Distr.
GENERAL

WP.13
25 April 2012

ENGLISH ONLY

UNITED NATIONS ECONOMIC COMMISSION FOR EUROPE (UNECE)
CONFERENCE OF EUROPEAN STATISTICIANS

EUROPEAN COMMISSION
STATISTICAL OFFICE OF THE EUROPEAN UNION (EUROSTAT)

ORGANISATION FOR ECONOMIC COOPERATION AND DEVELOPMENT (OECD)
STATISTICS DIRECTORATE

Meeting on the Management of Statistical Information Systems (MSIS 2012)
(Washington, DC, 21-23 May 2012)

Topic (ii): Streamlining statistical production

Leveraging Processing and Computing Improvements to Streamline Data Collection at the U.S. Statistics of Income Division of the IRS

Supporting Paper

Prepared by Barry Johnson and Melissa Ludlum, Internal Revenue Service, Statistics of Income Division, United States

I. Introduction

1. Statistical data derived from U.S. tax return information are used for a wide range of activities, including economic analysis, revenue estimation, and tax policy studies. The Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) has produced official statistics based on information reported by U.S. taxpayers for almost 100 years. Increasingly, taxpayers provide data to the IRS electronically, rather than on paper returns. In recent years, computing capabilities have also improved, with processing speeds accelerating and storage space becoming less costly. This paper will describe the ways in which SOI has leveraged these advances to improve its data collection processes. These improvements, which include fully or partially automated data collection procedures, more sophisticated quality verification, and expanded sample and item coverage, have allowed SOI to improve data collection efficiency, enhance data quality, and provide faster, more multifaceted products to its customers, while reducing overall costs.

II. Background of the Statistics of Income Division

A. SOI Mission and Products

2. SOI is the authoritative source of information on the federal tax system. Founded in 1916, coincident with the enactment of the U.S. income tax, SOI is charged with making data available annually on the function

of the tax system.[1] In fulfillment of its mission, SOI produces microdata files, summary tabulations, and written analyses.

3. SOI microdata are used by the U.S. Congress and Treasury Department for budget and tax administration. By statute, SOI also provides select data to other agencies; these data are used primarily in support of national statistics functions, including production of the U.S. Gross Domestic Product accounts.[2] SOI also produces a limited number of public-use microdata files, which are used widely by researchers outside the federal government for economic and tax policy related research.

4. SOI makes reports, tabulations, and some of its public-use data available through its Web pages, www.IRS.gov/Taxstats. There are currently more than 14,000 files available, including a complete digital archive of all its historical publications. These Web pages average between 6 and 7 million direct downloads annually. Users include academic researchers, the media, students, and the general public.

B. Studies and Data Collection Operations

5. SOI collects data from a broad range of administrative tax records filed with IRS. These records include tax returns filed by individuals, trusts, and domestic and international businesses, as well as information returns filed by charities and other types of tax-exempt organizations. At any given time, SOI may have more than 100 separate data collection projects, or studies, in various stages of development, production, or post-production processing. Most of these projects have, at their core, a stratified statistical sample of tax or information returns, selected on a flow basis after the IRS has processed them for tax administration purposes. All projects augment data items captured by the IRS during administrative processing with additional data items reported on supporting schedules and attachments. Data are also collected from information documents provided by financial institutions and employers.

6. SOI uses its own integrated computer network to capture and complete statistical editing of returns selected for its samples. Oracle provides the foundation of these data collection systems and the underlying databases. SOI developers produce customized data entry systems for each study, which are used to collect and test information provided on tax and information returns. Modular program designs allow applications to share standard components, yet provide the flexibility needed to ensure that each application includes specialized tools tailored to the specific nuances of the relevant tax code.

7. SOI field personnel, known as "editors," are located at 5 IRS processing centers geographically distributed throughout the U.S. Editors are specially trained to perform a wide variety of data collection and statistical editing functions. In addition to transcribing data items from all returns selected for the SOI samples, along with additional data from supplementary schedules and supporting documents, SOI editors perform a number of other tasks to improve the usefulness of the data for statistical analysis. For example, they assign analytically useful codes, such as occupation and industry codes. An important role is re-coding or re-classifying financial data reported in generic categories, for example “other assets,” to more specific categories, such as "stocks." Most significantly, they ensure that data for all projects are thoroughly tested for both internal consistency, verifying that all mathematical relationships are valid, and for agreement with provisions of the tax code. Working in consultation with SOI economists, who serve as subject-matter experts, editors correct any errors or inconsistencies that are found in order to ensure that the final data are of the highest possible quality.

