ACDEB
Advisory Committee on Data
for Evidence Building

# Advisory Committee on Data for Evidence Building:

## Year 2 Report Supplemental Information

October 14, 2022

# Contents

# 1. ACDEB Use Case Reports

During its second year, the Committee explored several use cases as a mechanism for gathering evidence and developing findings to support its recommendations. As part of this process, members considered each of the items below and addressed them, as applicable.

## Focus

- Review current and evolving approaches to accessing, linking, and analyzing data across the federal, state, and local levels with a consideration of how decisionmaking could be enhanced and facilitated.
- Investigate improvements for the current evidence-building ecosystem and weigh the possibilities of a National Secure Data Service.

## Rationale

- Evidence Act with an emphasis on Title III (that is, the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) 2018)
- Evidence Commission recommendations
- ACDEB Year 1 report (full Committee and subcommittee recommendations)
- Interagency Council on Statistical Policy (ICSP) and Office of Management and Budget (OMB) workstreams

## Key Points

- The value of data access, linking, and analysis for evidence building for data providers and users
- Potential data sources: Federal, state, local, and private sources; both official statistical products and administrative data
- Barriers, challenges, and gaps; for example:
  - Legal and regulatory barriers;
  - Cultural resistance;
  - Differences in metadata, data quality, and systems interoperability; and
  - Resource and capacity issues.

- Lessons learned, possible solutions, and opportunities; for example:
    - Coordination within and across levels of government and with the private sector;
    - Data standards, consistency, and interoperability; data quality improvements; and data discoverability and transparency;
    - Role of the academic community, communities of practice, training, and resource sharing; and
    - Technologies, tools, and advanced analytical methods
- Privacy and confidentiality: Privacy-quality tradeoff and privacy protections needed to comply with legal and ethical requirements

Table S1 summarizes each use case. This section also presents the detailed use case reports. These reports were finalized around July 2022 and may not reflect the current status of the models and examples presented.

**Table S1. ACDEB Use Cases**

| Project Title and Purpose | Models of focus | Champions | Originating Subcommittee |
|---|---|---|---|
| **Using Administrative Data to Track Project Impact**<br>Purpose: To explore the use of administrative data to track project outcomes, especially at the local level. | • Implementation of American Rescue Program and Infrastructure Investment and Jobs Act | • Christine Heflin | Legislation and Regulations |
| **Education and Workforce**<br>Purpose: To review current and evolving approaches to accessing, linking, and analyzing education and workforce data across federal, state, and local levels with a consideration of how decisionmaking could be enhanced and facilitated. This includes investigating improvements for the current evidence-building ecosystem and weighing the possibilities for the NSDS. | • Department of Education Ability to Benefit program<br>• The Midwest Collaborative | • Shawn Davis<br>• Gregory Fortelny | Governance, Transparency, and Accountability |
| **Health**<br>Purpose: To review current and evolving approaches to accessing, linking, and analyzing health data across federal, state, and local levels with a consideration of how decisionmaking could be enhanced and facilitated. This includes investigating improvements for the current evidence-building ecosystem and weighing the possibilities for the NSDS. | • National Center for Health Statistics National Vital Statistics System modernization efforts | • Brian Moyer<br>• Kimberly Murnieks | Governance, Transparency, and Accountability |
| **Labor Market Activity**<br>Purpose: To review current and evolving approaches to accessing, linking, and analyzing labor market activity data with an emphasis on (1) improving local labor market statistics, (2) leveraging unemployment insurance data for better evaluation/research/continuous improvement, and (3) producing new national statistics. This includes investigating improvements for the current evidence-building ecosystem and weighing the possibilities for the NSDS. | • Department of Labor Unemployment Insurance Equity Data Partnerships<br>• Regional state collaboratives | • Julia Lane<br>• Christina Yancey | Governance, Transparency, and Accountability |
| **Health Quality and Human Health**<br>Purpose: To explore approaches for better accessing, linking, and analyzing data for environmental condition and human health. | • Per- and polyfluoroalkyl substances contamination in drinking water | • Richard Allen | Government Data for Evidence Building |
| **Data Inventories and Metadata**<br>Purpose: To review current and evolving approaches to creating data inventories and managing metadata, with a consideration of how such tools could enhance and facilitate evidence-based decisionmaking at federal, state, and local levels. | • U.S. Chamber of Commerce Foundation Jobs and Employment Data Exchange | • Elisabeth Kovacs | Government Data for Evidence Building |

