NATIONAL ACADEMIES Sciences Engineering Medicine

Consensus Study Report
Highlights

Toward a 21st Century National Data Infrastructure

Mobilizing Information for the Common Good

Credible statistical information is foundational to the functioning of democratic societies. Just as bridges and highways facilitate the transportation necessary for commerce, the national data infrastructure informs decisions by governments, business enterprises, and individuals.

Historically, the U.S. national data infrastructure has relied on the operations of the federal statistical system and the data assets that it holds. Statistics created from surveys were essential to what we know about the well-being of the society and economy. They also created an infrastructure for vital empirical social and economic sciences research. Like other infrastructure, we can easily take these essential statistics for granted. Only when they are threatened, do we recognize the need to protect them.

Today, paradoxically, national statistics face both grave threats but, also, a historic opportunity. Declining survey participation poses a severe threat to the quality of statistical information. Yet, at the same time, the country produces unprecedented amounts of digital data about the activities of individuals and businesses.

To address these threats and explore the opportunities, the National Academies of Sciences, Engineering, and Medicine appointed a consensus panel with expertise in statistics, sociology, demography, survey methodology, economics, public policy, and the intersection of data, technology, and society to help guide the development of a vision for a new data infrastructure for federal statistics and social and economic research in the 21st century. This report is the first of three reports that were funded by the National Science Foundation to explore the many issues surrounding a new data infrastructure. This report describes the need for a new national data infrastructure, presents an initial vision, and describes expected outcomes and key attributes of a new national 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 full vision.

INITIAL VISION

The panel's vision for a new data infrastructure assumes that statistical agencies and other approved users (federal, state, tribal, territory, local government employees and researchers) will access and use data assets for solely statistical purposes that are relevant to the nation's information and research needs.

The United States needs a new 21st century 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. (Conclusion 2-1)

The panel identified key components, services, and needed capacities for a new data infrastructure. They include: (1) data assets; (2) the technologies used to discover, access, share, process, use, analyze, manage, store, preserve, protect, and secure those assets; (3) the people, capacity, and expertise needed to manage, use, interpret, and understand data; (4) the guidance, standards, policies, and rules that govern data access, use, and protection; (5) the organizations and entities that manage, oversee, and govern the data infrastructure; and (6) the communities and data subjects whose data is shared and used for statistical purposes and may be impacted by decisions that are made using those data assets.

The panel believes a new data infrastructure that includes the components and items above will provide benefits to policymakers, decision-makers, and the public as shown in Box 1:

These outcomes are possible because a new data infrastructure is guided by overarching values or

BOX 1
OUTCOMES OF A NEW DATA INFRASTRUCTURE

1. The nation's information resources are strengthened by blending data from multiple data sources and employing new methods, designs, capabilities, technology, and tools.
2. Critical information for decision-makers is made more timely, granular, and useful by expanding access to data from a broader set of data holders.
3. Researchers illuminate issues of national importance by accessing existing national data assets.
4. Enhanced evidence-based policy analysis informs federal, state, tribal, territory, and local governments.
5. Data holders are incentivized to share data for statistical purposes by providing them with tangible benefits that inform and improve their operations and activities.
6. A reformed legal and regulatory framework undergirds protections for both participants and authorities, permitting increased use of existing data resources for common good statistical information.
7. The national data infrastructure operates in a high trust environment, characterized by transparency, that balances expanded data use with strengthened privacy preservation and confidentiality protection, data security, legal compliance, and responsible and ethical data use.

attributes. The seven key attributes the panel identified are described below. Following each attribute is a description of selected key short-term actions that could help discern the best ways for the U.S. to progress towards the panel's full vision for a new data infrastructure.

SEVEN ATTRIBUTES OF THE VISION

1. Safeguards and advanced privacy-enhancing practices to minimize possible individual harm.
The social benefits of statistical information need not come at the price of increased threats to individuals' privacy and confidentiality. Any harm to individuals from building and operating this infrastructure should be minimized. New technologies and strong regulations can strengthen safeguards for individuals.

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. (Conclusion 3-1)

Key Actions: (1) Establish mechanisms to engage stakeholders (including data subjects, data holders and other responsible organizations) regarding data safeguard prerequisites for building trust; (2) develop strategy for ensuring key data safeguards are communicated effectively and transparently; (3) establish technical specifications for for privacy-preserving and confidentiality-protecting designs.

2. Statistical uses only, for common good information, with statistical aggregates freely shared with all.
Data resources produce non-identifiable aggregates, estimates, and statistics to create useful information for society and decision-makers without harming individuals. Data infrastructure operations and decisions are consistent with professional principles and practices, ethical standards, conducted by organizations free of political interference, and managed to ensure privacy and security. Confidential data cannot be used for enforcement of any laws or regulations affecting any individual data subject.

Key Actions: (1) Use pilots to promote wider understanding of "statistical uses"; (2) convene stakeholders to determine how best to describe new statistical products and distinguish them from privacy-threatening initiatives.

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.
A new data infrastructure should have access to relevant existing national digital assets for the creation of essential aggregates. The infrastructure should mobilize and leverage relevant data assets across different sectors.

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 the 21st century data infrastructure. (Conclusion 4-1)

This infrastructure includes a wider variety of data holders, data subjects, data seekers, and data users than in the past. Thus, the need to demonstrate the benefits of expanded data sharing becomes even more important and a prerequisite for support.

