USAFacts

Improving the Timeliness, Accessibility, and Usefulness of Tax Data

DATA POLICY REPORT

FEBRUARY 2026

Executive Summary

This USAFacts report makes recommendations to improve the timeliness, accessibility, and usefulness of tax data produced by the IRS's Statistics of Income Division (SOI). SOI provides vital data on American income, benefits, and taxes, all based on the millions of tax documents submitted by individuals and businesses to the IRS every year. However, SOI's data products lag multiple years, are not easy to find or use, and are developed using outdated technology.

To assess these challenges, USAFacts independently evaluated SOI's data products and interviewed a politically diverse group of tax data users and statistical officials. Based on their input and our assessment, we developed a series of recommendations for SOI to produce timelier tax statistics, publish more data, make its data products easier to access and use, prepare tax data for AI, and continue to preserve the privacy of individual tax returns.

Specifically, USAFacts recommends:

Limiting delays and publishing more data while preserving privacy and quality

1. Introduce internal APIs to accelerate data ingestion into the Research, Applied Analytics and Statistics (RAAS) Division's data warehouse
2. Modernize SOI's technology for accessing and analyzing the RAAS's data warehouse
3. Leverage modern technology to accelerate steps to strengthen data quality and preserve the privacy of tax returns
4. Deploy synthetic data files to speed up and improve the Public Use File

Making SOI data easier to access and more useful

5. Create a standalone website for SOI data products
6. Make SOI data products easier to access and download
7. Simplify concepts and provide user-friendly content based on the data
8. Improve the interoperability of SOI data products
9. Make details on the timing of data releases and the amount of historical data available easy to find
10. Make SOI data products AI-ready by integrating metadata, documentation, and evaluations

Addressing SOI's resource and institutional constraints

11 Congressionally authorize funding for SOI and appropriate its funding separately from the rest of the IRS

12 Increase funding for SOI

13 Elevate the SOI Director to IRS's leadership team

The combination of these changes would position SOI to bring its data products into the modern era. This will be particularly important over the next few years as lawmakers and analysts assess the effects of tax reforms introduced by the One Big Beautiful Bill Act (OBBBA).

Introduction

Each year, the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) publishes statistics on a range of tax topics, including tax rates, tax credits and deductions, corporate taxes, income, and more. By collecting over 160 million individual income tax returns each year in addition to corporate tax returns, W-2s for wage and salary workers, 1099s for self-employment income, 990s from nonprofits, and various other tax forms, the IRS sits on one of the richest sources of data in the United States. However, SOI's data products are challenging to navigate and can be delayed multiple years. Additionally, SOI does not publish key data that the IRS collects.

Making tax data timelier and more accessible would enable lawmakers and the public to better understand the taxes Americans pay as well as the effects of changes to the tax code. Now that policymakers have enacted major tax policy changes in the One Big Beautiful Bill Act (OBBBA),[i] reforms to SOI are needed to ensure people receive timely data to assess the effects of the law.

Over the past year, USAFacts identified opportunities to strengthen access to SOI tax data. This involved independently assessing SOI's data products and interviewing a range of tax experts, data users, and statisticians, including tax economists, policy analysts, and IRS data officials. In this report, we provide an overview of our research findings and recommendations to improve access to tax data.

The Importance of Tax Statistics

The IRS collects millions of individual income tax returns, corporate tax returns, health insurance filings, and numerous other tax forms each year, making it home to an extraordinarily rich dataset. The IRS collects detailed information about families, incomes, jobs, employer-provided benefits, government program benefits, corporate earnings, and several other areas of interest among the American public. If used properly, the IRS's data can be a significant resource to inform public policy and many of the pressing issues Americans face today.

Moreover, lawmakers have been increasingly enacting economic and social policy through the tax code. For instance, the employer-provided health insurance premium tax exemption and the Affordable Care Act premium tax credits subsidize the purchase of health insurance. The tax code also contains benefits for low- and middle-income families and children via the Earned Income Tax Credit and the Child Tax Credit, as well as childcare subsidies like the Child and Dependent Care Tax Credit and the 45F employer childcare credit. Additionally, the 45S employer credit for paid family and medical leave subsidizes employer-provided paid family leave benefits. The Inflation Reduction Act (IRA)[ii] introduced a series of tax credits that subsidize the use of clean energy, which OBBBA repealed. The list goes on, as the IRS now administers a major portion of government subsidies, benefits, and spending through our taxes.

As policymakers implement policy through the tax code, it is vital that they asses their outcomes with timely and useful data. IRS tax statistics should be a primary resource for evaluating those outcomes. While tax returns have limited demographic and household structure information, they measure sources of income and taxes paid to the federal government very effectively. Those statistics should also be an important resource for state and local government leaders who often implement their own tax policies on top of federal taxes.

Lastly, the 2025 enactment of the OBBBA heightens the need for timely and useful tax data. OBBBA implemented numerous tax code reforms, including extending, making permanent, or augmenting individual income tax provisions introduced by the Tax Cuts and Jobs Act of 2017 (TCJA),[iii] ending energy tax credits introduced by IRA, and temporarily introducing a series of new tax provisions, such as no tax on overtime and no tax on tips.