## Using Administrative Data to Track Project Impact

### Champion: Christine Heflin, *ACDEB Member*

### Project Description

This use case explores current and evolving approaches to using administrative data to track project impact. This includes investigating improvements for the current evidence-building ecosystem and weighing the possibilities of a National Secure Data Service (NSDS).

The use case considers how quarterly data on jobs and employment collected by federal agencies could be used to track economic recovery under the American Rescue Plan (ARP) and the Infrastructure Investment and Jobs Act (IIJA).

### *Rationale*

This use case highlights key aspects of the Evidence Act's vision for evidence-based decisionmaking. In addition, the use case advances the recommendations of the full Committee as well as of the Legislation and Regulations and Government Data for Evidence Building subcommittees, as outlined in the Year 1 report. The findings from the use case inform ACDEB's Year 2 report.

More specifically, the use case explores the following:

- Tangible information on the public benefit of making administrative data available for managing operations,
- The benefits and barriers to using various data sets for tracking economic recovery and project impact,
- The Five Safes framework,
- Privacy protections needed to comply with legal and ethical requirements, and
- Actions that would permit the NSDS to support agencies in using administrative data to monitor small economic area conditions and project impact.

### *Importance and Value Proposition*

Current and reliable information on local recovery and project effectiveness would be invaluable in identifying areas that are not responding to program interventions (that is, government assistance) and that need more assistance or a different approach. When areas are "left behind," trust in government, equity, and social cohesion decline. Being able to identify what interventions are most effective would increase the return on investment in government programming at all levels.

### *Background*

Unprecedented amounts of funding are being expended as part of the ARP and the IIJA. Federal, state, and local agencies want information on program impact and what types of projects are most impactful under what circumstances. For most of these programs, expected impacts include sustaining and expanding job opportunities and providing access to jobs through training or broadband access.

### Progress

***Data sources.*** Historically, the impacts of federal programs on employment have been based on projections. However, reliable data on jobs and employment are collected quarterly by federal agencies, and these data are available by county and, in some cases, by census tract. These data could be used to track economic recovery in relatively small areas. It may also be possible to use the data to perform analyses that attribute economic improvement to specific projects. Data sets will include but are not necessarily limited to Longitudinal Employer-Household Dynamics (LEHD), the Census Business Register, and the Quarterly Census of Employment and Wages (QCEW).

***Plans and priorities.*** The project leads will consult with data experts at the Department of Commerce (DOC) and the Department of Labor (DOL) to report what administrative and statistical data can be used to track recovery and project impact; what barriers (e.g., in policy, regulation, law, and systems capability) exist to using the data for tracking recovery and project impact; how the barriers can be reduced or eliminated; and what specific changes in practice, policy, regulations, and/or law are needed to reduce or eliminate the barriers.

### Challenges

The primary challenge with using statistical administrative data for project and program monitoring is with the controls in place (in law, rule, or policy) to ensure that analysis and reporting do not disclose personal or business confidential data. Increasing agency resources for screening project proposals to weigh risk and utility can address the challenge by reducing cycle time needed for important policy-relevant research. Additional resources are also needed for adapting the data for use through encryption, redacting data, or employing privacy-protecting technologies.

### Lessons Learned

***The benefits of using statistical and administrative data for program evaluation and monitoring.*** The DOC working group showed the benefit of using statistical and administrative data to monitor and evaluate the impact of the ARP and IIJA programs. The advantages of the approach are very significant. While evaluations done using survey information and data reported by grantees and contractors typically provide findings years after funding has been provided and require a verification protocol, statistical information from sources such as the QCEW are updated quarterly. This information is objective and does not require any additional reporting from program beneficiaries. The information is far more current than survey information and can therefore be used to make course adjustments as a program is administered. The approach also eliminates reporting requirements that discourage underserved communities and populations from applying for benefits.

