and statistical editing and cleaning. At the beginning, the SOI staff was almost entirely clerical, and the workforce was almost all female. While most managerial positions were help by males, a number of women emerged as early leaders in the Division's efforts. Unlike today, women seldom rose to senior management positions. Over the years, the SOI work force has seen an influx of diverse, highly-skilled employees. At the same time, the Division has adopted a variety of modernized work practices, including providing increased flexibility with work schedules and locations. One measure of the successful evolution of the Division's employment practices is the fact that employees who work for SOI tend to stay at SOI. It is not unusual for employees to spend their entire working career with the Division, with many, especially women, working at SOI for more than 40 years.

We do not have many details about how early SOI tabulations were compiled. It is likely that the SOI staff employed a three-step process even at the beginning, just like they did when I got there in 1963. The first step began with selecting the returns to be transcribed onto "edit sheets", so named because the material on the return was examined for consistency before being entered on an edit sheet. Typically, there was further checking using what were called consistency tests that ensured mathematical consistency. Initially these would have been manually performed, later they were automated. The manual processing was tightly controlled down to the hardness of the pencils (no pens were allowed). As I remember it, a Number 2 (medium hard) pencil was used. Initially, tabulations may have computed by hand, but we know that Hermann Hollerith, at the Census Bureau, began the Federal government's use of mechanical card tabulating machinery or "Tab" equipment for the 1890 Census.[5] IRS likely used this equipment very early in the SOI Program. I still remember seeing this equipment in the early 1960s when I began at SOI.

In the 1960's, there was a strong cooperative relationship between SOI and the US Census Bureau. SOI had purchased a one-third interest in the Bureau's Univac 1105 computer and SOI used it on the night shift for years.[6] The edit sheets were keyed onto 80-column punch cards, which could be sorted by the stratifiers and then tabulated. Lil Dorsey, one of many SOI heroes I was to know, often recounted how hot, loud, and prone to breaking down the Univac was in the days before transistors. Still, the Univac was a great advance.

The quality of the SOI tabulations was maintained by thorough training and later through the adoption of modern quality assurance/quality control sampling procedures. In fact, the formal quality SOI program seems to have been stepped up a notch when IRS centralized tax return processing through the introduction of 10 geographically dispersed processing centers. Consequently, SOI shifted its statistical data capture operations from Washington to the 10 centers. Decentralization of SOI operations increased non-sampling error due to differences in interpreting data collection instructions and taxpayer intent that arose across processing sites. This was addressed by implementing more data consistency checking and introducing stronger data quality processes. I introduced the methods championed by Edwards Deming and Joseph Jurand to try to incorporate quality improvement as a systematic component of SOI data collection processes.[7] The net effect of the use of more and more modern statistical quality control methods was to mitigate what could have, otherwise, been a serious lowering of data quality.

Enumerative versus analytic goals

In the beginning of the SOI program just tabular statistics were available. At first, that was all anybody wanted or could use. Of course, more timely and more detailed statistics were always in demand. At their best SOI publications usually came out several years af-

[5] The tabulating machine or Hollerith Machine was an electromechanical machine designed to assist in summarizing information stored on punched cards. It was the forerunner of the computer. See www.census.gov/history/innovations/technology/thehollerithmachine.

[6] The Universal Automatic Computer,or UNIVAC, was an electronic digital computer that used vacuum tubes and state-of-the art circuits to tabulate data and was a significant leap forward over the electromechanical tabulating machines. See www.census.gov/history/innovations/technology/univac1.

[7] SOI was already familiar with Deming's work through a report commissioned in the 1960's evaluating SOI sample processes: Deming, Edwards, (1963) "Review of the Sampling Procedures Used by the Internal Revenue Service to Produce Statistics of Income from Individual Tax Returns with Special Emphasis on Achievement of Quality." Reprinted in *Turning Administrative Systems into Information Systems: 1994*, Wendy Alvey and Beth Kilss eds, Washington, DC: Internal Revenue Service.

ter the reference tax year. Corporate tax return statistics, in part for structural reasons, chronically lagged individual tax return statistics. In good years, the individual statistics took two years to publish, corporation statistics three. There were, of course, points in SOI's history where other priorities (the Depression in the 1930's and then World War II) meant that statistical releases were even further delayed.

