A Dissenting View from Julia Lane

A comment on Gerald W. Gates's *How Uncertainty about Privacy and Confidentiality Is Hampering Efforts to More Effectively Use Administrative Records in Producing U.S. National Statistics*

Julia Lane*

The topic of the role of privacy and confidentiality in the use of administrative records for statistical purposes is an important one. In providing comments, I will draw extensively from my experience in working with fifty states and six federal agencies in creating the LEHD (Longitudinal Employer-Household Dynamics) program at the Census Bureau between 1997 (when I started my ASA fellowship) and 2004 (when the program became a fully-fledged Census Bureau program). However, I will also draw on my experience in building the LEED (Linked Employer-Employee Data) program at Statistics New Zealand, as well as my experience working with the data producers from many countries (both statistical agencies and administrative agencies) while in my World Bank consultancy. The final context is my experience with cyberinfrastructure initiatives within the United States, as well as in Europe.

I make three main comments here. The first is that the recommendations made by Gates are not necessarily consistent given his stated goal of advancing the use of administrative records within statistical agencies. The second is to suggest that the focus of his attention might be too narrow. Indeed, statistical agencies might well consider expanding their horizons beyond administrative data; the new opportunities presented by transaction data today are even more promising than the opportunities presented by administrative data when I first started doing research in this area 20 years ago. The third is to suggest a substantive course of action.

In my view, the recommendations in the paper are inconsistent with what is likely to succeed in practice. The focus on bureaucratic solutions, as well as the focus on the federal statistical community, does not seem to flow naturally from the discussion in the preceding sections of the paper and are certainly unlikely to succeed given my experience with data protection. Bureaucracy is no guarantee of better data protection; understanding human behavior and developing technological solutions is likely to be a much better solution. Indeed, Gates presents no evidence that changing the Privacy Act or expanding the role of OMB would enhance privacy or confidentiality; my hard-earned experience suggests that expanding bureaucracy will reduce the innovative use of administrative records, not increase it. In my experience, the legal counsel in administrative agencies was most convinced by statutory authorization and penalties that could be written into Memoranda of Understanding. There would be very little confidence that an additional layer of OMB oversight would provide sufficient data protection. This is particularly true since so many of the programs that generate administrative data are

*Program Director, Science of Science & Innovation Policy, National Science Foundation, Washington DC, mailto:jlane@nsf.gov.

run at the state level, not the federal. I am similarly skeptical that the establishment of data stewardship programs in agencies would result in anything but a negative effect. Adding yet another layer of review within each agency would create additional administrative burden, dilute lines of authority, and create additional turf battles. There is little evidence that the Census Bureau's data stewardship program was well received, or else it would have been adopted by other agencies. Finally, it is difficult to see how a public debate with the media could be effectively led by the OMB's statistical office which primarily functions to coordinate agencies and does not have a public affairs office. Technical, not bureaucratic, approaches to protecting privacy and confidentiality appear to have convinced the public that financial institutions can protect their data (as demonstrated by the widespread use of online banking). Statistical agencies can learn from best practices in other industries, rather than adding more bureaucracy.

I also disagree with the very narrow focus on administrative records to improve the data produced by US statistical agencies. The old model of statistical agencies as data producers should be revisited in the light of tremendous advances in cyberinfrastructure, both in data creation and data management. The model that Gates discusses is predicated on a vision in which statistical agencies are the producers of data (primarily survey) and that administrative data should be brought into the agencies in order to enhance this data. I would argue that the data world has changed. While it is true that the concerted efforts of a few academic researchers more than a decade ago led to many of the changes in statistical agencies lauded by Gates, transaction data—sensor, GPS, cell phone, financial, and email transactions—have substantially more potential to inform decision making in the next decades. For example, Chris Carroll and others[1] have pointed out the value of credit card data in producing data on consumer expenditures; I pointed out the value of transaction data in describing human and organizational behavior;[2] and Erik Brynjolfsson uses "nanodata" from Google search to predict housing starts faster, much more costeffectively, and more accurately than official statistics. It is no longer true that data need to be brought in and housed in a large scale data warehouse in order for statistical products to be developed and for confidentiality to be maintained.

What ARE some useful recommendations to promote the use of new kinds of data while protecting confidentiality? I would urge agencies to develop new approaches to protecting data that turn to technical and strategic, rather than bureaucratic, approaches. Some agencies have already done so; the expanded use of the NORC data enclave (`http://www.norc.org/dataenclave`) is an example that has been emulated abroad (see, for example, the recent launch by the UK's Office of National Statistics of a Secure Data Service[3]). In addition, agencies might look to the cyberinfrastructure community for new ways to manage data by the creative use of aggregation ("cloud") and federation ("grid") technologies.[4] Certainly the National Science Foundation has invested many millions of dollars in trustworthy computing[5] and other initiatives that

---

[1]`http://www.aeaweb.org/aea/2011conference/program/retrieve.php?pdfid=579`.

[2]Lane, J. (2010). "Administrative Transactional Data," Chapter in *Building on Progress: Expanding the Research Infrastructure for the Social, Economic and Behavioral Sciences.* 383–400. German Data Forum, Budrich UniPress Ltd.

could be leveraged by statistical agencies, and the White House is expanding investments in cybersecurity that can also help inform data protection policies.

I have one final comment—in the interest of keeping the record straight, I take issue with the statement on page 5: "While Internal Revenue Service (IRS) tax return information was originally intended to provide much of the source data for this project, LEHD managers could not get agreement with the IRS for adding such uses to the Tax Regulations that limit Census Bureau statistical use of tax data." In 1999, the Census Policy Office informed Census senior management that the proposed LEHD use of tax data was allowable under existing regulations, and transmitted formal correspondence from the Secretary of Commerce to the IRS Commissioner requesting such use. The formal denial of that request by the IRS Commissioner in 2000 was a direct result of IRS interactions with the Census Policy Office, not LEHD program managers. LEHD program managers then worked directly with IRS staff to resolve the legal difficulties, and IRS approval was finally achieved: see 26 CFR Part 301, posted in the Federal Register, vol. 68 No. 13 on Tuesday, January 21, 2003.

In sum, I found Gates's piece to be unconvincing. In my estimation, his recommendations would only increase the bureaucracy, inefficiency, and cost associated with acquiring administrative records for statistical purposes. The recommendations are unlikely to convince the public that the confidentiality of their data is better protected. Much can be done by looking outside the statistical system for new solutions, rather than adding more layers onto an existing bureaucratic system.

---

[3]http://securedata.data-archive.ac.uk/.
[4]Ian Foster http://anl.academia.edu/IanFoster/Papers/341141/Data_Management_and_Transfer_In_High-Performance_Computational_Grid_Environments.
[5]http://www.nsf.gov/funding/pgm_summ.jsp?pims_id=503326&org=CNS&from=home.