Given the benefits of using statistical and administrative data for project monitoring and evaluation, the procedure for performing this needs to be established for each individual Notice of Funding Opportunity. The best data source and approach to the analysis depends on the intended impact of a funding stream. However, most ARP and IIJA funding streams have a primary or secondary objective of increasing economic activity within an area. Data that show changes in the wages and earning of residents, employment levels of an area, and commercial activity are usually reasonable indicators of changes in the economic condition of an area. DOL, DOC, and the Small Business Administration have all used statistical data for evaluation. Adding commercially available data that tracks economic activity (e.g., credit card information and data from payroll processing firms) may provide additional useful information on near-real-time local economic conditions.

***The use of census tracts.*** In the IIJA report, the DOC working group recommends that grant awardees report on the geographic location(s) of the awardees, the point of service delivery, and where the program is expected to have an impact. The group also supports using census tracts as the standard for reporting. To track how interventions impact the economic conditions of census tracts, data on conditions are required because many projects are not designed to impact conditions at the county, or metropolitan area level. Efforts such as these to collect information on investments by a standard geographic unit of reporting, such as census tract, should be encouraged.

***The need for collaboration.*** Currently, some of the most useful data for project monitoring and evaluation is held by states, multi-state collaboratives, and state or university data centers. Typically, these data centers not only have the most current wage and earning data but data on many other factors that influence the economic status of areas, e.g., education, access to health facilities, crime, and incarceration information. One example is the Ohio Longitudinal Data Archive (OLDA), a collaborative arrangement between the State of Ohio and the Ohio State University. Operated jointly by the John Glenn College of Public Affairs and the Center for Human Resource Research, the OLDA stores data from five agencies (Education, Higher Education, Housing, Job and Family Services, and Opportunities for Ohioans with Disabilities) in Ohio. These data are available to government agencies as well as to external researchers. By providing access to both networks, Ohio creates a community focused on generating evidence-based research that is used by government for both research and public policy. The OLDA provides a useful model for services that a National Secure Data Service could provide. Other examples of state and collaborative models include the Midwest Collaborative and the states of South Carolina, Kentucky, and Minnesota.

***The costs of data preparation.*** Data cleaning and formatting for analysis are some of the most time consuming and expensive parts of the process of using the data for monitoring and evaluation. One major contribution of the OLDA is data preparation. Any approach intended to increase the use of data for evidence building must provide for the cost and expertise needed for this process. The cost of the process could be amortized by making the curated data available for several research projects. Additionally, the costs of data preparation could be reduced by establishing standards for file format, documentation, and syntax.

### *Other Observations*

If program evaluation is not a routine part of a government's business model, repeated investments may be made in interventions that do not work. To avoid catastrophic waste, rigorous evidence must be built on the best strategies for addressing economic events that inevitably recur nationally, regionally, and locally. However, the aspiration of routine, rigorous evaluation cannot be achieved using a model of occasional multi-year intensive reviews that produce findings after multi-million-dollar investments. To know which programs work and under what circumstances, program evaluation must be "routinized" and affordable. Using the statistical and administrative data of governments is the most promising avenue toward that end. Cross-government collaboration on evaluation may increase insight and further reduce costs.

## Education and Workforce

### Champions: Gregory Fortelny and Shawn Davis, *ACDEB Members*

### Project Description

The use case reviews current and evolving approaches to accessing, linking, and analyzing education and workforce data across the federal, state, and local levels with a consideration of how decisionmaking could be enhanced and facilitated. This includes investigating improvements for the current evidence-building ecosystem and weighing the possibilities of a National Secure Data Service (NSDS).

The primary models of focus for the use case are the Department of Education's Ability to Benefit (ATB) program and the Midwest Collaborative (MWC).

