



www.apdu.org

Association of Public Data Users

NEWSLETTER, June 2009, Vol. 32, No. 2

Welcome to the latest edition of APDU Newsletter. Please send your comments to [info@apdu.org](mailto:info@apdu.org).

---

## HIGHLIGHTS

- [The President's Corner](#), Andrew Reamer, The Brookings Institution
- [Increasing Government Transparency through Data.gov](#), Cindy Taeuber, CMTaeuber & Associates
- [Challenges Await Next Census Chief](#), Phil Sparks - Mary Jo Hoeksema, The Census Project
- [Agency IT Centralization and Ramifications for Statistical Agencies](#), Steve Pierson, American Statistical Association
- [2010 Census Advisory Committee Meeting Report](#), Ken Hodges, Nielsen Claritas

## The President's Corner

Andrew Reamer, The Brookings Institution

Welcome to the June 2009 issue of the APDU newsletter. Responding to member feedback on the April issue, we have sharpened our focus in two ways. Per member requests, we've reduced the number of articles to three. And we've sharpened their analytic focus. So you will see pertinent, timely articles by former Census Bureau staff Cindy Taeuber on the newly released Data.Gov website, by Census Project leaders Phil Sparks and Mary Jo Hoeksema on the challenges facing Bob Groves when he takes the reins of the Census Bureau, and by Steve Pierson, Director of Science Policy at American Statistical Association, on concerns raised by the consolidation of statistical agency IT systems with those of parent organizations. Steve takes a particular look at the situations facing the IRS Statistics of Income Division and the USDA Economic Research Service. We hope you enjoy these articles and encourage you to write the authors with any comments you might have.

The newsletter also include a summary report by Ken Hodges, APDU's representative on the Decennial Census Advisory Committee meeting in May, with a link to the full report on the APDU website. Among other matters, Ken tells us about controversies regarding the 2010 Census communications efforts and efforts to count prisoners, Katrina victims, and same sex partners.

As you know from APDU Data Updates, the president's FY2010 budget contained a number of positives, and some negatives, for federal statistical agencies. APDU has posted links to each agency's congressional budget justification at <http://www.apdu.org/resources/StatAgeBudReq.htm>.

Finally, I want to remind you about a number of upcoming APDU activities. Next week, you will receive a member survey, prepared by board member Paul Zeisset, that asks you to provide feedback on the variety of APDU services. As you know, APDU has expanded and reoriented a number of services, including the newsletter—please let us know what you think.

On July 16 from 2:00-3:00 EDT, APDU will sponsor our second webcast, "An Overview of the National Household Travel Survey (NHTS)," with NHTS manager Heather M. Contrino, Federal Highway Administration. The NHTS provides statistical measures of travel demand and travel behavior of the American public. The 2008 NHTS will capture the impacts of gas price volatility and the recession. I encourage you to tune in and learn more about this valuable, interesting effort, which is organized by board members Christine Pierce and Ed Christopher.

And then, of course, APDU will be sponsoring its annual conference on September 24-25 at the Brookings Institution, with the theme of "Measuring Recovery through Federal Data."

Conference co-chairs Bob Parker and John Kort have put together an exciting series of sessions regarding the ways in which the federal statistical system provides the nation with data that tracks economic conditions and efforts to improve them. Shortly, you will receive an announcement and a link to the agenda. I do hope to see you in Washington in late September for this important, exciting event.

In the meantime, I wish you a good entry into summer and look forward to writing you again in the next newsletter.

## **Increasing Government Transparency through Data.gov**

Cindy Taeuber, CMTaeuber & Associates

What organization has more data to analyze the results of federal activities than the federal government itself? Yesterday the easy answer was, “No one.” We may be at the start of a meaningful change to that answer. On May 21<sup>st</sup>, the federal government launched Data.gov, a one-stop online source to help the public find free federal data sets and to mix, match, and summarize data as they prefer.

