A Vision for a 21st-Century National Data Infrastructure: Implications for Librarians

Erica L. Groshen, Cornell U. –ILR

Beyond the Numbers

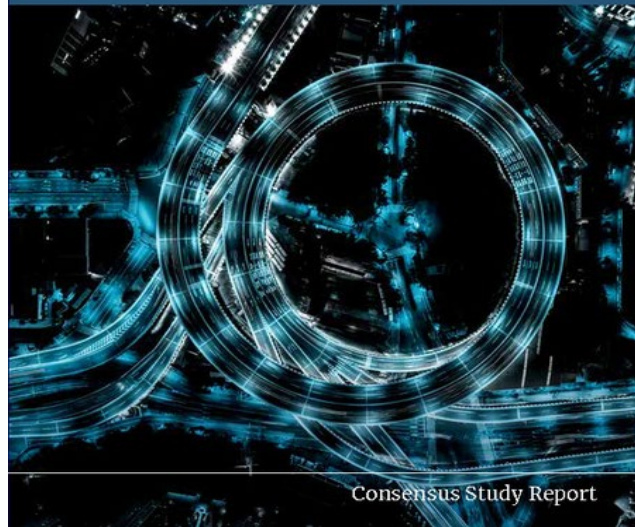Federal Reserve Bank of St. Louis

November 10, 2022

Agenda

- A new national data infrastructure?

- Vision, outcomes, and attributes

- More "blended" data

- Implications for librarians

NATIONAL ACADEMIES
Sciences
Engineering
Medicine

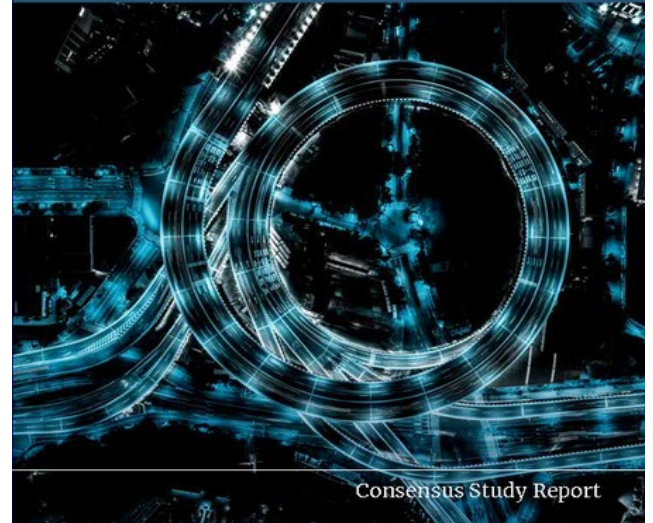Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good

Consensus Study Report

Cornell University
ILR School

"National data infrastructure"???

- Credible statistical info supports democratic society
  - Informs decisions by governments, businesses, and individuals
  - Similar to roads and bridges supporting commerce by facilitating movement of goods, services, and people
  - Examples: Unemployment, inflation, crime, and health statistics inform decisions regarding the economy, society, and peoples' lives

- Statistical agencies play critical role
  - Collect input data
  - Provide trustworthy statistics equally to all

Cornell University
ILR School

Building blocks for trustworthy statistics

Accurate. . . Getting it right

Objective. . . Free from bias

Relevant. . . Information you can use

Timely. . . Getting it out quickly

Accessible. . . Meeting you where you are

Cornell University
ILR School

Mid-20th Century data infrastructure

- Official statistics rely heavily on surveys
  - Few other sources
  - Sample designed and data collected to fit measurement needs
  - Response rates high
    - Other surveys rare
    - Public service valued
- Empirical social and economic research began with surveys

Cornell University
ILR School

Why we need a new 21st century data infrastructure

- Threat: All survey response rates falling
    - Raises costs and erodes reliability
- Opportunity: Big Data explosion
    - Digitized operations and records
    - Cheap powerful computers
    - Internet connectivity
    - Novel software (e.g., AI)

Cornell University
ILR School

Overall vision for a new 21st century data infrastructure

- What:
  - Improve quality, timeliness, granularity, and usefulness of national statistics
  - Facilitate more rigorous social and economic research
  - Support evidence-based policymaking and program evaluations

- How: Blending data from multiple sources
  - Overcomes limitations from single sources

Cornell University
ILR School

What is blended data?