### *Rationale*

This use case highlights key aspects of the Evidence Act's vision for evidence-based decisionmaking. In addition, the use case advances the recommendations of the full Committee as well as the Governance, Transparency, and Accountability and Government Data for Evidence Building subcommittees, as outlined in the Year 1 report.

More specifically, the use case explores the following:

- The value of education and workforce data access, linking, and analysis for evidence building for data providers and users;
- Potential data sources;
- Barriers to accessing, linking, and analyzing state, federal, and other data sources and possible ways to overcome these challenges;
- The privacy-quality tradeoff; and
- What data governance frameworks are currently in place or could be adopted to facilitate access to and linking of education and workforce data.

### *Importance and Value Proposition*

As described in ACDEB's Year 1 report:

> ***"Unprecedented changes in labor markets have led to fundamental changes in skill demands. Both sets of changes underscore the need to strengthen the connection between employment services, post-secondary programs, and workforce outcomes. Building these links will help individuals decide what education paths best meet their needs and will encourage high-return investments in skills that yield long-run economic security and mobility"***

The ATB program and the Midwest Collaborative exemplify the value proposition for leveraging education and workforce data for evidence building.

## Model 1. Ability To Benefit Program

### *Background*

In 2019, 2 million 16- to 24-year-olds were not enrolled in school and had not earned a high school credential. The ATB program allows a student without a high school credential (or recognized equivalent) who is enrolled in an eligible career pathway program to qualify for Title IV federal student aid. The career pathway program must enable the student to attain a high school credential.

Better data could help decisionmakers answer questions like:

- Which students take advantage of the ATB program?
- Did they concentrate in career and technical education (CTE) in high school?
- Do they complete a high school credential?
- Do they complete a postsecondary credential?
- What are their occupation and wage outcomes?
- Does their occupation align with their career pathway?
- Does their career pathway align with local labor market conditions?
- How mobile are ATB graduates across state and local markets?

### *Progress*

This is a potential project, so progress is reflected through identifying the value proposition, possible data sources, and barriers to data access, linking, and analysis (see below).

***Value proposition.*** Better access, linking, and analysis of data from the ATB program and other relevant data sources have the potential to provide myriad benefits for stakeholders.

The value proposition for states includes:

- Understanding the relative influence of high school CTE programs and supports on the ATB population to plan future investments in education.
- Discovering which industry-recognized credentials provide the most value to learners.
- Monitoring how CTE programs respond to local and regional labor market conditions.
- Building community awareness of positive effects of the ATB program to encourage more students to apply.
- Learning about student mobility across state lines through education and into the workforce.
- Learning the value of using a trusted third party for secure state data linking and analysis services to leverage Federal Student Aid Data, other federal data, and data from other states.

The value proposition for learners and consumers includes:

- Building awareness among school personnel and parents about CTE programs, career pathways, and the value of the ATB program for students who have not yet completed a high school credential.
- Encouraging more learners to apply.
- Providing adult learners who have dropped out a viable pathway to both a high school and a postsecondary credential.
- Understanding the earnings effect of participation in the ATB program using aggregate cohort earnings from Internal Revenue Service (IRS) data.

The value proposition for the federal government includes:

- Understanding the effect of Pell grant awards on the ability of ATB students to complete postsecondary credentials.
- Building community awareness of positive effects of the ATB program to encourage more students to apply.
- Establishing grant programs to support effective career pathways for ATB students to complete credentials and earn a living wage.

***Data sources.*** Data sources have been identified to realize the value proposition related to the ATB program, including the following:

- Federal: Individual student-level data from the U.S. Department of Education's (ED) Office of Federal Student Aid (FSA) (recipients of federal student aid student only); IRS and Statistics of Income Division (SOI) data for ATB student cohorts.
- State: Individual-level student data from Statewide Longitudinal Data Systems (SLDS); data from State Workforce Boards.
- Private: Individual student-level data, such as those from the National Student Clearinghouse.
- Private: Individual data on CTE certifications from credentialing boards.
- Others: Aggregate data on local and regional labor market conditions for CTE fields.