Data.gov became a reality just 120 days after President Obama signed a memorandum asking for recommendations towards building a foundation of transparency, accountability, and responsibility in the federal government. The Obama Administration’s new Chief Information Officer, Vivek Kundra, used his experience in providing public data about the workings of Washington DC’s local government to fuel his determination to bring “democratized feedback” to the federal government through the release of thousands of federal data files in standardized formats.

### **What is Data.gov?**

Data.gov represents the beginnings of a searchable catalog of publicly-available data sets and tools from the federal Executive Branch. The site is a data hub, not a data warehouse, as it provides links to federal agency databases and analysis tools.

The purpose of the website is to provide “raw” data and data tools from federal agencies so that users can summarize and categorize data as they please. At present, about 30 agencies are in the catalog, but some, such as the FBI, offer data tools only, not access to raw data. It would be less confusing if data sets that are accessed only through data tools were searched separately from those that provide raw data. In other cases, what the site calls “raw data” are actually public-use files that are a small sample of full files and provide no individually-identifiable data – such as the Social Security Administration’s one-percent sample of beneficiaries and the Census Bureau’s public use file of the American Community Survey.

As of this writing, there are over 100 data sets and tools available, the most active providers being the U.S. Geological Survey, the Internal Revenue Service, and the Weather Service. Data on tracking snow, rain, and hail, for example, are not likely to meet the stated objective of providing data to garner new ideas from the public on policy and to investigate how well government works. Data.gov expects to add scores more data sets within a month on topics from education to energy to prisons.

The data sets are offered in several standardized formats so that raw data can be manipulated or programmed according to the needs of data users. This includes text files for spreadsheets, XML for the exchange of data between Web services, and formats that feed data into mapping services.

The site currently provides links to government agencies with 28 tools. Some help data users make tables, graphs, and maps. Others are “widgets” (applications) that allow data users to paste data from a federal agency’s website on to their own site. There are also data extraction tools, such as the Census Bureau’s American FactFinder. Even though this tool does not help data users extract raw data from the Census Bureau’s data files, it does provide data in the form of maps, tables, and reports from a variety of Census Bureau sources.

Data users can rate the usefulness of the data sets that are available, although the difference in the meaning of two categories (“data utility” and “usefulness” of the data set) is not clear.

### **What are the challenges for Data.gov?**

Data.gov is a welcome investment in the nation’s information infrastructure—and yet it faces challenges.

One challenge is how to expand the number of data sets in a way that is cost effective. Some existing data sets, for example, while valuable, are from old systems and it would be expensive to modify them for online access. There are also issues of not duplicating work or taking over work that is better handled by the agency responsible for the data.

As agencies nominate data sets for inclusion, the managers of Data.gov must ensure national security and protect confidential information. They should consider the effects of the perception that the government may not be protecting confidential information when it lists “raw data,” which implies data on individuals. The Census Bureau, for example, is shown in the list of agencies because it has useful data tools and public-use files based on samples, not because it provides “raw” data on individuals who can be identified. Census has long struggled to convince a skeptical public that it does not release personally-identifiable data and it would be costly if the decennial census is entangled in any misperceptions of the meaning of the term “raw data.” Other agencies with data on individuals face similar concerns with this misleading terminology.

Beyond the obvious challenge to Data.gov of refreshing the site with additional data sets on a continuing basis, there are the challenges of supplying data that are meaningful as well as seeking and implementing input from data providers and users. Not surprisingly, the initial launch has data sets that are limited in number and noncontroversial.

The rollout of Data.gov was implemented very fast and, ironically for an initiative on open government, with limited input from federal agencies and data users. The developers say that the site’s feedback mechanisms now provide a more open process. Obviously, wider and organized

input is an essential next step—data users and providers can help the developers of Data.gov to better understand their needs and resources and thereby inform decisions about priorities. There are already mechanisms for doing this, including the Advisory Committees of statistical agencies. Additionally, they could create a community of practice around Data.gov through GSA’s Intergovernmental Solutions (<http://colab.cim3.net/cgi-bin/wiki.pl/>).