III. Electronic Filing

8. Traditionally, SOI programs worked with original paper-filed documents supplied by taxpayers, which were shipped to SOI editors after administrative processing, accruing significant shipping and

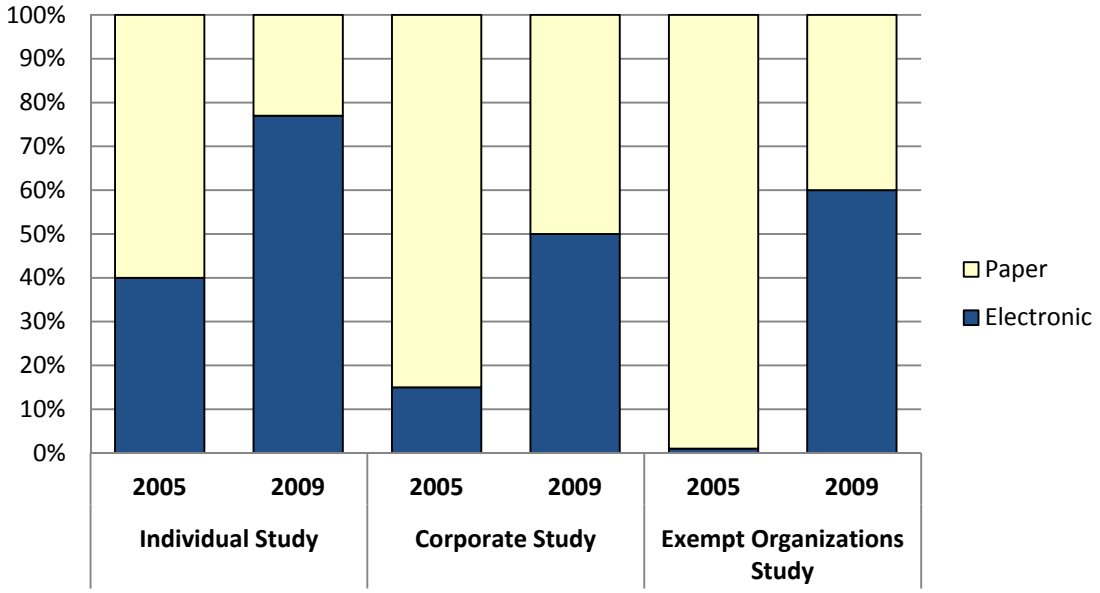
[1] See Title 26, Internal Revenue Code of 1954, Section 6108.
[2] Ibid., Section 6103.

controlling costs. Beginning in 2004, to save resources and speed processing, SOI transitioned to the use of digital images, made from paper documents, as the primary information source.

9. In 1996, the IRS introduced electronic filing, allowing some individual taxpayers to submit tax information directly to the IRS in electronic format, foregoing paper altogether. Today, more than 77 percent of individuals file electronically. More recently, the IRS Modernized Electronic Filing (MeF) program expanded electronic filing to corporations and tax-exempt organizations.

10. Electronically filed returns have become an integral part of SOI's statistical samples, replacing digital images as the primary information source for many studies. As shown in Figure 1, the percentage of returns in electronic format sampled for the individual, corporate, and exempt organizations studies has increased substantially in recent years. Electronically filed returns represented 40 percent of returns selected for the Tax Year 2005 individual sample and just 15 and 1 percent of the corporate and exempt organizations samples, respectively. This changed dramatically over the ensuing 5-year period. By Tax Year 2009, electronically filed returns comprised nearly 80 percent of returns included in the individual sample and at least half of the corporate and exempt organizations samples.

Figure 1: Percentages of Sampled Returns Filed by Paper and Electronically, Tax Years 2005 and 2009



11. The increased presence of electronically filed data has afforded SOI with opportunities to introduce new data collection procedures that have streamlined the statistical data collection process. In addition, overall improvements in computer technology and decreased costs of machine storage have also had a tremendous impact on the SOI programs. The sections that follow will discuss some of the ways that SOI is leveraging these changes to streamline the data collection process, invest in data quality initiatives, and expand the scope and expedite delivery of its products.