About SOI

Following the enactment of the Revenue Act of 1916, officials established SOI in 1917 to publish statistics on income and taxes in the United States based on IRS tax returns.[iv] Located within the IRS's Research, Applied Analytics and Statistics (RAAS) Division, SOI publishes a range of statistical products that inform Americans about the income distribution, the amount of taxes individuals and companies pay, and usage of various tax credits and deductions, among other topics.[v]

These statistics are vital to understanding America's tax system and informing tax policy. For instance, economists often use SOI data products to evaluate the effects of tax reform proposals and forecast future changes in tax revenue. Other analysts use tax statistics to evaluate America's income distribution and the implications of tax policies for individual households. Due to the importance of SOI's data products, the Office of Management and Budget (OMB) designated it as one of the nation's thirteen principal statistical agencies.[vi]

SOI is also a vital resource to other government agencies. SOI provides detailed tax return data directly to the Congressional Joint Committee on Taxation and the Office of Tax Analysis at the Department of the Treasury, which use that data to inform policy decisions in Congress and in the Executive Branch. Additionally, SOI is a crucial source of data for the Bureau of Economic Analysis (BEA), the Census Bureau, the Bureau of Labor Statistics (BLS), and the Federal Reserve, among others.[vii]

The IRS is also subject to strict standards to ensure that the tax information of individual families or companies remains private. As a result, SOI must produce tax return data while ensuring no individually identifiable information is made public.

Our Process

USAFacts analyzed SOI's data products while engaging experts to identify opportunities to improve access to tax data. Our independent assessment focused on the ease of access to SOI's data products, SOI's resources, and its operational processes. This included reviewing SOI's webpages and data products, analyzing SOI's funding and structure within the IRS, and evaluating the AI-readiness of SOI's products.

We also engaged stakeholders about strengthening SOI's data products. Through a series of interviews and a roundtable discussion, we gathered politically diverse input from stakeholders at the Tax Policy Center, Tax Foundation, Bipartisan Policy Center, and others. We engaged individuals who use tax data, such as tax analysts and economists, as well as those with experience producing tax data, including current and former IRS data officials.

Assessment of SOI Tax Data

Despite the importance of tax data, Americans and tax analysts cannot access and analyze SOI's statistical products in a timely and meaningful way. In particular, data published by SOI are often years delayed, SOI does not publish key data that the IRS collects, and SOI's data products are not user friendly or AI-ready. Operational and resource constraints prevent SOI from tackling these issues.

Publication Delays and Unpublished Data

Perhaps the most significant barriers to accessing tax data are SOI's publication delays.

Oft-cited tabulated topline tax data typically lags two to three calendar years. Currently, the most recent final individual and corporate tax data refers to income earned and taxes paid in 2022 and 2021, respectively.

Even more delayed is the public-use file (PUF), which provides detailed microdata on individual income tax returns. The most recent PUF is from 2015, and there is no public-use file on corporate income tax returns. Economists must use decade-old data to analyze tax returns as well as to forecast the potential effects of new policy changes a decade from now.

Moreover, despite the IRS gathering significant amounts of information from businesses, workers, families, and educational institutions, among others, SOI does not publish data on several key topics Americans care about, such as health insurance and student debt.

USAFacts identified multiple factors contributing to the two-to-three-year lag in topline data, the decade-long lag in the PUF, and SOI's inability to publish certain data.

Tax return filing timeline

One inherent limitation to producing tax statistics is the timeline of the tax filing season. The tax return deadline for a "tax year" is April 15 of the following year, but many taxpayers, particularly high-income households, commonly file for an extension and do not submit their returns until October 15th. In other words, many tax returns for 2024 were not collected until October 2025. So, it takes nearly a year for the IRS to receive all of the tax return data from which SOI publishes tax statistics. Despite this, policymakers could still significantly improve the timeliness of SOI's data and grow the agency's capacity to publish more data.

Ingesting data into the RAAS Division's data warehouse

The raw data the IRS maintains in its confidential, authoritative Master Files are neither standardized nor usable for statistical purposes. In order to prepare that data for statistical analysis, the RAAS Division runs computing programs weekly to extract data from the Master Files and ingest it into a data warehouse. In this process, the data is deidentified and standardized for statistical analysis.

While maintaining a separate data warehouse adds a step to the production process, doing so serves several important purposes. The RAAS data warehouse enables the IRS to have a historical record of tax returns, prepares that data for research, ensures the research system does not impact the filing season process and vice versa, and takes a vital step to preserving the privacy to tax returns. This way, the IRS is also able to provide research data suitable for advanced analytics like economic modeling.

However, RAAS lacks modern tools, such as APIs, to extract data from the authoritative Master Files and ingest it into the data warehouse. As a result, SOI staff often must do so manually, which takes considerable time and manpower.

SOI access to the RAAS Division's data warehouse

After RAAS ingests data into the warehouse and prepares it for statistical purposes, SOI prepares its statistical products by extracting data from the same data warehouse. However, SOI has a separate data infrastructure from RAAS, and its current software and database architecture are both outdated and fundamentally different. In particular, the RAAS data warehouse uses different operating systems, a different firewall, and a different database architecture than SOI. This separation and SOI's outdated technology prevents the statistical agency from easily using AI and machine learning to directly access RAAS's data warehouse to ingest and clean the data for statistical products. As a result, SOI staff must manually extract data from the warehouse, taking considerable manpower and adding to the time it takes to produce tax statistics.

SOI's steps to ensure privacy and data quality

Ingesting administrative tax data from the IRS Master Files is just the first step in ensuring high-quality statistics. SOI must take additional steps to ensure data quality and maintain individual tax return privacy.

SOI takes several steps to make tax data suitable for producing final official statistics. To start, the agency standardizes data to ensure that similar information is reported in the same way across all returns. This can require rearranging information that may be reported on the wrong lines of a form, correcting errors, or extracting additional details from documentation provided by taxpayers as separate attachments. SOI also codes information such as occupations, industry, and marital status, among other categories, to make the data statistically useful. Lastly, the agency imputes missing return information, particularly regarding returns filed after the close of SOI's sampling period. Only after completing these steps are data suitable for producing final official statistics.