### **What is the potential of Data.gov?**

APDU members may be scratching their heads at this point and asking, “What is the big deal?” The data sets currently shown on Data.gov are already available on federal websites, major social and economic data sets with summary data only apparently will not be shown, and there are other portals for accessing federal data sets.

In fact, there are benefits to Data.gov. It could become the single go-to source to search for a wide variety of federal data. Other portals, such as Data Ferrett, geodata.gov, and FedStats are more limited in the number and variety of data sets and do not have the same options for standardizing metadata and access to the data. It remains to be seen whether Data.gov can replace any of these.

Data.gov makes modest but important steps towards a uniform set of metadata and a standardized interface to data downloading. However, the metadata must be expanded to highlight when the same term is defined differently among data sets.

The goal of Data.gov is to eventually release thousands of government data sets online in their “raw” form so that the public can become more engaged in policymaking by analyzing information in their own way. While it remains to be seen just how much new data will become available, the White House is working with federal agencies to get it done. A useful upgrade would be to automatically notify data users as new data sets in a subject area of interest are added.

Kundra’s group is working to link data sets through unique identifiers, a move to meet Obama’s campaign promise to easily investigate, for example, the association between contracts and lobbying. Geographic codes are already unique identifiers; with a search function for geography, one could find every federal data set with information about MyTown USA.

In the end, Data.gov has the potential to mean easier access to much more data along with increased uniformity in the information available about federal data sets.

Comments?: Cindy Taeuber welcomes your comments at [cmtaeuber@comcast.net](mailto:cmtaeuber@comcast.net).

## **Challenges Await Next Census Chief**

Phil Sparks, Co-Director, The Census Project  
Mary Jo Hoeksema, Deputy Director, The Census Project

The next census is about counting and placing every person living in America as part of the constitutionally-mandated Census 2010. For the incoming director of the U.S. Census Bureau, overseeing a successful decennial census, however, will be akin to herding cats. The U.S. House of Representatives has finally awakened from its decade-long slumber regarding census oversight to discover its members' political fate is tied to a fair and accurate census. Some Republican members of the U.S. House and Senate are already charging the Obama Administration with "'politicizing" Census 2010—a year before the census is taken. The new Administration is still determining its priorities regarding not only Census 2010, but also the Bureau's other ongoing programs and surveys.

As if these challenges were not enough, the new Census Director will be forced to defend the bureau because of past problems with Census 2010 operational preparations and assure Congress that the next census will be non-political and effective. Congress is not the only entity questioning the Bureau's ability to conduct a successful Census 2010. Various population groups are also nervous about Census 2010. An organization of Hispanic ministers is urging a boycott of the next decennial due to fears that the census' confidential records will be compromised. Some gay rights organizations want a re-definition of married households on the census form to encompass the recent legalization of gay marriage in several states.

Upon his confirmation, the next Census Bureau Director, Dr. Robert Groves, will inherit all of these challenges as well as a Census 2010 operational plan he did not develop, but has to embrace and keep on track. In April, the Bureau initiated its Census 2010 plan by launching an initiative to verify the addresses of every household in the country. This first operational benchmark was accomplished by hiring hundreds of thousands of temporary Census Bureau employees. The final work of the next decennial census won't be over though until December 2010 when numbers are finalized.

We suggest that Dr. Groves, in addition to the daunting organizational tasks involving Census 2010, will face major challenges that could, if not handled well, undermine confidence in Census 2010. First, congressional committees are planning to dramatically increase their oversight of the Census Bureau. As a result, Dr. Groves should expect a higher level of congressional scrutiny and interest in the bureau's activities because of recent missteps, which he can't fix -- like the failure of the handheld computers -- and be prepared thoroughly to address them. Second, Dr. Groves should anticipate members of Congress pursuing changes to the decennial census form, not knowing that both the questions and the forms were approved months ago and are being printed. To avoid a backlash from key members of Congress and outside organizations who

advocate for these changes, Dr. Groves needs to develop a strategy. Third, Dr. Groves should expect more interest in the agency's partnership program and other communications plans.

