A Possible Data Future for the Observational Social Sciences

Robert M. Groves
US Census Bureau
November 2, 2011
5 Observations

1. The difficulties of measuring the busy, diverse, and independent American society and economy are increasing every year (that is, it costs more money to do the same things the Census Bureau has done for years).
2. The demands by American business, state, local, and community leaders for timely statistics on their populations are continually increasing.
3. New technologies are being invented almost daily that can be used to make it more convenient for the American public to participate in these efforts to inform us about the status of the country.
4. New digital data resources are being created both from Federal-state-local government programs, private sector transactions, and internet-related activities.
5. Near-term Federal government budgets are likely to be flat or declining.
Profound Conclusion

1. Higher costs
2. More demand for timely statistics
3. New technologies
4. New data resources
5. No new money

Conclusion: current practices are unsustainable
Outline

• The sample survey as a scientific instrument
• The rise of “organic” data
• Organic data and new statistical information
• A possible future
The Sample Survey

• Sometimes called the most important invention of the social sciences in 20th century
• Challenged on rigidity, blunt nature, lean measurement
• However, the principal source of social science research data
The Status of US Surveys

• Threatened coverage of frames
• Falling participation rates
• Increasing reliance on nonresponse adjustments
• For surveys with high response rate targets, inflating costs
Relative Sizes of Digital Data Production, c.1960

- Science
- Gov’t Statistics
- Journalism
- Private Sector Research
- Information Processing
Relative Sizes of Digital Data Production, 2010

Science

Gov't Statistics

Information Processing

Journalism

Private Sector Research
Changes in the Data World

• Digitization of administrative data
• Improved record-matching
• Continuous time process data
A Self-Monitoring Social and Economic Eco-System

• Organic data
  – Those produced auxiliary to processes, to record the process

• Designed Data
  – Those produced to discover the unmeasured
Examples of Organic Data

- Google searches (Google Flu)
- “Scraped” data from websites
- Tweets
- CCTV, traffic camera data
- Retail scanner data
- Credit card transaction data
- Data.gov
Common Features of Organic Data

• New data sources provide looks at interesting new phenomena
• They tend to be behaviors, not direct measures of internalized states
• The data are near real-time relative to the behaviors measured
• The data tend to be lean in variables
• They are grossly incomplete on coverage of usual target populations
Explore flu trends - United States

We've found that certain search terms are good indicators of flu activity. Google Flu Trends uses aggregated Google search data to estimate flu activity. Learn more »

National

States | Cities (Experimental)

Fight influenza

CDC urges you to take these steps to protect yourself and others from the flu:

1. Get vaccinated against flu – it's your best defense.
2. Cover your cough, wash hands often.
3. Take antiviral drugs if your doctor recommends them.

Centers for Disease Control and Prevention

Animated Flu Trends in Google Earth

Download and explore Flu Trends data in Google Earth. Need Google Earth? Download it here.
The Billion Prices Project is an academic initiative that uses prices collected from hundreds of online retailers around the world on a daily basis to conduct economic research.

This page shows our most recent research leveraging high-frequency price data, as well as the US daily inflation index (updated monthly on this page).

**BPP Geographic Coverage**

![Map showing geographic coverage of the Billion Prices Project](image-url)
Temporal patterns of happiness and information in a global social network: Hedonometrics and Twitter

Peter Sheridan Dodds,1,* Kameron Decker Harris,1 Isabel M. Kloumann,1 Catherine A. Bliss,1 and Christopher M. Danforth1,1
Approaching a Mixed-Source Future from the Survey Side

• In contrast to societies that have register systems, the US might approach a blended data world by building on top of existing surveys.
Measurement

- Construct
- Inferential Population
- Measurement
- Response
- Target Population
- Sampling Frame
- Sample
- Validity
- Measurement Error
- Processing Error
- Coverage Error
- Sampling Error
- Nonresponse Error

Representation

- Inferential Population
- Target Population
- Sampling Frame
- Sample
- Respondents
- Survey Statistic
- Edited Data
- Measurement
A Vision of the Future

- Multiple modes of data collection/acquisition
  - Internet behaviors
  - Administrative records
  - Internet self-report
  - Telephone, face-to-face, paper
- Real-time mode switch to “fill-in” missing data
- Real-time estimation
Administrative Records

Frame Variables on Sample

Ecological attributes of Sample

Imputation Engine

Nonlinked Digital Data

Estimation Engine

Collection Engine

Decision-Rule Engine

Administrative Records

Internet Self-Report

Telephone Self-Report

Face to Face Self-Report

Mail Self-Report
Administrative Records

Frame Variables on Sample

Ecological attributes of Sample

Imputation Engine

Nonlinked Digital Data

Estimation Engine

Decision-Rule Engine

Collection Engine

Internet Self-Report

Telephone Self-Report

Mail Self-Report

Face to Face Self-Report
Administrative Records

Frame Variables on Sample

Ecological attributes of Sample

Nonlinked Digital Data

Imputation Engine

Estimation Engine

Collection Engine

Decision-Rule Engine

Administrative Records

Internet Self-Report

Telephone Self-Report

Face to Face Self-Report

Mail Self-Report
Administrative Records

Frame Variables on Sample

Ecological attributes of Sample

Imputation Engine

Collection Engine

Estimation Engine

Decision-Rule Engine

Nonlinked Digital Data

Administrative Records

Internet Self-Report

Face to Face Self-Report

Mail Self-Report

Telephone Self-Report
Attributes of the Vision

• 24-hour cycles on mode-switch, imputation, estimation
• Empirical stopping rules for continued self-report efforts
• Statistical modeling to combine survey data with external, relevant other digital data
• Reduced cost, increased timeliness
Key Questions Facing Social Science

- Will “organic data” replace the designed data of surveys, given their low cost?
- Will new blends of organic data and designed data emerge?
- Will survey researchers blend or will IT masters add “designed data” to organic data?
- Whither “designed data”? 