Although RAAS officials deidentify the tax data when they ingest it into the data warehouse and process it for statistical purposes, that by itself does not provide privacy protection sufficient for SOI to publish statistical products. Information from tax returns, like income, number of children, and city, in combination with data readily available on social media or other public sources, can be used to identify taxpayers. So, SOI takes several additional steps to protect privacy in its tabulated topline data. For example, SOI may choose not to publish certain data if there is a high risk that it could be used to identify a specific individual's tax return information. This can mean eliminating a data item entirely from a table, collapsing specific cells within a table when too few taxpayers are represented, or grouping the data presented in coarser categories to reduce risk. SOI might also add noise in a structured way by, for example, using modified differential privacy methods — a statistical framework to release information about a dataset while protecting the privacy of individuals.

The decision to release tabulated data therefore requires weighing the risk to individual privacy against the utility of the final product. For example, choosing not to publish data for a particular item or grouping data in ways that reduce detail limits the usefulness of its statistical products. In some cases, like the PUF, methods for protecting individual privacy can harm data quality by introducing bias to the data, as described below.

While these steps SOI takes to produce data are vital to retain trust among both taxpayers and data users, its outdated technology platforms prevent the agency from doing so efficiently. SOI is not equipped to integrate AI or other forms of modern technology to accelerate the process of preparing the data for publication and adding privacy protections. Instead, SOI performs much of this work manually or using a set of 'R' programs, which is very time intensive, leads to significant publication delays, limits capacity to publish additional statistics, and is not easily responsive to emerging issues or changes in tax law.

Challenges with the PUF

SOI's PUF is a privacy-protected microdata file that represents the population of individual income tax returns in a given year. Independent, nongovernmental budget forecasters commonly rely on it to project the 10-, 20-, or even 30-year effects of tax policy proposals. Notably, the Congressional Budget Office relies on the PUF to forecast the effects of tax policy changes, as it does not have access to confidential tax microdata except for data on one limited provision in the tax code. Researchers can also use the data to more closely understand how previous tax policy changes impact taxes paid and post-tax income.

As stated earlier, the PUF's release is very delayed, with the most recent being from 2015. This means broad groups of tax economists did not have microdata to inform lawmakers of the impact of the 2017 TCJA prior to the scheduled expiration of the law's individual income tax provisions in 2025. As a result, when lawmakers worked to enact OBBBA, which extended, made permanent, and augmented those tax provisions, the public lacked detailed data to assess their impacts on workers, earnings, or investments.

Plus, PUF data utility concerns have increased over time because SOI has had to take increasingly significant steps to protect the privacy of the taxpayer data. These steps have included reducing content by eliminating geographic data, for example. SOI also collapses data on the highest income taxpayers into a single, aggregated record. Where there are extreme outliers, SOI has also suppressed data on an ad hoc basis. Each of these steps limits the ability of analysts to model certain economic behavior. The ad hoc nature of SOI's privacy protection strategy also makes it impossible for analysts to account for any bias these methods may introduce in their models.

Advances in technology and data science mean that bad actors require less information about an individual return to identify the household to which it belongs. Consequently, the legacy approach to preserving privacy in the PUF is no longer considered sufficient to protect individual identities.

Ease of Access and Usefulness of SOI Data Products

SOI's data products are not easy to find, access, or use online. To start, SOI does not have a standalone independent website. Instead, all its data products are hosted within the IRS's main website. Consequently, SOI is out of compliance with Title III of the Foundations for Evidence Based Policymaking Act of 2018[viii] and the resulting OMB Trust Regulation, which require statistical agencies to provide data through their own separately branded websites.[ix] For comparison, while SOI is located at irs.gov/statistics, the three other principal economic statistical agencies, BLS, BEA, and Census, have standalone websites that are branded separately from their parent agency.

Additionally, SOI's webpages do not meet modern user experience standards. One tax analyst noted to USAFacts that SOI's webpages have not changed since his career started two decades ago. There are no user interfaces to search for data. Instead, SOI's webpages commonly contain walls of downloadable excel files, with one excel file for each year of a single topic. Users must also have a strong working knowledge of the data tables, the US tax system, and any year-to-year tax law changes to accurately answer even simple questions about taxes and income.

USAFacts produces standards for federal data products to maximize ease of access to government data and has contributed to federal efforts to prepare government data for AI. These two evaluation frameworks can be applied to SOI to further illustrate how its data products are insufficient for disseminating statistics in the modern age.

Federal Data Excellence evaluation

In 2024, USAFacts and the Partnership for Public Service launched the Federal Data Excellence (FDE) program.[x] It is composed of two parts: the FDE Standards — a set of best practices to make government data products easy to access and use for the general public — and the FDE Awards, for which we applied the FDE Standards to evaluate more than 50 nominated government data products and recognized those with particularly user-friendly webpages and interfaces. The FDE Standards contain 22 criteria across the following four domains:

Data accessibility
Does the product offer data formats that are easy to download and interact with, and does it include clear descriptions that orient the user to the data product and collection program? This includes the ability to use filters, the existence of a landing page with information about the data product, and whether the text on a landing page is "clear, simple, meaningful, and jargon-free."

Interoperability
Does the data product have metadata and variables that are both human-understandable and machine-readable to facilitate integration with other data sources? This includes providing and clearly displaying methodology and posting variable names free of special characters.

Helpfulness to the user
Does the product provide features that meet common user needs? This includes providing variable definitions, a point of contact, and a suggested citation, among other features.

Temporal characteristics of data
Does the data product appropriately note and discuss the temporal characteristics of the data and timeliness of the data collection program? This includes posting and updating the publication cadence, posting the publication date, and discussing the appropriateness of comparing data over time.

