SOI at 100:

Traditionally Enumerative, Now More and More Analytic?

By Fritz Scheuren, retired SOI Director

0. A Personal Preface

For those of you who have not kept up with me (and why should you?), I am currently the Chief Editor of the Statistical Journal [1] of the International Association for Official Statistics (IAOS). That is part of the reason my SOI remarks and handouts have more of an international perspective than might have been expected. Forgive me please for the very personal nature of much of what follows.

Technically, the SOI legislation was not passed until October, 1916 [2]. This factoid, like many of the others provided in this paper, may not be new to you. Still the retelling of SOI history is personally and professionally satisfying and so I will go ahead in any case. My father, who was a great storyteller, used to begin a story that he had told many times. At some point, usually well into the tale, he would realize he had told this very story to these very same people before.

He would pause at points like that. Then, with a twinkle in his eye and a smile on his face, he would take a breath and continue anyway. So it is with me. I am my father's son, even to the twinkle. So forgive me please for this retelling and for places where my memory was garbled or had gaps.

Let's start then with some Pre-SOI history? There had been an income tax during the American Civil War [3] in the 1860s; but eventually it was considered unconstitutional [4], only to be re-instituted in 1912, as one of a series of progressive constitutional amendments [5]. Among other amendments were the shift in the way Senators were chosen (by direct election rather than by state legislatures) and the later repealed amendment that created Prohibition [6]. Enough on the legislation? Let me now turn to the technology available in 1916?

Hermann Hollerith, at the Census Bureau, began the Federal government's use of mechanical card tabulating machinery or "Tab" equipment for the 1890 Census [7]. IRS, likely, used this equipment from the inception of the SOI Program.

I still remember seeing this equipment in the early 1960s when I began at SOI. There certainly was a strong continuing cooperative relationship between SOI and the US Census Bureau when I arrived. IRS, after all, had purchased a one-third interest in the second Univac computer that Census bought and SOI used it on the night shift for years.

Lil Dorsey, one of many SOI heroes I was to know, some of whom like Pete Sailer are in this room today, told stories about the SOI/Census Univac computer in the days before transistors. Then, the Univac was very hot, loud, and always seemingly breaking down. Still, the Univac was a great advance. Lil told a story about how Census/IRS programmers wrote nonsense code during Christmas time. These programs turned all that noise into Christmas Carols. Only wished I had heard those

carols myself. One is always a creature of the age we live in and I was too young then, to have lived in that past, just like I am too old now for the future coming. Anyway, I will speculate about both some today.

At the beginning, the SOI operation was almost entirely clerical, and the workforce was almost all white and female. Usually as typically would be expected in those days, the men were the managers, although some women were what might, in another setting, be called “shop bosses.” Women seldom rose to senior management positions, though.

One exception was Helen Demond, who had just retired as an SOI Branch Chief when I came to SOI. Helen rightly deserved a shot at the SOI Director position. But was told, when she asked to be considered, that the position was not open, except to men.

Since those days, as we all know, there has been a general lowering in both gender and race barriers towards equal treatment. SOI was already out front on race, but not gender, when I got there.

Things are better now on both fronts. But, like the country, SOI in my opinion, has a long way to go. Still I am proud to have been in SOI for 17+ of my 31+ years of Federal service.

1. Introduction and Organizational Background

What eventually became the Statistics of Income (SOI) program began legislatively as an addition (Section 6108) to the Internal Revenue Code (Title 26) in 1916. The name of the program came out of SOI's statement of legislative purpose.

The original focus was primarily on picking up income details from the individual and business returns filed annually. Anyway, the SOI Program, soon to be in its 100th year, was begun on that basis. This paper looks back at that period and sketches where SOI is now and where, looking forward, SOI seems to be heading.

This paper will cover mainly the past. The future of SOI quite obviously has yet to be written. But I will speculate nonetheless. Maybe some of you reading this will make that history and then tell that story, when I am long done. Anyway, to begin?

2. Early SOI Days

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. Everything was done manually in those early days. This, after all, was well before the advent of electronic computers.

We do not have many details, so some of what comes next is speculative. I only wish back in the 1960s, when I began my career at SOI (now 50+ year ago) that I had

asked more questions of the people then working there, instead of guessing about so much, as I have to now.

Before the coinage of the words metadata and paradata were minted, "***what to do***" procedural manuals were being written. I wrote many such manuals myself.

The "***why do it this way***" documentation was seldom created, though; perhaps not even now. The archival storage of these procedures was very shaky too, so I do not have much to rely on, except my personal files.

Maybe the belief in stability overcame the normal caution in bureaucratic organizations to document what was done and why. Anyway, to say the least, there was to be far more change than expected!

3. SOI Enumerative Samples

The SOI individual samples were systematic and stratified, but for all intents and purposes they were treated as statistically stratified random samples [8]. The other return types, when studied, were not originally sampled at all; but based on the full population of the returns filed. Their populations were a lot smaller and most such returns, even back then, a lot more complicated, hence too hard to sample cost-effectively. As those of you who do sampling well know, there is a crossover point where the costs of the sample design work can exceed the cost of just processing the

entire population. I suspect this was a factor that slowed the sampling of these small populations.

