Private Sector Data: An Essential Component of a National Data Infrastructure

This issue brief describes why private sector data are an essential component of a new national data infrastructure, highlights the challenges and impediments of using private sector data for statistical purposes, and discusses the necessary attributes of private data accessed and used in a new data infrastructure as well as organizational implications. It concludes with action steps that will move the United States closer to realizing the promise of using private sector data to improve the nation's information resources.

Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good, a consensus study report by the National Academies of Sciences, Engineering, and Medicine's Committee on National Statistics, is the first in a series that seeks to build a vision for a new national data infrastructure. The report describes the need for a new data infrastructure, presents an initial vision, and describes expected outcomes and key attributes of a new data infrastructure. The report also discusses the implications of blending (combining) data from multiple sources and the organizational implications of cross-sector data access and use. The report concludes by identifying short- and medium-term activities that facilitate progress toward the vision and states that national statistics depending solely on statistical surveys are unsustainable.

The United States needs a 21st-century national data infrastructure that blends data from multiple sources to improve the quality, timeliness, granularity, and usefulness of national statistics, facilitates more rigorous social and economic research, and supports evidence-based policymaking and program evaluations. The private sector plays a critical role, both as an important data user and as a holder of key data assets. Businesses need official statistics that are more timely, detailed, and frequent to inform their forecasts, benchmarks, and market analysis and support better investment and location decisions. A new data infrastructure employing new technologies, capabilities, and partnerships provides a unique opportunity to meet these data needs by blending data from statistical surveys, government administrative agencies, and private sector enterprises to create new, more useful statistical products and facilitate research. By sharing their data assets for solely statistical purposes, businesses benefit not only from better national statistics but also through the reciprocal sharing of information and services that provide them with tangible benefits, informing their operations and activities.

CATALYSTS FOR INCLUDING PRIVATE SECTOR DATA

The current data infrastructure prevents us from realizing the promise of blended data. Three years after the signing of the Foundations for Evidence-Based Policymaking Act, laws and regulations remain major obstacles to accessing and using federal data assets for statistical purposes. Furthermore, the Evidence Act is silent regarding using private sector data for approved statistical purposes.

The Evidence Act's focus on federal data assets must be expanded and extended to include relevant data held by private sector enterprises, state and local governments, nonprofit and academic institutions, and others. While administrative data from federal, state, and local governments are an important source for statistical uses, a much larger set of data assets are being produced by private sector enterprises, including vast amounts of transactions data. These private data are not without weaknesses, but they often are more detailed, timelier, and may include information about segments of the population that are poorly represented in sample surveys.

Federal statistical agencies recognize that private sector data offer great promise and opportunities for improving national statistics. The panel's December 2021 workshops highlighted the extensive use of private sector data by federal statistical agencies. The Bureau of Economic Analysis described how the Health Satellite Account blends survey data from the Medical Expenditures Panel Survey with data from a private sector insurance company and Medicare claims data. The Census Bureau is using blended data to modernize its residential construction statistics program and to produce an experimental blended data product that provides modeled state-level estimates of retail sales. These examples demonstrate the possibilities of blending statistical, administrative, and private data, yet currently the federal statistical system lacks the robust relationships and mechanisms needed to effectively use private sector data.

Unlike federal and state administrative data sources, the statistical uses of private sector data are generally not limited by statute. No federal laws regulate businesses that buy and sell personal information. While access to private data may not be limited by statute, the protection and autonomy of data subjects are vastly different between the government and private sectors and there is significant diversity in mandated privacy protection across states.

Recent reports recommend the use of blended data to improve statistical information. The first report (2017) of the National Academies' Panel on Improving Federal Statistics for Policy and Social Science Research Using Multiple Data Sources and State-of-the-Art Estimation Methods recommended combining data assets of federal, state, and local governments with private sector sources. Beyond the National Academies, a growing number of foundations and professional associations have called for the use of blended data. For example, a 2021 Markle Foundation report also supported the use of blended data, including private sector data, to meet the needs of policymakers, researchers, and the public. The American Economic Association's Committee on Statistics, at its January 2022 conference, recognized the potential benefit (and challenges) of using high-frequency private sector data to modernize official statistics.

Other countries recognize the importance of private sector data. At the panel's December 2021 workshop, Statistics Canada, the UK Office of National Statistics, and Statistics Netherlands described their uses of private sector data and lessons learned. In addition, the European Commission has sought feedback regarding a proposal to make private sector data available for official statistics.