In 2000, more than a hundred thousand organizations, including religious organizations, business groups, community organizations and foundations, helped reach out to various communities to ensure a complete census count. The Census 2010 partnership program is a necessary link to communities. But, some of the agency's "partners" - like ACORN - will be seen as controversial. Dr. Groves will have to reassure the public and policymakers that the partnership program must reflect America's diversity if it is to be effective.

On the communications front, the agency's multi-million dollar national advertising program to encourage Census 2010 participation recently received a highly publicized vote of "no confidence" from advisory committees established by the Census Bureau. Dr. Groves, a technical expert on survey design and response, has little outreach and marketing expertise. He will need to be a quick study to assure his own advisory committees and Congress that he can successfully facilitate a solution and restore credibility and productivity to the Census 2010 communications plan

Census 2010 is a complex political and massive operational act vital to American democracy. It has been that way since the first census was taken in 1790. Dr. Groves' success as the next census director depends on his professional abilities to manage Census 2010. He also needs an acute political antenna to understand the growing political anticipation regarding Census 2010. Something akin to herding cats.

Comments?: Phil Sparks and Mary Jo Hoeksema welcome your comments at [philsparks@ccmc.org](mailto:philsparks@ccmc.org) and [paaapc@crosslink.net](mailto:paaapc@crosslink.net).

## **Agency IT Centralization and Ramifications for Statistical Agencies**

Steve Pierson, Director of Science Policy, American Statistical Association

In reaction to the Federal Information Security Management Act of 2002 (FISMA), government agencies are centralizing their information technology (IT) resources, a move that jeopardizes the activities of federal statistical agencies. As Title III of the E-Government Act of 2002, FISMA mandates several measures for each agency to ensure information security. Agencies must balance the FISMA requirements with those of the E-Government Title V, the Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA), which mandates confidentiality protection of statistical data collections. At least two statistical agencies are known to be affected but this could be just the beginning.

The statistical community was alerted in April that the IRS was acting quickly to take control of IT resources of its Statistics of Income Division (SOI), including hardware, software, databases, and 80-100 SOI staff. The Economic Research Service (ERS) at the U.S. Department of Agriculture (USDA) faces a similar problem, with a proposal to consolidate all IT resources into one of four data centers, termed "Enterprise Data Centers" (EDC's). The 2010 USDA budget also proposes the consolidation of seven leased facilities in the Washington, DC area, with the possibility of integrating all subject agencies' data centers into a single facility. Since the ERS details are still emerging and perhaps the decisions further off, I will focus largely on the SOI IT consolidation issue.

How the statistics community responds to these efforts may determine to what extent such IT centralization proposals in other agencies will affect their statistical units.

### **Statistics of Income Division**

SOI provides information on annual income, financial and tax data. Its products are used for a variety of tax policy work, including revenue estimation, economic baseline development, and distributional analysis. The primary SOI customers are the Congressional Joint Committee on Taxation, the Treasury Department, Bureau of Economic Analysis (for its National Income and Product Accounts), the Federal Reserve Board (for its Survey of Consumer Finances), and the Congressional Budget Office. SOI also responds to outside requests from congress, constituents, and a variety of private researchers.

Like other statistical agencies and offices, part of SOI's challenge is its unique mission within IRS, which creates the potential for SOI to be underappreciated and/or misunderstood within the larger organization of IRS. The current attempt to take control of SOI IT resources goes back more than a year and is actually the third such move. According to Fritz Scheuren (SOI Director, 1980-1994), IRS took over SOI's IT resources in 1964 and moved them to Detroit. After an

internal rebuilding of the IT resources in the 1980's, the IRS steps to dissolve SOI in the 1990's, an effort that was ultimately unsuccessful.

### **The Concerns about IT Consolidation**

While acknowledging and appreciating the IRS data security issues, the statistical community is concerned with the IRS IT consolidation for three principal reasons, all having to do with SOI's ability to carry out its mission effectively and efficiently.