IV. Automating the Data Collection and Statistical Editing Process

12. Before the introduction of electronic filing, SOI editors manually keyed taxpayer-reported data from paper returns (or digital images) into Oracle data collection forms as part of the data collection process. Automated consistency tests, embedded in the Oracle programs, prompted editors to complete statistical editing tasks and resolve inconsistencies during the data entry process. Using information obtained from electronically filed returns, SOI has nearly eliminated transcription for many of its programs and streamlined

many other edit functions. An obvious benefit of collecting data in this manner is the overall reduction in the amount of time required to perform data entry and editing.

13. A number of SOI studies now use fully automated batch processing for electronically filed returns. These programs automate both the data collection and statistical editing functions, greatly reducing the need for editor involvement. For the individual income tax study, almost 36 percent of the sample is edited using fully automated processes, including some error correction routines. SOI editors continue to correct errors manually in those cases where error correction cannot be automated due to the complexity of the returns. Likewise, for several relatively small SOI studies of international corporations and associated individuals, data testing and most error correction have been fully automated for the entire SOI sample, meaning that no editor involvement is required to collect data from or edit the returns. For the more complex corporate foreign tax credit study, editor involvement has been eliminated for the 78 percent of the sample that is electronically filed, and nearly 22 percent of all errors were auto-corrected during the most recently completed study; SOI economists resolved the remaining error conditions.

14. Unlike commercial tax preparation software available for individuals, software developed for U.S. businesses and tax-exempt organizations provides taxpayers with relatively little automated preparation assistance, which increases the potential for taxpayer errors and incomplete filing information. Thus, SOI's studies of U.S. corporations and tax-exempt organizations are hybrids of automated data collection processes and traditional statistical editing. For these studies, SOI extracts statistical data from the XML code provided by filers. Once extracted, these data are written directly to databases and used to pre-populate Oracle editing forms. Manual intervention is limited to complex editing functions and error correction, although data transcription is still required for paper-filed returns.

15. Advances in technology, including faster processors and reduced storage costs, have allowed SOI to introduce numerous tools to increase efficiency and quality. For example, a newly developed automated coding routine has been introduced on SOI’s Sales of Capital Assets (SOCA) project, a study that collects information on sales of investment assets, such as stocks, bonds, and mutual funds. Editors classify each asset into one of 22 different categories and assign an asset type code, a process that has been very costly. Beginning in 2012, a significant portion of taxpayers is now providing this information to the IRS electronically, allowing SOI to automate this coding process. The process takes advantage of an SOI-created database, made up of over 1 million asset descriptions and asset codes developed from earlier SOI studies. A matrix of PL/SQL routines uses this master database to assign a preliminary asset code to each reported transaction. Editors then review the assigned code and make corrections as needed. The new coding system accurately codes 75 percent of assets included in the SOCA project and has reduced costs by 34.7 percent per transaction.

16. SOI has also developed automated tools to improve existing controls and monitoring for data collection and statistical editing processes. SOI programmers have created “dashboards,” which display real-time metrics for return processing, such as remaining workload, program completion progress, and a variety of accuracy and quality measures. The dashboards can be accessed at any time by managers and editors. These tools provide precise tracking of editing progress and allow managers to make informed decisions to meet production milestones. For example, the dashboards display the actual number of returns that must be completed daily by the editors to meet all target program completion dates. This information is compiled based on actual production statistics and is updated dynamically. Editors and managers also use dashboard information to monitor the quality of work so that interventions, such as additional training or improved guidelines, can be introduced as soon as problems arise. In addition to allowing more efficient workload planning and quality monitoring, the study dashboards have replaced more cumbersome ad hoc reports.

V. New Quality Enhancing Initiatives

17. The radical reduction in data transcription due to the availability of electronically filed data has allowed SOI editors to focus more intensely on data quality and value-added activities. For all studies, returns without a tax liability and those filed only for regulatory purposes frequently require the most editing. Here, improved technology has allowed SOI to leverage historical data to verify or correct ambiguous, missing, or inconsistent information, saving time and improving final products. For some studies, including those of tax-exempt organizations and U.S.-owned foreign corporations, editors can use the Oracle data collection programs to access data from a previous reporting period in order to perfect current-year data. In addition to using these prior-year data to validate taxpayer provided information, they can be used to impute missing values. This use of longitudinal editing has reduced the time spent researching or contacting taxpayers, saving resources, and reducing respondent burden.