In the panel's judgment, private sector data are an essential component of a new national data infrastructure, but challenges and impediments must be addressed.

CHALLENGES AND IMPEDIMENTS

For all of their promise, private sector data assets have limitations. Like administrative data, private sector data are collected for purposes different from data collected

for official statistics. Private sector data often include information only about customers, may be difficult to link to statistical records, or may lack adequate documentation. Private sector data are bespoke and often costly without any inherent sustainability. Most private enterprises, like other federal, state, and local government data holders, have little incentive to share data for common good statistics. Privacy-protecting behaviors of private sector data holders are highly variable and largely unregulated, and there is little transparency and accountability for private sector data use.

Even though federal statistical agencies currently are using private sector data for statistical purposes, data acquisition, access, and use across the federal statistical system is siloed, inefficient, and largely uncoordinated. At this time, the United States has no cohesive, coordinated plan to ensure that blended private sector data become an essential and growing source of public information and research.

As one workshop participant stated, private sector data are not "gold dust" but rather like sand, abundant and requiring significant effort to make them useful for official statistics. To realize the promise, explicit values or attributes must guide the operations of a new national data infrastructure.

KEY ATTRIBUTES FOR PRIVATE SECTOR DATA USE

A new data infrastructure will provide benefits to policymakers, decisionmakers, data holders, and the public because it is guided by overarching values or attributes. Highlighted below are seven key attributes and the associated implications of including private sector data in a new national data infrastructure.

1. *Safeguards and advanced privacy-enhancing practices to minimize possible individual harm.* The new data infrastructure requires strong, uniform data protection and needs to account for the presence, interests, and rights of data subjects. Ethical treatment of data subjects demands attention to how data use affect the data subject's life (minimize harm); provides data subjects with autonomy about how their data are used; ensures data subjects understand how they benefit from expanded data uses; and confirms that the activities of the new data infrastructure respect data subjects. Currently, the autonomy of data subjects regarding data access and use differs significantly between the private and government sectors.

2. *Statistical uses only, for common good information, with statistical aggregates freely shared with all.* The new data infrastructure has the sole aim of serving the common good by producing statistical aggregates and facilitating advanced research that create shared information useful to the nation. Infrastructure operations and decisions are consistent with professional principles and practices, conform to ethical standards, are autonomous and free of political interference, and are managed to ensure privacy and confidentiality. No use should allow identification of an individual, business, or data subject. Data cannot be used for the enforcement of any law or regulation affecting any individual data subject.

3. *Mobilization of relevant national digital data assets, blended in statistical aggregates to provide benefits to data holders, with societal benefits proportionate to possible costs and risks.* The new data infrastructure should mobilize and leverage relevant data assets across sectors, but data acquisition should be limited (granularity and frequency) to what is necessary to satisfy prespecified statistical uses. The new infrastructure includes a wider variety of data holders, data subjects, data seekers, and data users than in the past, necessitating new relationships and partnerships. It is essential to demonstrate to data holders the direct, tangible benefits of expanded data sharing and societal benefits must be proportionate to data holders' costs and risks. Tangible benefits incentivize all data holders, including private sector enterprises, to share their data for common good statistical purposes.

4. *Reformed legal authorities protecting all parties' interests.* The current legal framework limits which data

assets can be shared, with whom, and for what purposes and does not satisfy the demands of a modern data infrastructure. Changes are needed to incentivize beneficial sharing, provide uniform privacy protection, preserve confidentiality, ensure autonomy, and prevent abuse of data-sharing arrangements. Thus, legislative, and regulatory reform is needed.

5. *Governance frameworks and standards effectively supporting operations.* In the panel's judgment, data governance is crucial and must address the blending of multiple data sources, the need to respect the rights of data subjects, reflect changing notions of privacy and consent, and recognize that the risks and benefits of data sharing may vary depending on data use context and purposes. The use of private sector data raises additional ethical issues warranting further study. Establishing an effective data governance framework and establishing standards require the active engagement of data holders, data subjects, and responsible data infrastructure entities.

6. *Being transparent to the public about analytical operations.* At any time, the public, data holders, and data subjects should be able to know how their data are used, by whom, for what purposes, and to what societal benefits. Multistakeholder participation is needed, and data holders and data subjects must be given a voice in decisions that affect them. In the panel's opinion, transparency must be a stated requisite in the legal basis of a new data infrastructure as well as part of the data governance framework.