First, without control of its IT resources and direct access to its data, SOI's responsiveness and nimbleness would be at risk. While details of the consolidation are unclear, the general fear is that any SOI IT requests to a centralized IRS IT department would be superseded by non-SOI requests. One scenario apparently under consideration is that SOI staff would have to file a "Unified Work Request" to make changes to applications code, a process that would clearly bog down SOI work.

Direct and unencumbered data access is essential for SOI staff to produce their reports and ensure the accuracy and completeness of return data. Such data access is also vital to SOI clients who use SOI data in "dynamic engagement" (with SOI personnel) to investigate data anomalies, augment records, and create data extracts.

The second concern of IT centralization is that SOI, without control over its databases, would have a more difficult time collecting additional data necessary to produce its reports. SOI must frequently request from sampled taxpayers – primarily multinational corporations – data not provided in the original tax filings and does so in letters from the SOI director. To encourage the corporations to provide the data voluntarily, SOI must earn a data provider's trust, which is done by promising to protect the respondent's confidentiality. Without control of its IT resources, SOI would be unable to assure potential data providers that a firewall exists between SOI and the rest of IRS.

The third concern for IRS IT consolidation is the impact the loss of SOI personnel would have on timeliness and efficacy. The human capital and the accumulated institutional knowledge of a statistical office, while intangible, are some of the greatest resources in a statistical agency. To develop this resource, statistical agencies invest an enormous amount of time in the education and training of new employees and the professional development of all employees.

In the case of SOI, its statistical "culture" has been built over decades for the highly specialized work carried out by its economists, IT personnel, statisticians and other professionals. In a centralized IT structure, SOI personnel could be undervalued for their subject matter expertise, especially because attention in a broader structure could be primarily on their programming skills.

Indeed, while data security is regarded as a main driver for IT consolidation, there is also speculation that IRS wishes to tap the SOI expertise and talent. Already, SOI processes a substantial part of IRS administrative work, in addition to its own work. External auditors confirm the quality of SOI's IT operation. One review is reported to have found SOI IT several times more efficient than that of the Chief Information Officer's organization.

## **Principles and Practices**

The National Research Council's recently released *Principles and Practices for a Federal Statistical Agency* (fourth edition) echoes the first two concerns above, in its discussion of the principle of a "strong position of independence":

- i) "Other characteristics related to independence are that a statistical agency has the following: ... Authority to control information technology systems for data processing and analysis in order to securely maintain the integrity and confidentiality of data and reliably support timely and accurate production of key statistics." (p. 23)
- ii) "[The trust of its data providers and data users] is fostered when a statistical agency has control over its information technology resources and there is no opportunity or perception that policy, program, or regulatory agencies could gain access to records of individual respondents... A statistical agency also needs control over its information technology resources to support timely and accurate release of official statistics, which are often produced under stringent deadlines." (p. 23)

## **Calls for Action**

IRS IT consolidation has significant implications for the ability of SOI to carry out its critical work. Members of the SOI Advisory Panel, ASA, Congress and others have been active in addressing the threat to SOI's autonomy.

- Congresswoman Carolyn Maloney, the chair of the Joint Economic Committee, urged IRS Commissioner Shulman to exempt SOI from IRS IT consolidation.
- Congressional staffs for key committees have made inquiries and are monitoring the process. Prominent members of the tax policy community, including SOI Advisory Panel members, and I have met with the congressional staff to delineate the ramifications of an IRS IT consolidation for SOI.
- ASA's President Sally C. Morton has written officials at Treasury, OMB, IRS and elsewhere urging that SOI be allowed "to maintain authority over its IT functions and personnel ..."
- *Tax Notes* did a story on this issue, interviewing newly retired SOI Director Tom Petska, SOI Advisory Panel members and the IRS Research, Analysis, and Statistics (RAS) Division Director Mark J. Mazur, and covering many of the concerns mentioned above.