### *Challenges*

While there are clear benefits to accessing, linking, and analyzing education and workforce data, like those described above, there are also known barriers to these activities.

For example, barriers to using data from SLDS include:

- Key data systems (like the SLDS) are owned and managed by states.
- Multiple state agencies contribute to the data system, and each agency controls its own data.
- Each state agency data contributor must approve the use of its own data.
- If the contributors do not agree to share their data, then those data are not accessible for cross-agency or cross-state analysis.

- Data in SLDS systems are protected by the Family Educational Rights and Privacy Act.
- A less-than-50-state SLDS solution would not provide many of the value propositions cited above.

Likewise, there are barriers to ED's FSA office data, including the following:

- ED owns and governs individual-level data on federal student aid recipients, prioritizes the privacy and security of its individual-level data, and would need to agree to let a trusted third party use these data.
- The Higher Education Act contains a provision prohibiting nongovernmental researchers and policy analysts from accessing personally identifiable information (PII) within the National Student Loan Data System, where some key data elements are stored. Historically, this has narrowed the scope of potential evidence-building initiatives.

There are also barriers related to IRS data, including the following:

- The SOI Division in IRS has partnered with ED to provide aggregate earnings data for the College Scorecard.
- Title 26 contains provisions limiting the disclosure of federal tax information.
- It is unclear whether the ATB program data could also be considered for tax administration purposes.

And barriers for other data:

- Data on federal unaided students (such as those managed by the National Student Clearinghouse) are available but at a cost.
- Multiple credentialing organizations provide industry-recognized certifications for CTE fields. In health care alone, they are legion. Identifying all the relevant organizations and establishing data sharing agreements is likely not feasible.

***Privacy-quality tradeoff.*** Finally, the privacy-quality tradeoff must be addressed. For example, exploring possibilities for the ATB program depends on linking multiple data systems that contain sensitive PII. The governance issues are the main barrier to realizing this use case. Rather than physical linkage, this use case might be a good candidate for secure multiparty computation.

### *Lessons Learned*

***Provide value.*** Critical data for evidence building about the connection between employment services, postsecondary programs, and workforce outcomes are currently closely held in states and often controlled by individual state agencies. While groups of states are realizing the value proposition of secure data sharing for evidence building (see the Midwest Collaborative description below), only a 50-state solution will meet federal needs for evidence on programs and policies. Achieving this nationwide perspective first requires demonstrating to each state that such a solution meets their own needs for evidence about education and the workforce.

***Centralize requests.*** Understanding the connections between education and the workforce requires sensitive and secure federal data on student aid and income/employment. Any evidence-building efforts using federal student aid data or IRS data must be approved by those agencies and conform to allowable purposes under statute. Establishing a central point of access for evidence-building projects using these data could support multiple use cases to inform policies and programs in this area.

***Share resources.*** Data from the ED's Federal Student Aid office only cover students who apply for federal student aid and then follows the enrollment status of only those who are aided. Understanding the matriculation, transfer, and completion of all postsecondary students, aided or not, is essential to modeling public policy and developing effective programs. Developing a governmentwide agreement with a private organization that tracks student enrollment, such as the National Student Clearinghouse, could provide scaling efficiencies to serve multiple social service agency use cases for evidence building.

## Model 2. Midwest Collaborative

### *Background*

The Midwest Collaborative (MWC) is a regional collaborative of Midwest states that joined together to share education, training, and workforce data through a value-driven approach to building data infrastructures. The MWC governance structure consists of an executive committee that oversees an Administrative Data Research Facility (ADRF). The ADRF consists of a secure, cloud-based platform, a policymaking body, and a technical advisory body. The interim administering organization is the National Association of State Workforce Agencies (NASWA), and the interim platform organization is the Coleridge Initiative.