- Collect data from multiple sources
  - Surveys
  - Government agencies
  - Private aggregators and companies
  - Crowdsourcing
- Combine them
  - Match/merge
  - Model
  - Use predictive analytics
- Examples: GDP, productivity, state JOLTS…

Cornell University
ILR School

Outcomes of a new data infrastructure

1. US information resources are mobilized in a coordinated manner

2. More timely, granular, and useful information

3. Enhanced research insights

4. Enhanced evidence-based policy analysis
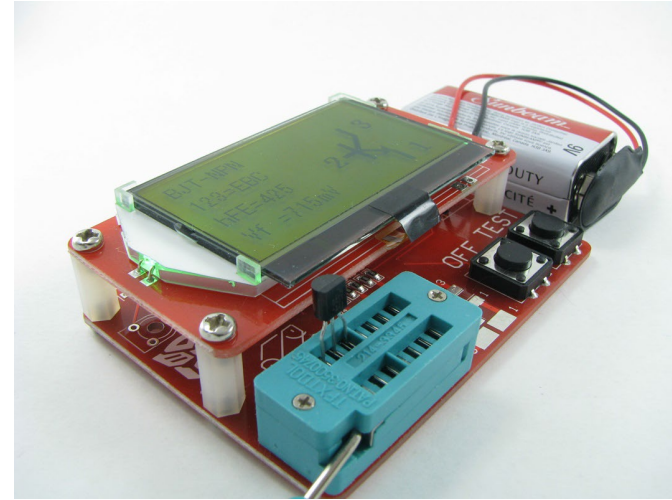
Cornell University
ILR School

Outcomes of a new data infrastructure, cont'd.

5. Data holders incentivized to share data for statistical purposes

6. Reformed legal framework enhances privacy protections

7. National data infrastructure operates in a transparent, high-trust environment

Cornell University
ILR School

Key components of new data infrastructure

1. Data assets
2. Data-related technologies
3. People with expertise to manage, use, and understand data
4. Standards governing data
5. Organizations to manage data infrastructure
6. Trust of data subjects and holders

Seven attributes of the vision for the statistical system

1. Safeguards and advanced privacy-enhancing practices to minimize possible individual harm

2. Statistical uses only, for common good information, with statistical aggregates freely shared with all

3. Mobilization of relevant digital data assets, blended in statistical aggregates, providing benefits to data holders, with societal benefits proportionate to possible costs and risks
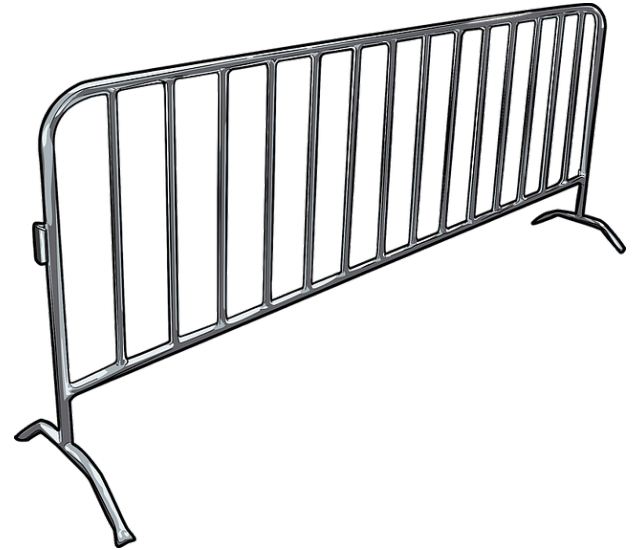
Seven attributes of the vision for the statistical system, cont'd.

4. Reformed legal authorities protecting all parties' interests
5. Governance framework and standards effectively supporting operations
6. Transparency to the public about analytical operations using the infrastructure
7. State-of-the-art practices for access, statistical, coordination, and computational activities, continuously improved to efficiently create more secure, more useful information

Cornell University
ILR School

Barriers

- Resources
- Legal impediments
- Learning curve
- Fragmented statistical system
- Privacy concerns

Progress toward the vision

1. Evidence-Based Policymaking Commission, Evidence Act, and ACDEB
   - Increase statistical agency access to federal administrative data
   - Create National Secure Data Service to facilitate blending