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 abstracted onto "edit sheets," so named because the material on the return was examined for consistency before being entered on an edit sheet. 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.

The edit sheets would then be keyed onto 80 column punch cards, which were later sorted by the stratifiers and the required tabulations made. Typically, there was further checking using what were called consistency tests.

Individual Income Tax Returns were manually sampled, using a stratified design based on District Office and size of total income, later adjusted gross income. 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. Only the individual return sample selections were shipped to Washington for further processing, not the whole population.

Other returns, (e.g., business, estate and gift tax returns) were not generally sampled until the 1950s; and, hence, all were shipped to Washington for edit/abstraction. Remember there were no service centers in those days.

All the Corporation returns were statistically processed each year in the IRS National Office, as was the sample of the Individual returns. Traditionally, the other tax returns were neither sampled nor processed annually.

4. Initial SOI Work Structure

In the beginning of the modern tax system, IRS was highly decentralized and conducted by Internal Revenue Agents around the country [10]. In fact, until the 1950s the IRS approach appears to have been structured in a manner similar to the way Indian Agents were monitored during the 19th and early 20th Century [11]. Only a few summary items were consolidated in Washington and provided to the public in a Commissioner's Annual Report (i.e., number of returns, total income, and tax). For most of the nearly 100 years of the SOI program, as is true now, the Commissioner's Annual Report was a responsibility of the Division, at that time called the Statistics Division.

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 only big difference is that for resource and timing reasons a sample was selected instead. The

enumerative sample goal [e.g., 11a] is to describe a population, not as might be the case in an analytic sample [e.g., 11b] to use the sample to look at how the population arose.

Books full of tables were the main products produced initially. For example, weighted tallies of items from individual returns, with subtotals by district office (or state) were produced early, along with some summary measures of economic activity. Examples of economic activity were total or adjusted gross income; and, for corporations, total assets, gross receipts, and net income/loss.

Special SOI studies were undertaken at Treasury's request with greater and greater frequency. Such special studies continue today, albeit, reduced for budget reasons and/or found unnecessary because of changes in the tax system.

5. **SOI in the Depression and until 1960**

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 regular SOI Corporate program, until today.

Business returns were stratified by industry and size initially for tabulation purposes and later for sampling purposes as well. Corporate returns were first sampled in 1951 but with great difficulty. In fact, the uncertainly introduced by the sampling of corporate returns endangered the time series by industry and size of total assets.

As the story goes, Ernie Enguist, then, SOI Director pitched in himself week after week, week-ends too, and worked to smooth out the transitions. He was an economist who had come from the Economic or Business Statistics side of the Census Bureau. And he knew his stuff.

He may have been responsible for the decision to begin sampling the corporate returns, although the number of corporation returns had grown greatly and something had to be done. Having made the decision he helped fix the unintended consequences.

I was to work for Ernie. Another SOI hero. Perhaps the greatest SOI Director. Sorry Tom [Petska], if it is any consolation to you I do not consider myself in his league either. Of course, Barry is stepping up nicely and may be the best of us all!

6. Moving away from Tradition

Regular delivery of electronic microdata files to SOI customers in the Office of Tax Analysis began with the 1960 Individual Income public-use file. The file went to Brookings to be statistically matched for use with the 1960 Census public-use 1/1000 sample developed by Jack Beresford [12] at Census.

Joe Pechman from Brookings successfully urged the then, Statistics Director Ernest Enguist to provide Brookings with SOI Individual tax return data, so Brookings could match the SOI data statistically to the 1960 Census Public use file [13].

The statistical matching program was started under Enquist. He also knew Joe Steinberg from his Census days and supported a record linkage effort that Joe started involving data from IRS, SSA and the Census Bureau's Current Population Survey (CPS). The disclosure and bureaucratic challenges of this ambitious project were greater than envisioned. But eventually the work was completed by a staff led by me and including Pete Sailer. At that point I was at SSA [14]. I had not worked on that effort at IRS

I was to work on the 1962 file and on the team that designed the 1964 public-use. I well remember flying punch cards up to the Detroit Data Center for model tabulation/simulation runs that the Office of Tax Analysis at Treasury wanted, working onsite at the IRS Detroit Data Center to tune the simulation model runs, and, then, returning with bound paper tabulations to "walk up the street" to the Office of Tax Analysis.

The 1964 tax microsimulation model that SOI had developed went live with the effort by Treasury to study and then introduce graduated withholding. Although methods were to improve later, my recollection is that I personally may have gone back and forth perhaps a dozen or more times to Detroit on this one project.

We did not have virtual VPN technologies then; and, hence, the Service's severe disclosure restrictions forced us to employ difficult workarounds. Of course, to be

honest I was young and had a ball. Doing something new, important, and interesting! What more could a guy still in his 20s ask for?