FDE results for SOI's individual income tax data

To understand the accessibility and usefulness of SOI's data products, USAFacts applied the FDE Standards to one of SOI's most important data products: individual income tax data.[xi] Overall, SOI's individual income tax data performs well on helpfulness to users but falls short in most other regards, particularly data accessibility.

Data accessibility

SOI's individual income tax data is not easy to navigate or download. SOI's individual income tax data lacks filters or any data navigation tool for customizing data to a user's needs, does not have an easy-to-understand landing page, and lacks an API or FTP server for bulk data access.

Helpfulness to the user

SOI's data does contain some information to help orient users to the data. It achieves this by providing variable definitions, addressing uncertainty in the data, and providing accompanying reports, suggested citations, and a point of contact. SOI also publishes useful cross-sections, takeaways, and views of the data in the SOI Bulletin, a quarterly publication covering a range of tax topics.[xii]

However, although the data product has helpful variable definitions, a user must pair them with a strong working knowledge of provisions in the tax code to comprehend the data itself. Importantly, an increase in the standard deduction or other changes in the tax code can influence year-to-year changes in SOI statistics, such as taxes paid or income earned. While SOI's major reports and quarterly publications describe annual changes to tax law, that information is difficult to find when users access the data tables directly. Consequently, many users do not have sufficient context to understand when those changes take effect. This could leave users, particularly those who do not closely follow changes to tax laws, to misinterpret SOI's individual income tax statistics.

Additionally, while the SOI Bulletins contain useful cross-sections, takeaways, or views of the data to guide the user, they are not very user-friendly. The articles are often dense, written with jargon, and only published as downloadable PDFs. As a comparison, the BLS's Spotlight on Statistics provides a series of user-friendly blog posts and data visualizations that illustrate the meaning of their statistics. For example, a June 2024 Spotlight on Statistics post provided a cross-sectional overview of the transportation and warehousing industry, including the distribution of workers by each sub-industry and their demographic characteristics.[xiii]

Interoperability

SOI provides methodology for individual income tax data and makes it findable by the user. Transparent methodology that details how the data came to be provides vital context that lets a data user (or an LLM) ask the right questions and learn from the data.

However, the data product does not contain any other metadata, variable names are not short and meaningful, and variable names contain special characters. As a result, the data products do not facilitate integration with other data sources because they are challenging to work with programmatically.

Temporal characteristics of data

The webpage hosting SOI's individual income tax data contains the date of the most recent update.

However, the amount of available historical data is not explained and the appropriateness of comparing data over time is not discussed. SOI's release schedule could also be more effectively discoverable. This could leave users with questions about when the next year of data will be published and how to interpret year-to-year changes in the data. Given that the data product also does not flag tax law changes, users could misinterpret trends in the data.

SOI data products are not AI-ready

Many Americans are consuming information in entirely new way thanks to advancement in artificial intelligence (AI). Government data providers, including SOI, are not yet prepared for this. In particular, government data are frequently presented in ways unsuitable for incorporation into information provided by today's AI platforms. Large language models (LLMs) are often unable to cite government statistics in answers to users' questions. Moreover, when they do cite government statistics, they do so inaccurately or without proper context.

In 2024 and 2025, USAFacts partnered with the Department of Commerce to develop guidance on preparing government data for AI. When applying that work to SOI, it is apparent that SOI's data products, like the data products in most other statistical agencies, are not AI-ready. While SOI's products may meet the standards required to be machine-readable, they lack proper documentation and metadata needed for its data to be machine-understandable. As a result, even if an LLM like GPT-5 or chat application like ChatGPT can read SOI's data, it does not have the context needed to use or discuss the data accurately. So, an LLM response that uses SOI data may misinterpret what the data means.

Additionally, SOI does not create, curate, or publish its own AI evaluation datasets for public use. Evaluations ensure that AI responses from popular chat interfaces and models, like ChatGPT, are accurate. For data providers, they highlight any issues with data or metadata retrieval and comprehension. Finally, they provide AI researchers with insight into how data publishers anticipate usage and help highlight caveats or limitations in the datasets.

Resource and Operational Constraints

Underlying these issues are resource and operational constraints that prevent SOI from making upgrades to its internal technology and external data products. After accounting for inflation, SOI's funding has been decreasing for decades. Additionally, SOI is organizationally buried within the IRS, preventing its needs from being an institutional priority.

