NATIONAL ACADEMIES *Sciences Engineering Medicine*

Consensus Study Report
Policy Brief

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

Brief for Policymakers

This consensus study by the National Academies of Sciences, Engineering, and Medicine's Committee on National Statistics is the first in a series that is intended to build a vision for a new national data infrastructure. This 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 data from multiple sources as well as 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. Future reports will explore blending multiple data sources, data equity, technology and tools, and other data infrastructure–related challenges and opportunities. The authoring panel had expertise in statistics, sociology, demography, survey methodology, economics, public policy, geography and geo-spatial data, and the intersection of data, technology, and society.

This policy brief presents 10 key takeaways and 5 action steps for national, state, and local leaders who shape policies, priorities, and funding related to strengthening, improving, and transforming the ways the United States uses and benefits from better informational resources.

10 TAKEAWAYS

1. **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.** To meet the demands for credible, trustworthy, and timely statistical information, the United States should mobilize the nation's ever-expanding data assets. It should blend data from multiple sources to improve national statistics, promote social and economic research, and support evidence-based policymaking while preserving privacy and protecting confidentiality.

   The current data infrastructure prevents us from realizing the promise of blended data. Data acquisition, access, and use are siloed, inefficient, and largely uncoordinated. Two and a half 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. 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.

2. **It is ethically necessary and technically possible to preserve privacy and fulfill confidentiality pledges regarding data while simultaneously expanding the statistical uses of diverse data sources.** A new data infrastructure must respect data subjects' rights and interests, minimize any possible individual harm, and actively engage stakeholders and communities. These are ethical necessities that contribute to the legitimacy and trust building required for a successful new data infrastructure. Privacy and security are key features of a trusted system; data are safeguarded and secured, while privacy is preserved, and confidentiality is protected. A strong set of safeguards ensures data will not be used to harm any individual or data subject. A combination of modern cybersecurity, encryption, and secure access protocols can provide greatly enhanced data security, while new privacy-enhancing technologies are essential for protecting confidentiality. With strong protections in place, the societal benefits of statistical information critical to its welfare will not come at the price of increased threats to privacy and confidentiality.

3. **Data from federal, state, tribal, territory, and local governments; the private sector; nonprofits and academic institutions; and crowdsourced and citizen-science data holders are crucial components of a 21st-century national data infrastructure.** A new data infrastructure must mobilize the nation's relevant data assets expanding the scope of the Evidence Act and the 2017 Commission on Evidence-Based Policymaking report. Acquisition of data assets, data elements, data granularity, and frequency should be limited only to the information needed to satisfy the proposed statistical use. Data assets will not be stored or retained in a data warehouse; a new data entity will provide a shared service that permits authorized users to conduct temporary linkages for exclusively statistical purposes.

4. **Data sharing is incentivized when all data holders enjoy tangible benefits valuable to their missions, and when societal benefits are proportionate to possible costs and risks.** Data holders will be more likely to share their data assets for approved statistical purposes if they understand the tangible benefits that expanded sharing provides both to themselves and to society. The benefits to data holders should go beyond improved statistics to reciprocal "information sharing," where tailored insights extracted from data assets and analysis flow back to data holders, informing their activities and operations. A new data infrastructure must ensure that the societal benefits justify the costs and risks of data sharing.

5. **Legal and regulatory changes are necessary to achieve the full promise of a 21st-century national data infrastructure.** The current legal framework that limits which data assets can be shared, with whom, and for what purposes does not satisfy the demands of a modern data infrastructure. The current framework prohibits beneficial sharing and lacks consistent requirements to preserve privacy, protect confidentiality, ensure autonomy, and prevent abuse of data-sharing arrangements. Thus, legislative reform is needed.

6. **Effective data governance is critical and should be inclusive and accountable; governance policies and standards facilitating interoperability include key stakeholders and oversight bodies.** The data governance framework includes guiding principles, authorities, structures, directives, policies, and procedures for acquiring, accessing, using, managing, and protecting data assets. Standards facilitate the acquisition, use, and organization of data and interoperable exchange across sectors. A new data infrastructure needs to adopt existing data standards, when appropriate, and to promote the creation of new standards when necessary. Data governance requires the active engagement of data subjects, holders, users, responsible infrastructure organizations, and oversight bodies.

7. **Trust in a new data infrastructure requires transparency of operations and accountability of the operators, with ongoing engagement of stakeholders.** At any time, the public, data holders, and data subjects should be able to understand how their data are used, by whom, for what purposes, and to what societal benefit. Transparency is a prerequisite for accountability, enabling the public to express concerns, seek redress, and oversee compliance with a new data infrastructure's stated mission. Transparency is also a prerequisite for public trust. With trust and transparent procedures, the credibility of the statistical information produced through a new data infrastructure will be enhanced.