18. In addition to longitudinal editing, improved technology enables editors to access related tax and information returns to improve data quality. SOI integrated both longitudinal and related information return data on a recently developed study of a form used by tax-exempt bond issuers to claim a credit payment from the federal government. For this study, the SOI edit system displays the form used to request the credit payment, as well as the information return filed at the time the original bond was issued. It also provides editors with access to previously edited credit payment requests filed for the same bond and to administrative information collected after the current credit payment return was filed. Editors use this additional information to verify and correct data reported by the filer, and supplement data on the credit payment form with additional information from the original filing, which increases the analytical value of the final data file.

19. At the close of statistical editing, SOI economists historically have conducted post-edit reviews to verify the statistical consistency of the data files. Although each return in the data file has already been carefully checked, these post-edit reviews examine the entire sample together in order to identify outliers, missing records, and inconsistent intertemporal changes at both the aggregate and microdata levels. In the past, these reviews were performed manually at the close of the statistical editing process. Recently, SOI has automated much of the post-consistency test process, introducing batch programs to identify complex data quality and statistical consistency issues. "Statistical assistants" who work onsite with editors to monitor production and quality, periodically run these batch reports and make the appropriate data edits. Only the most complex cases are referred to economists for additional intervention. The batch programs allow post consistency testing and data correction to be completed more quickly and uniformly. Periodically assessing the final data during statistical data processing also allows timely feedback and guidance to editors. This can improve the overall quality of the edited statistical data and allow SOI to deliver final data files earlier.

VI. Product Improvements

20. Decreased costs of statistical editing processes and the increased availability of electronically filed data have allowed SOI to deliver improved products to its customers. Many SOI programs now collect significantly more data items than in the past. Figure 2 compares the overall number of data fields collected for the Tax Year 2009 studies to those collected for Tax Year 2005. Over those 5 years, SOI expanded the individual program to include more than 400 additional data fields collected from 15 new forms and schedules. Similarly, the corporate program added more than 400 new fields, while the exempt organizations programs more than doubled to include 600 new data items.

21. Cost savings realized from incorporating data from electronically filed returns have allowed for expanded sample coverage for many studies, improving the coverage of sub-populations or smaller geographic regions. For example, compared to 2005, the Tax Year 2010 individual sample includes approximately 100,000 additional returns, while supporting data collection for two concurrent panels and the annual cross-section sample. The Tax Year 2009 exempt organizations sample includes nearly 1,000

additional returns filed by tax-exempt hospitals. Likewise, the private foundations sample has expanded to include all organizations that report certain excise taxes related to their activities. The increased sample sizes for these studies have allowed SOI to make more detailed information available to customers while improving overall reliability of the estimates.

Figure 2: Data Fields Collected for SOI Studies, Tax Years 2005 and 2009

Program	Number of fields		Percentage change
	2005	2009	
Individual Studies	2,266	2,679	18.2
Corporate Studies	1,960	2,365	20.7
Exempt Organizations Study	400	1,002	150.5

22. Processing efficiencies realized through incorporating electronically filed returns and improved technology have also reduced overall costs and improved the timeliness of SOI products. For example, the SOI corporate study file delivery has been advanced by 2 months. The expedited file delivery has, in turn, saved additional resources by eliminating the need for a preliminary data file, which previously had been created and delivered to major customers in advance of the main data file. Through other technical innovations that leverage electronically filed information, SOI has also augmented the corporate file with additional data reported for international business transactions, making these data available more than 9 months earlier than in the past, which has allowed customers to integrate these data into time-sensitive estimates.

23. Technology advances are also transforming the products that SOI makes available, both to customers who receive microdata files and to the general public. In the past, annual administrative data files for the more than 143.6 million U.S. individual income tax filers were too large for SOI to effectively process for most statistical purposes. In addition, there were significant lags between the dates taxpayers filed their returns and SOI's ability to access administrative data collected from them. In recent years, however, SOI has begun using these data on a regular basis to create new products and to improve existing offerings. Some of these, such as recently produced state-level tables, focus on small-area estimates or estimates for relatively rare subpopulations. They are also being used to improve SOI's very popular migration files, which track population inflows and outflows at the county level. In the past, the migration files were produced for SOI and were limited in coverage to only those taxpayers whose tax return information was received and processed by the IRS during the first 37 weeks of a calendar year. With increased computing capacity and easier, timelier access to data, SOI has successfully piloted creating these products "in house," using a full calendar year (52-week) file, improving the utility of the data. And, because of processing improvements, the data can be made available to the public sooner than in the past.