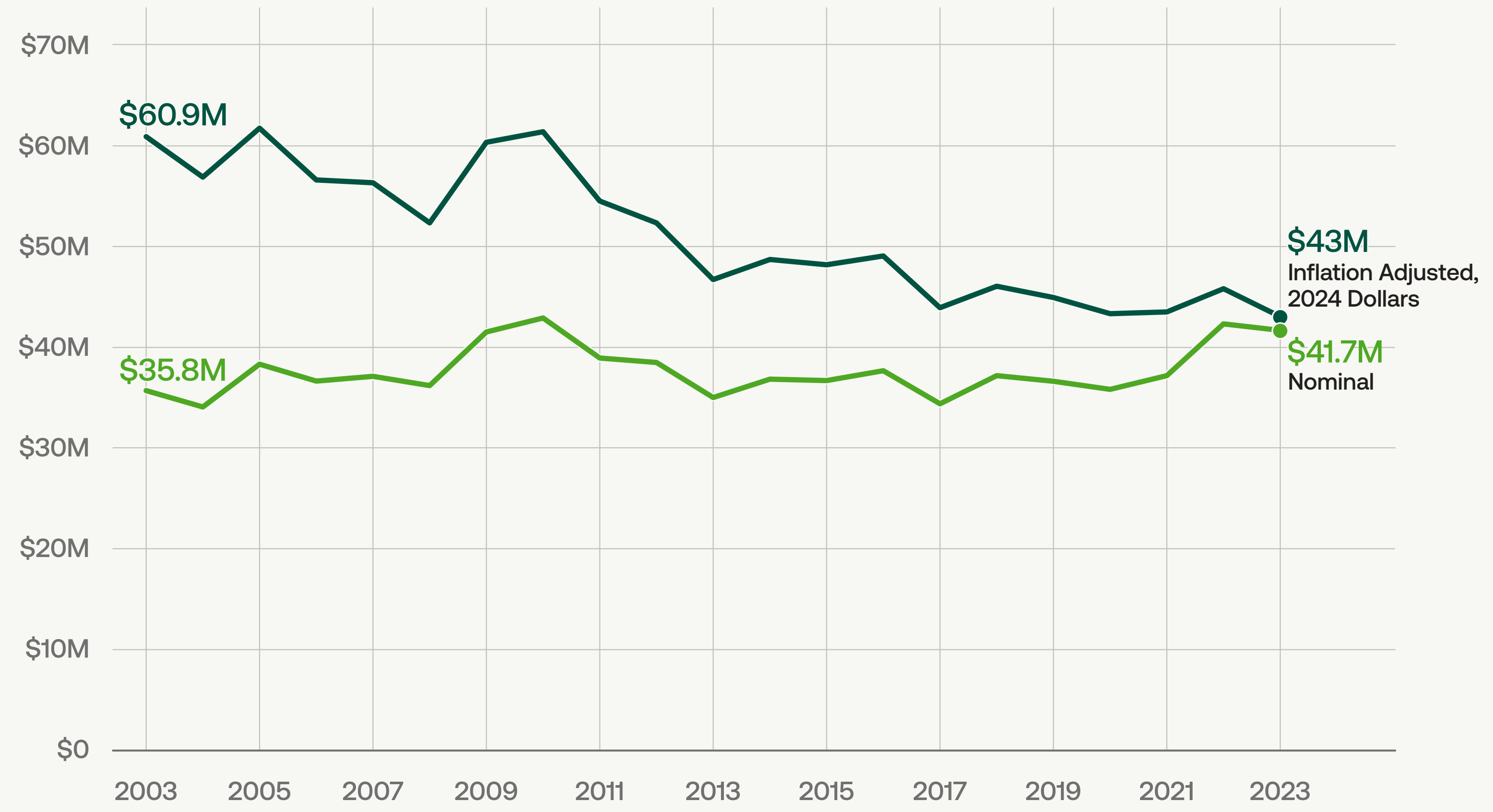
Resources

Although the Revenue Act of 1916 requires the IRS to publish statistics based on taxes, no legislation authorizes funding for SOI and Congress does not separately appropriate its budget each year. Instead, SOI's funding falls within a broader appropriations category: the IRS's operational budget. In effect, SOI's budget is largely at the discretion of IRS leadership and Congressional oversight of the statistical agency is limited.

This situation is unique relative to SOI's peer economic principal statistical agencies. Congress has enacted authorizing legislation and appropriates funding for the Census, BEA, and BLS separately from their parent agencies, the Departments of Commerce and Labor.

Without a separate appropriation, SOI's funding has decreased over the past two decades after accounting for inflation. Chart 1 contains SOI's funding from 2003 to 2023, with 2023 being the most recent year of data.

Chart 1: SOI Funding, 2003-2023 ($ millions)[xiv]



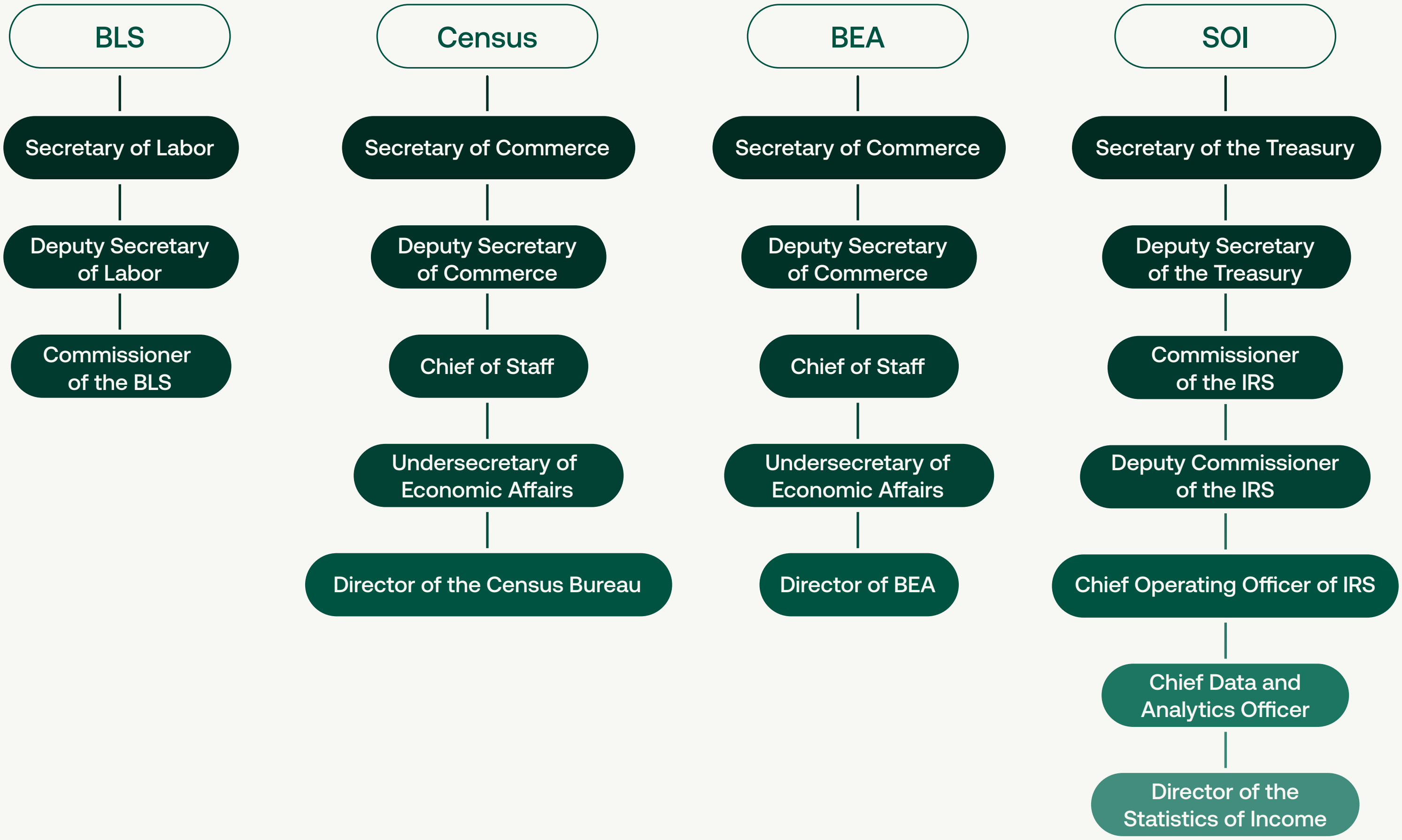
In nominal terms, SOI's budget increased from $35.8 million in 2003 to $41.7 million in 2023, or 16.5% over twenty years. However, that increase did not keep pace with rising prices. After accounting for inflation,[xv] SOI's budget, in 2024 dollars, decreased from $60.9 million to $43.0 million, or 29.4%. So, the purchasing power of SOI's budget has fallen by more than a quarter over the past two decades. For comparison, during that period, the real, inflation-adjusted budget of the BEA and Census increased by 20.2% and 59.5%, respectively, while BLS's inflation-adjusted budget declined by 14.1%.