2. Statistical agencies, academics, and tech firms pursue initiatives

3. CNSTAT reports develop overall vision and near term steps
   - Include private sector and state and local administrative data
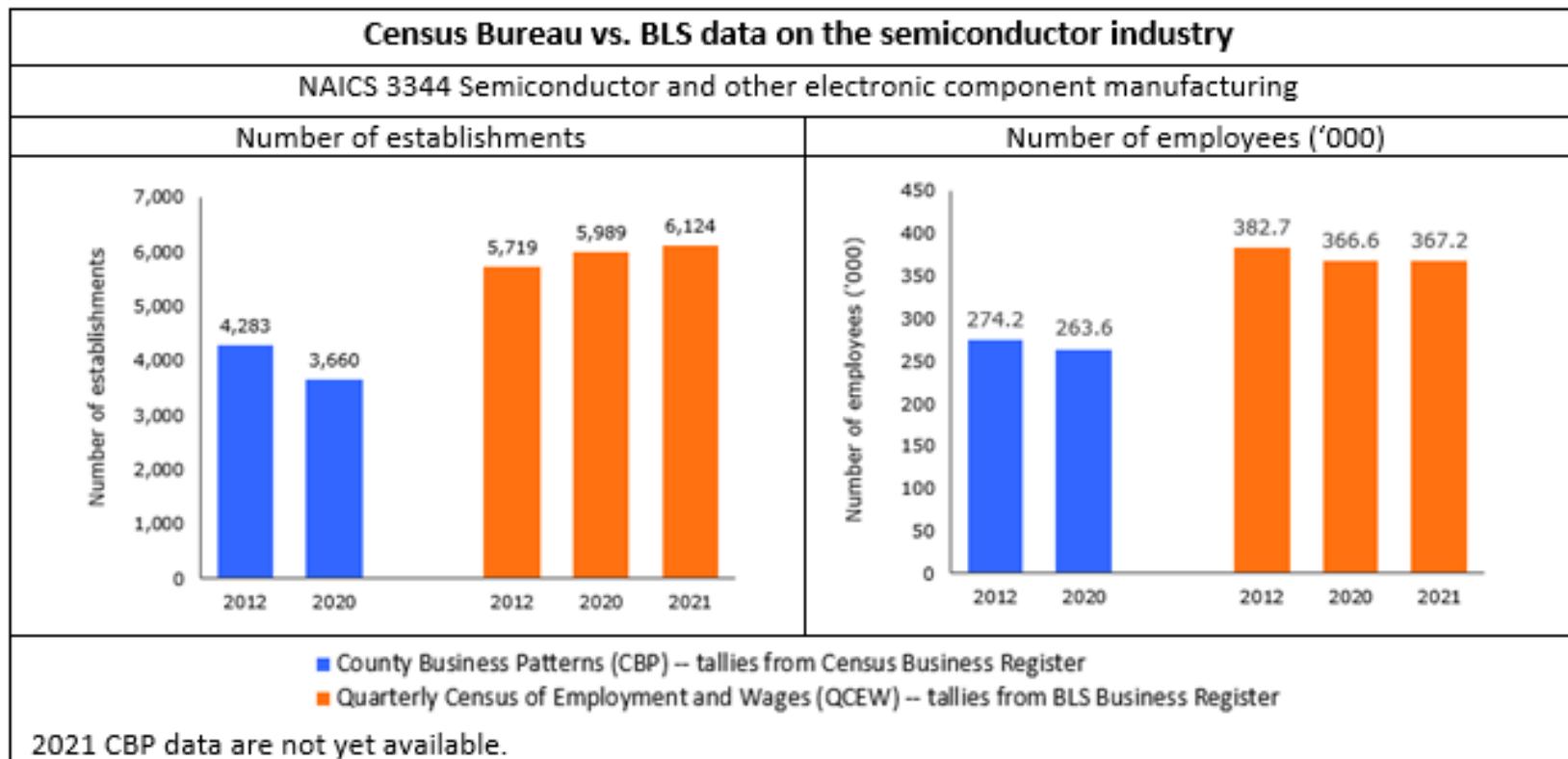
Cornell University
ILR School

Implications for librarians

- Identify data sources to match needs for blending
  - Example of difference between Census and BLS employment by industry

BLS and Census employment discrepancies

Census Bureau vs. BLS data on the semiconductor industry

NAICS 3344 Semiconductor and other electronic component manufacturing

Number of establishments — Number of employees ('000)

■ County Business Patterns (CBP) -- tallies from Census Business Register
■ Quarterly Census of Employment and Wages (QCEW) -- tallies from BLS Business Register

2021 CBP data are not yet available.

Implications for librarians

- Identify data sources to match needs for blending
- Inform users of advantages and limitations of sources

Cornell University
ILR School

How do private sector and official indicators compare?