***Governance structure.*** The key components and features of the MWC governance structure include the following:

- ***MWC Executive Committee.*** The MWC Executive Committee determines final approval on all policy recommendations and project proposals and consists of state representatives from the MWC Council and MWC Data Stewards Board.
- ***MWC Council.*** The MWC Council is the policymaking body for the collaborative. The goal of the Council is not to prevent states from doing what they wish with their own data but instead to provide a set of rules of engagement to allow states to work with one another more easily. The Council provides a means for states to focus on the core questions for educational workforce needs by providing a request for proposal approval process and standardized disclosure forms and by helping manage the review process for expedited access for states and researchers.
- ***MWC Data Stewards Board.*** The Data Stewards Board provides technical advice for the collaborative and consists of staff members who are subject matter experts regarding the data within the ADRF. The board additionally provides best practices for data use as well as advice on how to link data sets.
- ***Administering Organization (NASWA).*** The administering organization engages states to communicate the value proposition of the MWC as well as to determine states' needs. Other duties include enhancing interstate collaboration and development as well as implementation of governance arrangements.
- ***Platform Organization (Coleridge Initiative).*** The role of the platform organization is to provide and support the ADRF as well as provide needed technical training and advice to stakeholders.
- Administrative Data Research Facility. The ADRF is currently a FedRAMP-certified environment built on top of Amazon Web Services (AWS) GovCloud, whose capabilities include data ingestion, data documentation, data analytic tools, and data stewardship. Authorized participants in the collaborative receive browser-based access to databases, file systems, and external websites.

From a security perspective, the ADRF is fully encapsulated within AWS on a segregated network with no public access, and the platform requires multifactor authentication. Databases are encrypted at rest, and each research team is assigned a separate database and shared drive. Researchers are provided AWS AppStream virtual desktops that are temporary and do not retain information once a session has been completed. Overall, researchers can only access read-only data for approved and authorized projects.

Though an administrative user with access to multiple databases could be compromised, MWC's separation of duties mitigates that risk.

The ADRF has incorporated the Five Safes principles, which include safe projects, people, settings, data, and outputs. The purpose of these principles is to increase utility while maintaining acceptable risk.

- ***Approval process.*** The MWC is working on a streamlined project approval process for two or more states with data in the ADRF that wish to collaborate. First, the states submit a project description to the executive committee. If approved, their data stewards allow access, and the MWC offers structural context and support. A continual focus for the collaborative is to add value for states and help them to socialize.

  External researcher access is part of a request for proposal process in development. States identify priorities and research interests, and a proposal is submitted based on these priorities. The proposal starts at the policy council then moves to the data stewards who ensure it aligns with allowable uses. Each data source has a data steward who makes that decision, and the executive committee gives final approval. Once approved, a researcher can perform matching and linkages within the ADRF between approved data sets.

- ***Privacy.*** The MWC uses a cell suppression size of 10 and hashes direct identifiers. Parties approved to access the ADRF are heavily vetted, sign confidentiality and non-disclosure agreements, agree not to use screen scraping software, and are under contract not to attempt to re-identify data. Researchers are also not allowed to import into the ADRF their own data that could potentially be used for unauthorized matching or re-identification. The ADRF is primarily used for data access, as opposed to allowing exports. Any exports are statistical results proofed for confidentiality, as opposed to microdata containing personally identifiable information. All export requests undergo a privacy and security review as well. Data destruction and retention periods are based on individual contract language, as approved by the data owner for each project.

### *Progress*

***Benefits.*** Through this collaboration, states are realizing many benefits for improving access, linking, and analysis of education and workforce data, including the following:

- Developing standard, state-driven reports and multi-state products;
- Establishing the ability to link needed data across state lines while protecting privacy; and
- Providing an adept training program to develop teams of researchers who can solve real-world, practical problems.

***Data sources.*** The MWC harnesses a variety of high-value data sources provided by state partners, including the following:

- State unemployment insurance (UI) wage records,
- National data clearinghouses,
- Postsecondary data from 2-year and 4-year technical colleges,
- Statewide Longitudinal Data Systems (SLDS) that contain individual record-level data on public school students, and
- Rehabilitation and housing data.