The dawn of the 1960s brought changes to America. Wide adoption of computers and advances in computer languages led to improvements in the IRS Master File and the clients' abilities to use microdata. This led to a demand for SOI data over SOI statistics/analytics. Regular delivery of electronic microdata files to SOI customers in Treasury's Office of Tax Analysis (OTA) began with the 1960 Individual Income Tax Public-Use File. I was to work on the 1962 file and on the team that designed the 1964 public-use file. An early use of the SOI data was in developing the 1964 tax microsimulation model used by OTA to study and then introduce graduated income tax withholding.

Increased use of tax data led to requests to link the data to information from other sources in order to improve analytic power of the data. Joe Pechman from Brookings successfully urged former SOI Director Enguist to provide Brookings with the SOI Individual tax return data, so Brookings could match the SOI data statistically to the 1960 Census Public-Use File, which was composed of a 1 in 1000 sample.[8] This led Director Enguist to institute a statistical matching program that included a record linkage effort involving data from IRS, SSA and the Census Bureau's Current Population Survey (CPS). The privacy protection and bureaucratic challenges of this ambitious project were greater than envisioned. But eventually the work was completed by a staff led by me (then at SSA) and SOI's Peter Sailer.[9] Unfortunately, these early SOI efforts did not become routine and the methods used remained *ad hoc* for far too long.

Role of Big Data and analytic data sets going forward

I would argue that the public release of the individual public-use microdata, as well as the creation of new data through statistical matching of the SOI data to files produced by other agencies, marked the beginning of the analytic support role that SOI needs to emphasize going forward. The micro-data sample files produced for Treasury and for the Congress, along with the public-use files used by SOIs' external customers including other government agencies, think tanks and major universities, remain the cornerstone of most tax policy analysis in the U.S. Making public-use microdata directly available to SOI's customers was a good decision. After all, allowing an open data-driven discussion of tax policy issues is central to our democracy, and the need for transparency is very important, then as much as now. While the practice of making anonymous, privacy-protected micro-data available has continued, the disclosure prevention methods required for its release have grown progressively more stringent in response to wider and growing availability of electronically linkable data elsewhere and cheaper computing technology. These increased application of disclosure prevention methods, while absolutely required to protect taxpayer privacy, somewhat reduce the utility of the files. Inevitably, we will reach a point where the data loss required to protect privacy will make the files unsuitable for most uses, requiring SOI to explore alternatives, for example synthetic data or restricted access protocols.[10]

Most statistical agencies are coping with the 'Big Data' revolution, which offers much opportunity for SOI to learn from and adopt and/or adapt what others are researching or doing already.[11] For SOI, relatively easy access to the growing amount of electronic data captured for large segments of key filing populations means the days of SOI enumerative samples are numbered and I believe that number is small.[12] These samples will be replaced by smaller, analytic samples in-

[8] For a discussion of this work, see Pechman, Joseph (1985) *Who Paid Taxes 1966–1985*, Washington, DC: Brookings Institution.

[9] See for example: Crabbe, Patricia; Sailer Peter, and Kilss, Beth (1984) "Taxpayer Data Used to Study Wage Patterns by Sex and Occupation: 1969, 1974 and 1979," In *Statistics of Income and Related Administrative Record Research: 1984*; Wendy Alvey and Beth Kilss eds, Department of Treasury: Internal Revenue Service.

[10] For a discussion of privacy concerns in public-use data, see: *Privacy, Big Data, and the Public Good: Frameworks for Engagement* (2014) Julia Lane, Victoria Stodden, Stefan Bender, and Helen Nessenbaum eds, Cambridge University Press.

[11] For example, National Academies of Sciences, Engineering, and Medicine (2017) *Innovations in Federal Statistics Using New Data Sources While Respecting Privacy*, Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods, Robert M. Groves and Brian. A. Harris-Kojetin, eds. Committee on National Statistics, Washington, DC: The national Academies Press. Also: Feyen, Michelle (2015) Transforming How We Produce Statistics: An Inside Perspective, *Statistical Journal of the IAOS*, Vol 31, pages 59–66.