24. Longitudinal panels can provide insights into the ways individuals or businesses change their economic behavior over time or in reaction to shocks. For years, SOI has constructed very useful prospective panels of individual taxpayers. To create such a panel, SOI would select a sample and then process all subsequent returns filed by the sampled individuals for a pre-defined time period. Producing these panels required lengthy planning and data collection processes, and data were not available for analysis for many years after start-up. High costs also kept the sample sizes relatively small, making it difficult for economists to control for regional or local influences on behavior. Increasingly, SOI is using archived population data to create longitudinal panels. By using these population data, SOI can draw panels retrospectively in order to react quickly to requests for information. Archived population data are also well-suited for constructing considerably larger panels. Recently, researchers created a 15-year longitudinal panel representing the entire population of individual taxpayers. This file has already been used to produce papers examining the long-run effects of tax credits targeted toward low-income taxpayers, focusing on outcomes of both recipients and their children. Researchers are just beginning to tap the potential of these data.

25. In another significant improvement to SOI products, data reported on information documents provided to the IRS by third parties, such as employers and financial institutes, are being more widely incorporated into SOI study files. Each year, the IRS receives and processes nearly 2 billion of these documents, which contain information such as wages, interest earnings, capital gains, and pension contributions. Matching data from these documents to SOI files is now used to verify, as well as augment, data that are reported on sampled tax returns. In some cases, these data, which represent actual transaction values, have replaced imputed estimates, greatly improving the accuracy of data used for economic modeling and analysis. For example, for tax purposes, an individual is required to report as income the value of retirement savings removed from an institution, although, in most instances, these amounts are subsequently reinvested and a tax credit is later claimed. By using information documents, analysts can identify these types of transactions and remove them from income estimates, improving overall accuracy. In addition, research is currently underway to develop methods to use information document data to represent wage earners whose incomes are below the tax-filing threshold in order to provide more complete estimates of U.S. income distributions. If successful, this will provide a valuable new source of information that will support a wide range of economic research.

VII. The Future

26. Continued increases in the availability of electronically filed data and reduced costs of working with very large data files will have an impact on SOI in the future. SOI has already realized significant decreases in data collection costs, and these costs should continue to decline over time. More profound, however, will be the transformation of SOI products – microdata files, tabulations, and analyses. This will include increased production and use of longitudinal panel files, new files that link across different types of economic activity in order to more fully describe economic behavior or sectors, and real-time advanced or preliminary estimates that will provide policy makers and analysts with up-to- the-minute estimates essential for some types of research.

27. Of particular importance will be the increased use of population data to shift SOI products from a focus on particular types of taxpayers to one that describes larger sectors of the economy or tax-filing population. For example, SOI has recently assembled a special longitudinal panel of income data, linked to asset information reported on estate tax returns filed for relatively wealthy decedents. These data will be used for studies of charitable giving and investment return, and they may ultimately refine economists' understanding of lifecycle savings behavior. In the recent past, these wealth data have also been linked to trust income data to better understand the effects of bequests on beneficiaries. In the future, tabulations and analyses based on this type of file will become standard SOI products. Another potential new product, based on a recently piloted project that combined pass-through income information reported for small businesses, sole proprietorships and partnerships to more fully describe the U.S. small business sector, also has great promise. Because these types of businesses sometimes change form in reaction to changes in the regulatory or tax environment, this file could prove invaluable for better understanding and predicting behavioral responses to a wide range of shocks. This work may also be expanded to more fully map complex ownership networks, often used by companies to evade both domestic and foreign tax liabilities. This type of file, and others that that more fully describe business behavior in the face of increased globalization, will likely be important components of future SOI work.

VIII. Conclusion

28. Improvements in technology and increased provision of tax data electronically have had, and will continue to have, a profound effect on the Statistics of Income Division of IRS. Already, SOI has realized cost savings, which have been used to increase content and coverage, as well as improve quality. Data are being provided more timely and are being augmented with population data from related information documents and tax forms to provide analysts with more comprehensive data. These data have the potential to transform the

understanding of behavioral reactions to tax policies and economic shocks. Ultimately these changes have the potential to impact tax, as well as broader government, policies.