8. **The operations of a new data infrastructure would benefit from the inclusion of continually evolving practices, methods, technologies, and skills to ethically leverage new technologies and advanced methods.** The technical aspects of a new data infrastructure are likely to be highly dynamic. New developments in remote access, cybersecurity, cryptography, and computational approaches will emerge and operations within a new data infrastructure must continually innovate and improve. The dynamic nature of all of these features demands continuous refreshing of the skill mix of the infrastructure's operational staff.

9. **Many organizational structures can support a new data infrastructure.** The current organizational structure of the Federal statistical system did not anticipate a proposed 21st-century data infrastructure which should tap relevant assets from all sectors, including private sector enterprises raising new challenges. Alternative ideas have been promulgated for organizing statistical operations within a new data infrastructure. The key new data entity (or set of entities) is not a data warehouse, but rather a computational service for linking data files in diverse ways, to produce blended statistics. Several potential organizational models can be envisioned for this new entity: within the federal government, outside the federal government, or as a new public–private partnership. To identify the best option for the United States, the panel suggests that a widespread dialogue should begin, involving the many stakeholders of a new data infrastructure.

10. **In building a 21st-century national data infrastructure, early success may come first from integrating data that are relatively easily available, demonstrating the utility of improved statistical information of national importance, and constructing effective partnerships for necessary legal change.** There are many ways to achieve the vision of a 21st-century national data infrastructure, and it is too early to identify each step necessary to achieve it. The committee suggests leveraging existing initiatives, both domestically and internationally, and looking for early examples of success. This will require forging new partnerships with data holders, key data infrastructure entities, and interested stakeholders. The report includes 40 short-term and 25 medium-term activities that can help discern the best ways for the United States to progress toward the committee's full vision.

5 STEPS TO TAKE NOW

1. Assess implications of cross-sector data sharing 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 a NSDS demonstration project at the National Science Foundation.)

2. Implement Evidence Act regulations and rule making (i.e., statistical agency access to federal program and administrative data, access to restricted Confidential Information Protection and Statistical Efficiency Act data assets, and data sensitivity).

3. Identify legislation and regulatory priorities regarding Commission on Evidence-Based Policymaking state-related recommendations (access to state income and earnings data, National Directory of New Hires); catalogue regulatory features that impact sharing for statistical purposes.

4. Develop legislative strategy for data synchronization bill (permits the U.S. Census Bureau to share limited business tax data with the Bureau of Labor Statistics and the Bureau of Economic Analysis).

5. Identify legal options that would incentivize data holders to share data for statistical purposes.

THE PROMISE OF A NEW DATA INFRASTRUCTURE

Current Projects Demonstrate the Benefits of Blending Diverse Data Sources

The recent modernization of the U.S. Census Bureau's residential construction statistics program was possible only by blending multiple data sources. Rather than collecting residential housing permits, which would include data from 9,000 permit-issuing organizations, the U.S. Census Bureau now receives data from third-party sources and introduced a small cutoff sample to supplement the third-party data. Satellite imagery (using geolocation and georeferenced data dimensions), rather than data collected by telephone interviewers, is used to identify the start of construction. This approach was adopted after collaborating with Statistics Canada. Effective with January 2022 release, the U.S. Census Bureau now publishes more granular statistics, including construction permit statistics for every jurisdiction in the United States, rather than just for states.

New Data Infrastructure Capabilities

Pilot research projects have led to a new vision of how data might be shared between companies and statistical agencies. Because company data holdings are so large, they cannot be transported from the data holder to the data-seeking statistical agency, active work is focusing on how data can be usefully and safely accessed and processed at the data holder's facility. Software residing in the data holder's domain acts on the data holder's existing data to produce aggregates that serve as the statistical building blocks that a federal statistical agency might use directly or blend with other survey or census data. Developing interoperable sharing mechanisms across distributed data sources will be key to a new data infrastructure vision.

NATIONAL ACADEMIES Sciences Engineering Medicine
Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good
Consensus Study Report

To read the full report, please visit https://www.nationalacademies.org/our-work/toward-a-vision-for-a-new-data-infrastructure-for-federal-statistics-and-social-and-economic-research-in-the-21st-century

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** (*Senior Program Officer*); **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 Consensus Study Report *Toward a 21st Century National Data Infrastructure: Mobilizing Information for the Common Good* (2022). 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. Copies of the Consensus Study Report are 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*