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. (Conclusion 3-2)

Key Actions: (1) Seek researcher input regarding Standard Application Process implementation as an access tool; (2) monitor activities of Interagency Council on Statistical Policy working group on private sector data; (3) monitor "data-connecting" pilots collecting data at the data holder's site; (4) publish criteria for prioritizing new data assets.

4. Reformed legal authorities protecting all parties' interests.
Federal statistical agencies have the right, under the Evidence Act, to use federal program data for statistical uses only, *unless directly prohibited by law.* However, there are many laws and regulations that *do* prohibit

federal statistical agencies from utilizing existing data for statistical purposes. The panel assumes the legislative and regulatory recommendations stemming from the Evidence Act will be initiated, but more needs to be done to bolster data safeguards and broaden data access.

Legal and regulatory changes are necessary to achieve the full promise of the 21st century national data infrastructure. (Conclusion 3-3)

Key Actions: (1) Legislation establishing the the design, authorities, and funding for the NSDS; (2) implement Evidence Act regulations and rule making; (3) identify legislation/regulatory priorities regarding CEP state-related recommendations; (4) develop legislative strategy for a bill that permits the Census Bureau to share limited business tax data with the Bureau of Labor Statistics and the Bureau of Economic Analysis.

5. Governance framework and standards effectively supporting operations.

The "data governance" framework includes guiding principles, authorities, structures, and directives for acquiring, accessing, using, managing, and protecting data assets. Data governance involves active stakeholder engagement, oversight protocols, open and transparent communications, and accountability. Standards in data definitions and access protocols are critical to provide interoperability across partners and sectors.

Effective data governance is critical and should be inclusive and accountable; governance policies and standards facilitating interoperability include key stakeholders and oversight bodies. (Conclusion 3-4)

Key Actions: (1) Convene potential data-sharing organizations; (2) document current practices in data access; (3) document existing ways data are curated, protected, and preserved; (4) identify priorities for standards development.

6. Transparency to the public about analytical operations using the infrastructure.

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 benefit.

Trust in a new data infrastructure requires transparency of operations and accountability of the operators, with ongoing engagement of stakeholders. (Conclusion 3-5)

Transparency enables the public to express concerns, seek redress and oversee compliance with the stated mission of the infrastructure. Transparency is thus a prerequisite for public trust in the infrastructure and associated statistical products.

Key Actions: (1) Identify communication priorities regarding transparency; (2) sponsor public discussion regarding alternative oversight structures to achieve transparency; (3) engage stakeholders to evaluate alternative approaches.

7. State-of-the-art practices for access, statistical, coordination, and computational activities, continuously improved to efficiently create more secure, more useful information.

New developments in remote access, cybersecurity, cryptography, and computational approaches are constantly emerging. Thus, the operations inside a data infrastructure must continually innovate and improve. Similarly, a data infrastructure must have the talent to blend data together for more insightful research and statistical products. The acquisition, access and use of diverse data assets held by different organizations in different sectors will involve new partners with divergent experiences and expertise. This dynamism demands continuous refreshing of the skill mix of data infrastructure staff.

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. (Conclusion 3-6)

Key Actions: (1) Exchange knowledge about needed staff skillsets to support new operations of infrastructure; (2)

build communities of practice to catalyze the technical skills base

MULTIPLE ORGANIZATIONAL STRUCTURES CAN SUPPORT A NEW DATA INFRASTRUCTURE

There have been alternative ideas proffered for how to organize statistical operations within a new data infrastructure. The key new entity (or set of entities) needed is not a data warehouse, but rather a computational resource for linking data files in diverse ways to produce blended statistics. The report catalogues several organizational models for this new entity: inside the federal government, outside the federal government, or in a new public-private partnership. To identify the best option for the United States, the panel suggests the beginning of widespread dialogue involving the many stakeholders of a data infrastructure.

Key Actions: (1) Monitor America's DataHub Consortium capabilities regarding regional and sectoral data sharing; (2) clarify data infrastructure roles and responsibilities; (3) identify NSDS provided services and capabilities; (4) clarify federal statistical research data centers services and responsibilities; (5) sponsor bipartisan, multisector dialogues on how best to govern private sector data use for national statistical purposes.

BUILDING A NEW DATA INFRASTRUCTURE

This report is written at a time of unusual change. There are ongoing research and statistical agency initiatives blending data from multiple sources. These will undoubtedly inform future activities. Some new laws and regulations have been enacted, but more needs to be done. Indeed, some of the initial building blocks of the 21st century data infrastructure are being constructed, but without a coordinated vision.

There are many ways to achieve the vision, and it is too early to identify each of the steps that should be mounted to achieve it. The panel suggests an approach that leverages the many ongoing initiatives, both domestically and internationally, and looks for early successes. This will require forging new partnerships with data holders, key data infrastructure entities, and interested stakeholders. The panel identifies a set of short-term and medium-term activities associated with each of the seven key attributes and possible organizational models. These tasks provide a possible roadmap that permits progress toward the full vision.

In building a 21st century data infrastructure, early success may come first from integrating data that are readily easily available, demonstrating the utility of improved statistical information of national importance, and constructing effective partnerships for necessary legal change. (Conclusion 5-1)

The United States is capable of building this new national data infrastructure. With appropriate design of operations using the data, the American public and decision-makers would enjoy more timely, granular, and accurate information and a more robust research infrastructure.

FOR MORE INFORMATION
This Consensus Study Report Highlights was prepared by the Committee on National Statistics and the Committee on Population based on the Consensus Study Report *Toward a Vision for a New Data Infrastructure for Federal Statistics and Social and Economic Research in the 21st Century* (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*