Operational constraints

Relative to the leaders of other economic principal statistical agencies, the SOI director's role is not very senior within the IRS's organizational structure. It is particularly far removed from its cabinet level secretary (the Secretary of the Treasury). The SOI director is also not on the IRS's Senior Executive Team, which serves an advisory role to the IRS Commissioner.

Currently, the SOI director reports to the IRS's chief data and analytics officer and there are a total of five offices between him and the Secretary of the Treasury. By comparison, only one position separates the BLS commissioner from the Secretary of Labor (the Deputy Secretary of Labor), while the Census director and the BEA director each have three layers of leadership between them and the Secretary of Commerce (see the organizational chart on the following page).

Chart 2: Organizational Charts: SOI Compared to Census, BEA, and BLS


BLS
Secretary of Labor
Deputy Secretary of Labor
Commissioner of the BLS
Census
Secretary of Commerce
Deputy Secretary of Commerce
Chief of Staff
Undersecretary of Economic Affairs
Director of the Census Bureau
BEA
Secretary of Commerce
Deputy Secretary of Commerce
Chief of Staff
Undersecretary of Economic Affairs
Director of BEA
SOI
Secretary of the Treasury
Deputy Secretary of the Treasury
Commissioner of the IRS
Deputy Commissioner of the IRS
Chief Operating Officer of IRS
Chief Data and Analytics Officer
Director of the Statistics of Income

The absence of SOI's perspective from IRS and Treasury leadership could make the needs of the statistical agency more difficult to prioritize. With the enactment of OBBBA, the IRS is now charged with implementing several major tax code reforms. Without SOI's voice on the IRS's Senior Executive Team, the agency may not get the input necessary to implement these tax changes in a way that would enable efficient data collection and production of tax statistics. This could introduce more barriers to producing timely and relevant tax statistics as Congress and the American public consider the legislation's impact.

Recommendations

Based on USAFacts' assessment of SOI's limitations, below are a series of recommendations to help SOI produce tax statistics in a timely, relevant, and accessible way. These recommendations include internal technical solutions, improvements to the dissemination of SOI statistical products, and mitigation of the agency's resource and operational constraints.

Limiting Delays and Publishing More Data While Preserving Privacy and Quality

Recommendation 1: Introduce internal APIs to accelerate data ingestion into the RAAS's data warehouse

The RAAS data warehouse serves an essential function of preparing tax data for research and preserving the privacy of individual returns. However, the process for ingesting data into the data warehouse is long and burdensome for IRS officials because RAAS lacks modern tools and often must do so manually.

USAFacts recommends enabling APIs to IRS's authoritative Master Files. This would modernize and streamline the process of ingesting varied confidential data sources into RAAS's data warehouse. It would ultimately quicken the first step needed to produce tax statistics.

Recommendation 2: Modernize SOI's technology for accessing and analyzing the RAAS's data warehouse

SOI's data infrastructure needs updating. SOI's current database architecture and software prevents the use of machine learning or LLMs to access RAAS's data warehouse to ingest and clean data for statistical products.

SOI's data infrastructure should be re-architected to a modern state that is more compatible with RAAS's data warehouse and that allows the statistical agency to use AI, machine learning, and LLMs. This would enable SOI to access RAAS's data warehouse more efficiently to produce statistical products.

Recommendation 3:
Leverage modern technology to accelerate steps to strengthen data quality and preserve the privacy of tax returns

While tax data is deidentified when ingested into the RAAS data warehouse, SOI officials must take additional steps to add noise to the data while ensuring data quality. Lack of modern technology at RAAS slows down deidentification and ingestion into the RAAS data warehouse, but it also slows down the subsequent steps SOI must take to prepare the data for publication and preserve the privacy of tax returns.