7. *State-of-the-art practices for access, statistical, coordination, and computational activities; continuously improved to efficiently create increasingly secure and useful information.* The technical aspects of a new data infrastructure are likely to be highly dynamic and must continuously innovate. Partnering with diverse data holders, including private sector enterprises, will be essential in incorporating new technologies, methods, and capabilities. An exemplar of such a partnership is the pilot project investigating the feasibility of accessing and processing data at the company's site to produce statistical aggregates for official statistics.

MULTIPLE ORGANIZATIONAL STRUCTURES CAN SUPPORT A NEW DATA INFRASTRUCTURE

The current organizational structure of the federal statistical system did not anticipate a new data infrastructure that would tap relevant data assets from all sectors, including the private sector. The panel believes it is essential to consider the implications of including private sector data on data infrastructure organizational options, organization types, and organization placement. To identify the best option for the United States, the panel suggests beginning a widespread dialogue among the many infrastructure stakeholders, including private sector representatives.

ACTION STEPS RELATED TO PRIVATE SECTOR DATA

The Evidence Act is silent regarding the statistical uses of private sector data. In contrast, the panel sees the merit of blending private sector data with relevant statistical, administrative, nonprofit, and other data sources to produce more granular, timely, and relevant information about the economy and society. All data assets have weaknesses, but careful blending of data from multiple, complementary sources can emolliate the weaknesses of any single data source. The following steps move the United States closer to realizing the promise of using private sector data to strengthen, improve, and transform the way the nation uses and benefits from better information.

1. Begin a dialogue with the private sector to better understand the incentives and disincentives affecting their willingness to share information for statistical purposes and imbue both private sector data holders and other data infrastructure stakeholders with a mutual understanding of the value of expanded data sharing.

2. Identify legal options that would incentivize private sector enterprises to share data for statistical purposes.

3. Discuss criteria and key characteristics of data assets to include in a new national data infrastructure. The report included a suggested set of criteria/key characteristics for consideration:

    - Assets are fit for intended statistical purposes.
    - Acquisition of information is limited and minimized to satisfy pre-specified purpose.
    - Access and use respect data holders' and data subjects' interests and privacy.
    - Uses should prioritize easily acquired data assets that provide tangible benefits.
    - Available, usable metadata facilitates statistical uses.

4. Monitor ongoing "data-connecting" pilot projects with the private sector that are exploring the feasibility of collecting data at the data holder's site. The variety, size, and proprietary nature of relevant data sets preclude moving the data set from private enterprises to a statistical agency or entity. Understanding the objectives, requirements, needed expertise, challenges, and outcomes of these pilots inform needed future data infrastructure capabilities.

5. Use existing examples of blended statistics produced by statistical agencies as case studies for documenting the costs, risks, and benefits of expanded sharing and blended data.

6. Assess implications of including private sector data on the organization, responsibilities, and functions of a future National Secure Data Service (NSDS). (The recently enacted CHIPS and Science Act included a provision to establish an NSDS demonstration project at the National Science Foundation.)

7. Sponsor a bipartisan, multisector dialogue on identifying options for governing private sector data use for national statistical purposes.

PANEL ON TOWARD A 21ST CENTURY NATIONAL DATA INFRASTRUCTURE: MOBILIZING INFORMATION FOR THE COMMON GOOD **Robert M. Groves** (*Chair*), Office of the Provost, Georgetown University; **danah boyd,** Data & Society; **Anne C. Case,** School of Public and International Affairs, Princeton University; **Janet M. Currie,** Center for Health and Wellbeing, Princeton University; **Erica L. Groshen,** Cornell University School of Industrial and Labor Relations; Upjohn Institute for Employment Research; **Margaret C. Levenstein,** Inter-university Consortium for Political and Social Research, University of Michigan; **Ted McCann,** American Idea Foundation; **C. Matthew Snipp,** Department of Sociology, Stanford University; **Patricia Solís,** School of Geographical Sciences and Urban Planning, Arizona State University

STUDY STAFF **Thomas Mesenbourg** (*Study Director*); **Michael Siri** (*Associate Program Officer*); **Katelyn Stenger** (*Associate Program Officer*); **Joshua Lang** (*Senior Program Assistant*)

FOR MORE INFORMATION
This issue brief was prepared by the Committee on National Statistics based on the report *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good* (2023). The study was sponsored by the National Science Foundation. Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project. The Consensus Study Report is available from the National Academies Press, (800) 624-6242, or https://www.nap.edu/catalog/26688.

Division of Behavioral and Social Sciences and Education

NATIONAL ACADEMIES *Sciences Engineering Medicine*