[12] Kitchin, Rob (2014) *The Data Revolution: Big Data, Open Data, Data Infrastructures & Their Consequences*. Sage.

tended to validate and interpret the aggregates that will be produced using the administrative data.

Improved access to comprehensive administrative tax data offer expanded opportunities to build on the experience of those pioneering statistical matching studies in ways that can benefit the entire federal statistical system. One way in which tax data have traditionally been helpful has been through the creation of sampling frames to support work at the Census Bureau and National Agricultural Statistical Service and even the Federal Reserve Board's Survey of Consumer Finances.[13] Statistical agencies that have historically relied on surveys to collect information are now facing formidable challenges as response rates decline and data collection costs increase, providing an opportunity for SOI and tax data to make even greater contributions.

One long-recognized need is for expanded use of tax data provided by businesses to improve consistency among the business registers used by the Census, the Bureau of Economic Analysis (BEA), and Bureau of Labor Statistics (BLS). Current legal limits prohibit direct sharing of tax data with BLS, however the BLS data are foundational to the developing of BEA's National Income and Product Accounts (NIPA); NIPA also incorporate tax data from the IRS and Census. Finding a way to help eliminate structural differences in the registers used by these agencies, perhaps through special research projects or new joint SOI/Census products that would also benefit the SOI mission, would provide a great service to the larger statistical community.[14] There are many other examples, some, like work with the Census for improving the decennial Census and the American Community Survey, are already under discussion or being piloted. More collaboration among the agencies is essential to the long-term health of the federal statistical system.

There is also a role for SOI in filling the metadata/paradata gaps for data collected during administrative processing. After all, the metadata produced for operational purposes makes the data quality fit for that purpose, but not necessarily for an analytic or policy research application. There appear, however, to be some real fitness (quality) gaps when the data are used for analytic policy applications. SOI's place shifts in this scenario from primarily data delivery to a metadata and paradata service role. The samples that are selected by SOI are, then, used to strengthen the data's interpretation and use. In this new formulation, SOI would, thus, be the "dirty hands" partner, close to the raw data that comes from tax and information return filers, fixing the raw data's weaknesses when they might endanger a central policy use, and, at minimum, developing an understanding of the data's weaknesses so SOI's clients can employ the data anyway, when fixing them is not affordable. Sometimes fixing just parts of the raw data is all that can be afforded.

Here the approach might resemble metaphorically fording a stream that you cannot afford to bridge, building pile after pile of boulders spaced so that one can jump from one pile to another in the hope that one can eventually get to the other side without getting wet or at least without drowning. This "roughly right" world, as the metaphor implies, requires a lot of statistical literacy skills by everyone involved, not just SOI staff, but also all the various user and producer communities of which SOI is a part.

Closing thoughts

In my view, SOIs future lies in partnerships with its customer and fellow data suppliers, linkages across disparate data systems and building differing units of analysis, in the cross-section and longitudinally. Some of these and other possibilities for SOI's future role are developed and demonstrated elsewhere in this publication in selected papers and in comments from Arthur Kennickell. But nearly all of us are likely to underestimate the distance to the goal and may not even see the rivers (or oceans?) that have to be crossed. There is usually the fog of detail to be seen through or stumbled into and out of without taking a bad fall, losing one's way, or drowning. However, being guided by a sense of adventure and not a map, this need not be too daunting, even for a 100-year old organization, especially one that already has a lot experience evolving. Anyway, what are the alternatives?

[13] Statistical uses of tax data by Census and NASS are outlined in Internal Revenue Code §6103(j). For detailed descriptions of the evolution of the use of tax data in support of the Survey of Consumer Finances, see (2017) *Statistical Journal of the International Association for Official Statistics, Constant Focus: Engaging to Measure Wealth*, Volume 33.

[14] For a detailed description of this issue, see National Research Council, (2006), *Improving Business Statistics Through Interagency Data Sharing: Summary of a Workshop*, Caryn Kuebler and Christopher Mackie, Rapporteurs. Steering Committee for the Workshop on the Benefits of Interagency Business Data Sharing, Committee on National Statistics, Division of Behavioral and Social Sciences and Education, Washington, DC: The National Academies Press.