Introducing modern AI and machine learning technologies at SOI could speed up both data quality and privacy measures. It would enable SOI to quickly identify and correct errors in administrative tax data that might be introduced by taxpayers or during IRS processing. The same technology could also enable SOI to spot data anomalies and high-income outliers who are more easily identifiable if their information were made public. Lastly, it would help SOI be more responsive to emerging issues and changes in tax law.

For example, RAAS has experimented with using recommender systems, like those used by Netflix to recommend content, to identify outliers in the data. This approach could enable SOI to quickly spot and address errors and/or easily identifiable tax return information. This approach, and other AI-based approaches, can speed up the data cleaning and coding processes necessary to prepare data for statistical processes.

Recommendation 4:
Deploy synthetic data files to speed up and improve the Public Use File

Accelerating current efforts to produce synthetic PUFs and a validation server would give broad groups of researchers more timely access to microdata. One example of this: the IRS is currently working with the Urban Institute to develop synthetic PUFs that use advanced statistical methods to approximate the actual tax data that will be released as a PUF. This effort is also working to develop a validation server that would give analysts an automated way to run the models they develop with synthetic data on the actual confidential IRS data. There are several advantages to the proposed synthetic files:

- By approximating, rather than using, confidential data, synthetic PUFs would guarantee that the privacy of tax returns is maintained.
- Synthetic data files would be much quicker for SOI to release, enabling the agency to publish more data, and faster.
- Synthetic PUFs could be released with error estimates, enabling analysts to understand bias in the sample and account for it.
- The validation server would ensure that the results of the models built with synthetic data are validated by the actual confidential data.

The principal disadvantage of the synthetic PUF is that users would not be able to work with the actual microdata of individual tax returns, resulting in less precise estimates. While somewhat limiting, the validation server largely addresses this concern by giving researchers the ability to run statistical models developed using the synthetic PUF on the actual confidential IRS data. Moreover, given that the most recent existing PUF is from 2015, researchers today largely do not have access to any recent microdata.

The current synthetic data project is underway but the aforementioned technical limitations at SOI and the RAAS data warehouse present obstacles. Ensuring the project has proper resources and continues would make tax statistics more accessible to broad research audiences. Moreover, improving technology used to ingest data into the data warehouse and access it for statistical purposes would also accelerate the development of synthetic PUFs.

Making SOI Data Easier to Access and More Useful

Recommendation 5: Create a standalone website for SOI data products

Developing a standalone website, independent of the IRS website, would make SOI data easier to find and provide more functionality to tailor webpages to the needs of SOI's products. It would also ensure that SOI complies with the Foundations for Evidence-Based Policymaking Act and OMB's Trust Regulation.

Recommendation 6: Make SOI data products easier to access and download

Deploy data filters and/or data navigation tools that allow a user to customize and shape the data according to their interest, develop easier to follow data product landing pages that explain the data in plain language and act as central hubs of useful information about the data products, and introduce an API or FTP server to enable bulk downloads. The combination of these changes would make SOI's data products easier to navigate, work with, and access.

Recommendation 7: Simplify concepts and provide user-friendly content based on the data

Simplify terms and concepts so that a user does not need a strong working knowledge of provisions in the tax code to comprehend the data. This would ensure that SOI complies with the Plain Writing Act of 2010, which requires government agencies to publish in understandable language.[xvi] SOI data products should also include descriptions of changes to tax law that are effectively discoverable to prevent users from misinterpreting year-to-year changes in tax statistics.

Additionally, make the cross sections, takeaways, or views of the data in SOI Bulletins more user-friendly. These could include shorter blog posts and data visualizations that are jargon-free and help users interpret tax data.

Recommendation 8: Improve the interoperability of SOI data products

Integrate metadata into data products and adjust variable names to make them short, meaningful, and free of special characters. This would help facilitate integration of SOI data products with other data sources and improve both machine- and human-readability.

Recommendation 9: Make details on the timing of data releases and the amount of historical data available easy to find

Post the publication cadence of the data products on the products themselves, explain the amount of historical data being made available to the user, and discuss the appropriateness of comparing data year to year. This would help users anticipate when SOI will publish new data and how to interpret year-to-year changes in the data.

Recommendation 10: Make SOI data products AI-ready by integrating metadata, documentation, and evaluations

Current SOI spreadsheets and tagged PDF documents meet government standards for being machine-readable. However, they are not accompanied by enough metadata and other documentation to be machine-understandable. As a result, they cannot be correctly interpreted by external AI tools like ChatGPT.

SOI should provide the context necessary for LLMs to accurately retrieve and summarize data. This means providing comprehensive documentation and metadata through data dictionaries, knowledge graphs, and semantic layers, among others.

Additionally, PDFs require extra processing steps to be ingested in AI context pipelines and can lead to errors. Where possible, SOI should instead use plain text files published in HTML or .txt files and make them accessible to web-crawlers. When publishing charts, SOI should use image files (such as PNG or JPG) and JavaScript objects, which are more accessible to modern multimodal LLMs than PDFs. Accompanying SOI charts and figures with plain-text descriptions or captions would also enable better integration with AI.

Lastly, in line with the recent AI strategy guidance published on AI.gov, we encourage SOI to create, curate, and publish its own AI evaluation datasets for use by the public.

Addressing SOI’s Resource and Institutional Constraints

Recommendation 11: Congressionally authorize funding for SOI and appropriate its funding separately from the rest of the IRS

Enact legislation that authorizes funding for SOI, gives the statistical agency more independence, and empowers Congress to appropriate its annual funding levels separately from the IRS’s operational budget. This would make Congress responsible for setting SOI’s resources, just like it is for Census, BEA, and BLS, enhancing its oversight into the IRS’s statistical agency. Moreover, SOI would be more able to anticipate its resources and set its priorities with a predictable, more stable annual appropriation.