Generally reassuring letters in reply from Commissioner Shulman and Michael Mundaca, the Acting Assistant Treasury Secretary for Tax Policy have not soothed stakeholder concerns. At present, there is little news on the status of IRS's actions to centralize the IT resources.

### **Economic Research Service**

The USDA IT consolidation poses many of the same concerns for ERS as the IRS IT consolidation does for SOI. The motivation is less clear however. The FY10 request for consolidating the Washington area USDA facilities details a cost savings as the principle motivation: "The Department expects to avoid \$62 million of lease costs over the 15-year life of the consolidated lease." The consolidation does not seem motivated by the Enterprise Data Center memos.

In addition to concerns similar to those for SOI, ERS's joint activities with other statistical agencies could be jeopardized. ERS shares a data enclave with National Agricultural Statistics Service for the Agricultural Resource Management Survey as well as IT resources with the National Center for Health Statistics for National Health and Nutrition Examination Survey (NHANES). Given the predominating CIPSEA concerns and that ERS wouldn't have control over its IT resources, it is plausible that ERS's partners would hesitate to continue the joint activities, or at least that potential partners in future joint activities would be dissuaded.

### **Opportunity for Federal Statistics Community?**

Rahm Emanuel, the President's Chief of Staff, is frequently quoted to the effect, "don't let a crisis go to waste." Such advice could be applied to the threats posed by IT centralization on federal statistical agencies. Leadership of the departments that host statistical agencies rarely hear from the statistical agency's stakeholders and the same is true of OMB, GAO, and Congressional oversight committees. By reaching out to these audiences, we in the statistical community can introduce ourselves, educate them of the importance, functions and needs of statistical agencies, and initiate constructive dialogues between the various parties. Such dialogues are particularly important with a new administration in place that has made abundantly clear its appreciation of statistical data and the federal statistical agencies that produce these data.

Comments?: Steve Pierson welcomes your comments at [pierson@amstat.org](mailto:pierson@amstat.org).

## 2010 Census Advisory Committee Meeting Report

Ken Hodges, Director of Demography, Nielsen Claritas

The 2010 Census Advisory Committee (CAC) met May 7-8 at the Census Bureau, and the meeting was attended by APDU representative Ken Hodges and alternate representative Bill O'Hare.

The meeting included updates on 2010 Census progress, and as usual, Census Bureau officials were upbeat in their description of census preparations. Several pointed to address canvassing, in which 140,000 temporary workers with handheld computers, have pounded the pavement (as well as dirt and turf) to verify and update the address list for the 2010 census. Due to the recession, the operation has benefited from a workforce that was easier to recruit and more skilled than expected, and which worked longer hours and with less turnover. As a result, canvassing has taken much less time than planned. The CAC also heard presentations on the Census Bureau's National Processing Center, and on plans for census coverage measurement, coverage follow up, and evaluations of the effectiveness of the communications campaign.

The good news on census preparations was welcomed, but many CAC reps were concerned about the census communications campaign, and reports that a panel of advisory committee reps had been so unimpressed with preliminary ads that they issued a "vote of no confidence" in DraftFCB, the campaign's primary contractor. Census and DraftFCB officials assured that these concerns would be addressed in revised ads, but with Census and the contractor eager to move on with a tight production (filming) schedule, the CAC reps had to press hard for an opportunity for further review and input.

CAC reps also pressed Census Bureau officials on the counting of specific populations. For example, there was considerable disagreement with the plan to recode persons reporting same sex "married" to same sex "partners." We also heard renewed recommendations that prisoners be counted at their pre-incarceration residence, and that populations displaced by Hurricane Katrina be counted at the neighborhoods from which they evacuated. Such topics make for lively discussions.

The meeting concluded with a formal CAC recommendation that the Census Bureau reconsider the plan to recode same sex "married" to same sex "partners," and a request for further information on why the pre-census letter does not include "in language" instructions for requesting a census form in a language other than English.

The full meeting report is available at <http://www.apdu.org/advocacy/dcac.htm>.