5. Where are we now?

The micro-data public-use files continue today and are now used by SOIs' customers directly, as the Individual public-use file is available fairly widely outside of IRS and main Treasury. These files are made available to policy analysts not only at Treasury but also to Congressional staffs and executive government agencies and major Universities.

The SOI practice of making disclosure-proofed micro-data available has continued, albeit the disclosure prevention methods that are required for its release have grown progressively more stringent, with the wider and wider availability of electronically linkable data elsewhere and cheaper and cheaper computing [14].

Making restricted-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. Making available the Individual public-use file, in a scrubbed version with a low re-identification risk remains. The public release of the individual public-use microdata, we would argue, was a big advance in moving SOI from an enumerative to a more nearly fully analysis program.

6. 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. At their best SOI publications usually came out several years after the reference tax year. There were, of course, points in the history of SOI where other priorities (the Depression in the 30s and then WWII) took the focus off SOI altogether.

At those times, statistical releases could be 8 or more years after the reference tax year. Corporate tax return statistics, in part for structural reasons chronically lagged individual tax return statistics [15]. In good years the individual statistics took two years to publish, Corporation statistics three.

The quality of the SOI tabulations was maintained by thorough training and then later modern quality assurance (QA/QC) sampling procedures. The formal quality SOI program seems to have been stepped up a notch with the growing decentralization of the statistical data capture operation. Deming was brought in to confirm this and he did [16].

Bottom line, at the era when SOI operations were decentralized the problem that arose of differences across processing sites, with different implementations, was addressed with stronger data quality procedures and more consistency checking.

The net effect of the use of more and more modern statistical quality control methods was to mitigate but likely not completely reverse what could have, otherwise, been a serious lowering of data quality. If memory serves the centralization of the complex corporate abstraction was begun in the early 1960s, moving it from Washington to various IRS Service Centers and, when it was opened, the Detroit Data Center.

This change was a good thing, unlike the decision to separate the SOI computer program staff and move a big part of it to Detroit in the 60s. That staff move certainly slowed down the ability of SOI to cope then with the increasingly changing IT world. That IRS upper manager decision, perhaps partly political, held up the SOI program for perhaps a decade or even two. A major blow. Narrow cost savings, if any, purchased at the price of major benefit losses to SOI customers.

There were further opportunity costs, however, to the move of the SOI computer staff to Detroit. Certainly that move was to slow the analytic efforts that the SOI program needs to satisfy its customers in the future. To summarize, so much SOI energy had to be spent on these imposed changes that modernization was greatly impacted. Additionally efforts to improve the main SOI data sets over time or over institutions via linkages were, therefore, slower than desired too.

As noted, the Individual public-use file was developed in part to be statistically matched to the then new 1/1000 sample from the 1960 decennial census. Unfortunately, these SOI efforts did not become regular and the methods used remained ***ad hoc*** for far too long [17]. These efforts were, however, the beginning of the analytic support role that SOI needs to emphasize going forward.

7. Role of Enumerative v. Analytic Data Sets Going Forward

With the growing amount of routinely 100% captured (CDW) revenue processing data, the days of SOI enumerative samples are numbered and that number is small. SOI samples, analytic but no longer enumerative are part of what is wanted. These samples would be much smaller than now. Their goal would be to validate and interpret the aggregates that can now be produced by the revenue processing system.

There is a role for SOI in filling the metadata/paradata gaps for data that comes out of Revenue Processing. After all, the metadata done for operational purposes makes the data quality fit for that purpose, not necessarily for an analytic or policy research application.

Revenue processing is of high quality when the data are for the fitness of that tax administration use. 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, as noted 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; at a minimum developing an understanding of the data's weaknesses so SOI's clients can employ the data anyway, when fixing them is not affordable.

There are lots of microsimulation models now in government policy shops, like the Office of Tax Analysis (OTA), that are actioned by data inputs. SOI needs to examine these applications outside Treasury for examples. The TRIM processing done for Health and Human Services would be one example [18] when a full fix is wanted. 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, takes a lot of statistical literacy skills, by everyone involved not just SOI staff but also all the various user and

producer communities that SOI is part of. We are all still learning this by doing it together.

Guided by a sense of adventure and not a map need not be all bad, even for a bureaucrat. We are all still young enough to enjoy this! Right? Anyway, what are our alternatives?

8. An Afterword?

In my view, SOIs future lies in partnerships with its customer and fellow data suppliers, linkages across disparate data systems, building differing units of analysis, in the cross-section and longitudinally.

Other possibilities for SOI's future role might be mentioned, even though these are complex and emotionally charged. For example, there are "turf" issues that could impede "satisficing" relationships for SOI. To be blunt, entities (individuals and organizations) outside IRS see very clearly the goal --maybe better than SOI does.

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 thru/stumbled into and out of without taking a bad fall, losing one's way, or drowning.

Other official statistical organizations, not just SOI, have been coping with the 'Big Data' revolution [e.g., 19]. This, too, offers much for SOI to learn from and adopt

and/or adapt what others are doing already or researching. But that is for another paper to be talked about and speculated on when we meet in June.

Cited References (to be provided separately)