Recommendation 11 (alternative):

An alternative to legislatively authorizing SOI and separately appropriating its funding is creating an entry for SOI in the Department of the Treasury's budget. This means that the Treasury would establish a baseline funding level for SOI that the IRS would be required to provide. This would ensure that SOI would always have enough resources to meet the Treasury's needs (specifically, the Treasury's Office of Tax Analysis).

Recommendation 12:
Increase funding for SOI

Provide upfront funding for SOI to invest in new technologies that would expedite the statistical production process, increase capacity to publish more data, continue to preserve privacy, improve the accessibility and usefulness of its data products, and prepare its data for AI.

In addition to upfront investments, SOI needs the capacity to be able to continuously improve its products and processes. So, we recommend increasing SOI's annual budget as well. As a starting point, lawmakers could restore SOI's funding to levels comparable to what it received in the mid-2000s (after accounting for inflation). This would translate to increasing SOI's funding by about $20 million, from the around $40 million it receives annually today to about $60 million dollars.

Recommendation 13:
Elevate the SOI Director to IRS's leadership team

Change the IRS's reporting structure to have the SOI director report to the IRS Deputy Commissioner and appoint the SOI director to the IRS's Senior Executive Team.

By reporting to the IRS Deputy Commissioner, the SOI director would be in a position comparable to those who lead the economic statistical agencies at the Department of Commerce (BEA and Census). Additionally, adding the SOI director to the Senior Executive Team would enable SOI to provide input to IRS leadership so that tax filing processes are carried out in a way that facilitates data collection and production. This will be particularly important over the next few years, as the IRS implements OBBBA's tax reforms and lawmakers and the public will need timely statistics to evaluate the impacts on taxes and incomes.

Conclusion

SOI is responsible for producing statistics on taxes and income based on millions of tax documents submitted to the IRS. With Congress increasingly enacting public policy through the tax code over the past several decades, lawmakers and the broader public vitally need accessible, timely, and useful tax statistics. However, currently, SOI's statistical products lag multiple years, miss key data, are not easy to find or use, and are not AI-ready. As the IRS implements tax reforms introduced by OBBBA, there is new urgency to invest in upgrades to SOI's internal technology, updates to its webpages, and the development of new data products. Doing so would enable SOI to publish more, timelier data and make its data products more user- and machine-friendly, while at the same time continuing to preserve the privacy of individual tax returns.

End Notes

i. H.R.1, One Big Beautiful Bill Act, 119th Congress, https://www.congress.gov/bill/119th-congress/house-bill/1/text.

ii. H.R.5376, Inflation Reduction Act of 2022, 117th Congress, https://www.congress.gov/bill/117th-congress/house-bill/5376.

iii. H.R.1, An Act to provide for reconciliation pursuant to titles II and V of the concurrent resolution the budget for fiscal year 2018, 115th Congress, https://www.congress.gov/bill/115th-congress/house-bill/1/text.

iv. Bettye Jamerson and Robert A. Wilson, "Statistics of Income: 75 Years of Service," Internal Revenue Service, https://www.irs.gov/pub/irs-soi/75yearssoi.pdf.

v. 1.1.18 Research, Applied Analytics and Statistics Division, Internal Revenue Manual, Internal Revenue Service, November 8, 2022, https://www.irs.gov/irm/part1/irm_01-001-018.

vi. Taylor R. Knoedl, "The Federal Statistical System: An Overview," R48161, Congressional Research Service, August 19, 2024, https://www.congress.gov/crs-product/R48161.

vii. 1.1.18 Research, Applied Analytics and Statistics Division, Internal Revenue Manual, Internal Revenue Service, November 8, 2022, https://www.irs.gov/irm/part1/irm_01-001-018.

viii. H.R.4174, Foundations for Evidence-Based Policymaking Act of 2018, 115th Congress, https://www.congress.gov/bill/115th-congress/house-bill/4174/text.

ix. Fundamental Responsibilities of Recognized Statistical Agencies and Unites, Offices of Management and Budget, 5 CFR Part 1321 (2024), https://www.federalregister.gov/documents/2024/10/11/2024-23536/fundamental-responsibilities-of-recognized-statistical-agencies-and-units.

x. Federal Data Excellence, USAFacts, https://usafacts.org/research-and-initiatives/fde/.

xi. SOI tax stats – Individual statistical tables by size of adjusted gross income, Statistics of Income, Internal Revenue Service, November 8, 2024, https://www.irs.gov/statistics/soi-tax-stats-individual-statistical-tables-by-size-of-adjusted-gross-income.

xii. SOI Tax Stats – SOI Bulletins, https://www.irs.gov/statistics/soi-tax-stats-soi-bulletins.

xiii. Mary Dorinda Allard and Kennedy Keller, "Keeping America Moving: Employment in transportation and warehousing industries," Bureau of Labor Statistics, July 2024, https://www.bls.gov/spotlight/2024/keeping-america-moving-employment-in-transportation-and-warehousing-industries/home.htm.

xiv. Statistical Programs of the United States Government, Office of Management and Budget, Fiscal Years 2003-2023. All data is actual funding amounts except for 2022, which has appropriated levels.

xv. Consumer Price Index, All Urban Consumers, Bureau of Labor Statistics, 2003-2023, https://www.bls.gov/.

xvi. H.R.946, Plain Writing Act of 2010, 11th Congress, https://www.congress.gov/bill/111th-congress/house-bill/946.