- Government and private initiatives are complements, not competitors

Official	Private
• Transparent, known statistical properties • Access to comprehensive, sensitive government data • High survey response rates • Objective • Long history	• Proprietary methods • Speedy production • Quick innovation • Access to transactional data • Tailored to special needs

Cornell University
ILR School

Implications for librarians

- Identify data sources to match needs

- Inform users of advantages and limitations of sources

- Reinforce high-trust environment and support resources and response rates for federal surveys

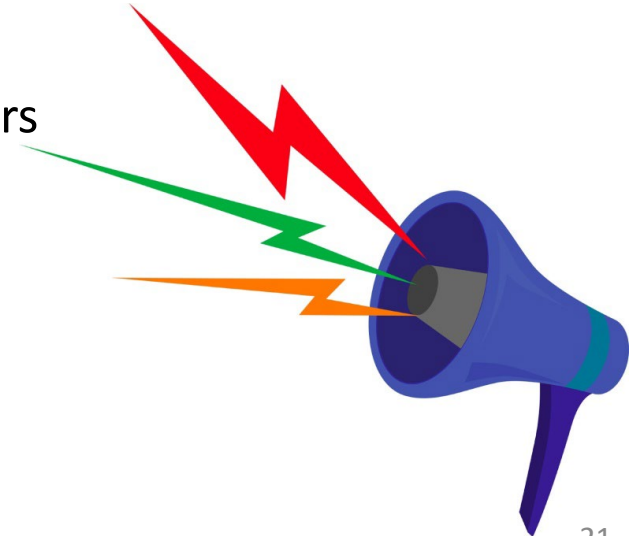**Cornell University**
ILR School

Support independence and modernization for stat agencies

- Decry political interference; independence necessary for trust
- Don't free ride—speak up for evidence and official statistics
  - Attest to trustworthiness; debunk attacks
  - Cite your sources
  - Participate in federal surveys; encourage others
  - Support funding and modernization
  - Join Friends of BLS

    www.friendsofbls.org
    LinkedIn: "Friends of BLS" group
    Twitter: @Friends_of_BLS

Cornell University
ILR School

Takeaways

- Current national data infrastructure is ill-equipped to meet 21st century data needs
- US needs a data infrastructure that blends data from multiple sources to
  - Improve national statistics
  - Facilitate more rigorous social and economic research
  - Support evidence-based policy and program evaluations
- Progress will require new partnerships, a shared vision, legislative action, consensus regarding activities that move us toward the vision...**and help from librarians**

Cornell University
ILR School

22

Thank you.

Erica L. Groshen
Cornell-ILR
erica.groshen@gmail.com

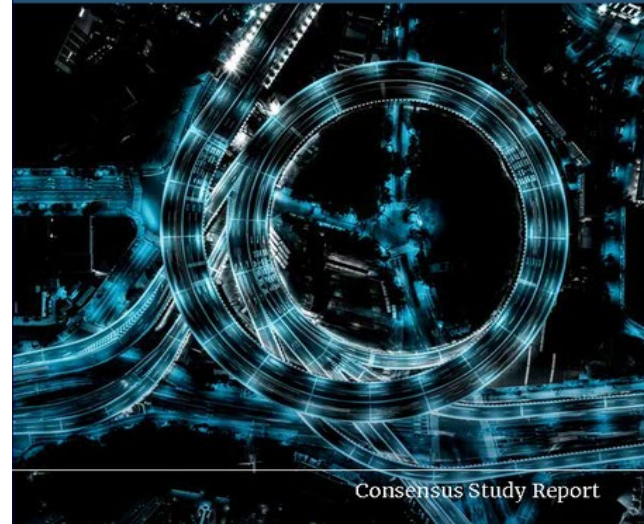https://nap.nationalacademies.org/catalog/26688

Text, Report Highlights and Policy Brief available for viewing/download. An interactive website for this report (and the two that follow) will be built. It will include report content in an accessible, digestible format as well as an FAQ section.

NATIONAL ACADEMIES Sciences Engineering Medicine

Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good

Consensus Study Report

Cornell University
ILR School