## Council of Professional Associations on Federal Statistics

http://www.copafs.org

# COPAFS Brief:

## March 1, 2013 COPAFS Quarterly Meeting on “Big Data and Federal Statistics”

“Big Data,” while on the tip of everyone’s tongues, has yet to find a commonly accepted definition. Speakers at the March 1, 2013 COPAFS meeting offered explanations that highlight several characteristics of big data. Mike Horrigan (Associate Commissioner for Prices and Living Conditions at the Bureau of Labor Statistics) says that it is “non-survey data,” which includes administrative data from federal or other programs. Bob Groves (Georgetown University Provost and former Director of the Census Bureau) defines it as data sets so large that we have to go to their sites to use them. All agree that big data are dynamic – always changing and updating. Big data include data on Twitter subjects, Google searches, and E-Bay transactions, as well as Medicare records, stock exchanges, or cash register transactions. Big data sets are immensely useful in business decision making, and are portals to a better understanding of complex human behavior. But what roles do, can or shouldn’t they play in generating, supporting, or complementing official federal statistics. That was the question of the day.

***Can’t live with them; Can’t live without them.*** Bob Groves describes three stylistic options for federal statistical agencies: (1) Ignore ‘Big Data” and march traditionally onward; (2) Supplant official statistics with the more timely “big data; or (3) Blend big data and survey-based data to improve timeliness while retaining the statistical reliability and privacy protection of federal statistics. The third option is most appealing and most realistic. But blending poses big challenges since there is no way to link big data to survey response records. Mike Horrigan gives high marks to big data for its timeliness and relevance. However, when it comes to objectivity, accuracy, and lack of bias, big data are questionable. Beyond big data quality issues, there are also concerns about confidentiality, lack of transparency and replication problems. Mike Horrigan, Groves and Bill Bostic (Census Bureau Associate Director for Economic Programs) agree that assessment of and research on blending big data and federal statistics need to occur before it becomes a prevalent practice.

***Statistical Agencies are Exploring Big Data Applications***. The Census Bureau is researching big data solutions to problems of timeliness and list updating. Bostic reports that non-sample big data on foreclosures, new construction, and building permits are seen as a way to improve census address lists. It is hoped that they can supplement the building permit and construction data the Census Bureau collects already, but there are standardization and other issues to consider. The Census Bureau also is exploring the possibility of acquiring retail and service statistics from private suppliers to fill gaps in business data products. However, Bostic cautioned that using such data requires further research to address concerns related to confidentiality, definitions, and data quality. Knowing that big data are in its future, the Census Bureau has formed an agency-wide team to systematically explore options for consistent use across the agency.

At the Bureau of labor Statistics (BLS), development of the Consumer Price Index is aided by webscraping (software-assisted collection of information from across the Internet) to track product characteristics in order to adjust for quality changes in such products as televisions and cameras. Another example given by Mike Horrigan is the use of Google searches to track flu outbreaks and tweets related to job loss as predictors of unemployment claims. Horrigan described additional applications of private big data on moves from data and how big data can help BLS improve sample construction, price estimation, and imputation based estimates. He warns that the use of private sources requires agreements with the firms that provide such data, and statistical agencies need to be cautious about the confidentiality and liability implications of contracts.

***A Love/Fear Relationship with Consumers.*** John Horrigan, who directs the Media and Technology Institute at the Joint Center for Political and Economic Studies, describes big data as data generated by the activity of individuals. In his experience, big data provide a means to unleash innovation, empower consumers, personalize health care, and inform policymakers. The sharing of personal information is key to many of these innovations and, as Horrigan noted, that requires *trust*. Describing big data as a "factor of production," John Horrigan says they are being used in the operation of many businesses, often to consumers' advantage. For example, big data enable consumers to gather information on consumer goods, and then provide opinions on the perceived value of products and services. Considering what could go wrong with big data applications, Horrigan noted consumer concerns about privacy and unwanted access to personal information. But in a paradox ripe for study, he also notes evidence that people who express concern about privacy still use the Internet to supply information about themselves. Whether and how consumer concerns affect consumer behavior, and what that means for the representativeness of big data remain important questions. He concludes with the need to build in consumer protections, as well as to elevate levels of digital literacy across the population.

***Future Needs for New Institutions, Different Expertise.*** The complex interconnectivity of business, consumers, and federal government interests and constraints makes it difficult for statistical agencies to move ahead with big data. Groves, in suggesting the need for a new institution, indicates how difficult it is now for statistical agencies to contract for big data from the private sector. Both the agencies and private big data owners have concerns about liability and other consequences of confidentiality breeches beyond their control. Private big data producers may also be worried about competitors learning something about their business or using their information to develop other proprietary product. They may require unique incentives for access to their data. Groves proposes a public/private partnership institution to develop the rules and broker the arrangement of access by statistical agencies and others to private big data resources.

Finally, progress in the federal statistical system depends upon development and recruitment of a new labor force in a new specialty area Bostic refers to as "data scientists" – individuals trained in the intricacies of capturing data from the Internet, blending non-sample big data with survey-based data, and meeting the simultaneous goals of efficiency, timeliness, accuracy and reliability.

*See the March 13 meeting agenda, speakers' biographical information, and Bostic's and Mike Horrigan's PowerPoint slides at: http://www.copafs.org/meetings/march_2013.aspx*

*Special thanks to Ken Hodges of the Population Association of America whose notes on the entirety of the COPAFS March 1, 2013 meeting (available on the COPAFS website) substantially informed this brief but who is not responsible for any errors or misinterpretations made herein.*