***Questions answered.*** The MWC has leveraged its governance framework to answer critical questions with the education and workforce data provided by its members, including:

- What is the impact of government programs in terms of how access to certain support pays off for earnings?
- Are state programs and institutions effective?
- How are special services for people with justice system involvement paying off?
- Are states meeting educational attainment goals?
- Are there wage gaps among racial and ethnic groups due to a lack of access to postsecondary enrollment options and concurrent enrollment programs?
- Are there barriers in postsecondary attainment due to attending secondary schools with less experienced teachers (possibly leading to a lack of postsecondary preparation during high school and resulting in the need for remediation during college and a lower likelihood of completing a degree program or credential)?
- Is enrollment in baccalaureate programs aligned with the needs of local employers?
- Is there evidence of job sorting mechanisms that create segregation in the labor market?
- Are aid and scholarship programs in postsecondary education resulting in student success?
- Are students able to make better-informed decisions about which programs or occupations may benefit them?
- Are institutions and states using evidence-based data to improve programs?

***Data products.*** Examples of MWC products include the following:

- Multi-State Postsecondary Report
- Unemployment to Reemployment Portal

### *Challenges*

***Barriers.*** While significant progress has been made, challenges remain, including the following:

- There is not currently a process in place for investigations to ensure that researchers are using data for disclosed purposes, which would pose a greater concern if nonstatistical data or microdata were allowed to be exported.
- While the MWC makes a case-by-case effort to reduce the chances of re-identification when certain combinations of indirect identifiers are present, there does not appear to be a standardized protocol for doing so.
- Replacing NASWA or the Coleridge Initiative would be difficult due to concerns that new administrative or platform organizations would not have the same richness of resources and experience.

***Opportunities and gaps.*** Opportunities for addressing these challenges and expanding the MWC projects, products, and processes include the following:

- Further audits and risk assessments of potentially re-identifiable data (such as indirect identifiers that could be combined or geolocation information) are necessary should export of such data ever be allowed.
- Potential investigative mechanisms to ensure authorized uses are being followed should exports containing potentially re-identifiable data ever be allowed.
- Even though exports are generally not allowed, there is still a risk of screen scraping or screenshotting of sensitive data. A tiered access protocol for highly sensitive data could require monitored access either at a secured facility or remotely with a webcam proctoring solution.
- A formal labeling system for different sensitivity levels of data may be needed. Currently, state survey data are qualified by individual PII, company PII, and allowable use.
- The federal government has been sharing data between agencies for years, but interstate sharing is still new for some states. A single agreement to cover data sharing would be helpful.
- It is difficult to simplify the complexity of different mandates that govern allowable uses of data while integrating the data into a single platform that meets state expectations.
- Increasing socialization across states and scaling up collaboratives across the nation will likely lead to a need for national governance to ensure continued alignment. Funding at the federal level would be helpful to support such a national entity. Covering fixed and marginal costs would lead to greater participation from states.
- States must be able to identify priority areas and be actively involved in the proposal process to ensure their priorities are reflected in the work being conducted.

- Further state participation in the governance structure is needed.
- There is a potential for a data concierge, as the current library of data elements is mainly self-service; states can contact technical staff directly if needed.
- One-offs for projects or research questions are common, but future similar needs require continual renegotiating.
- More real-time data would be beneficial for addressing public policy issues.
- Further interstate data sharing would allow states to track graduates who gain employment across state lines.
- Self-employed people, people in the military, or those who are federally employed do not show up in UI workforce records.
- UI workforce record matching relies on social security numbers to match postsecondary students to the workforce. Students who move from K-12 education directly into the workforce often cannot be matched as a social security number is not generally collected in K-12. States need to determine a way to allow linkage of these students while preserving privacy and confidentiality.
- One of the most significant limitations is that UI workforce records do not include occupational information. Some employers feel reporting such information would be a burden due to the necessity of setting up and assigning occupational codes. In addition, it is difficult to determine if an individual is fairly compensated for part-time employment, underpaid for full-time employment, or if they are employed in a typically high-paying or low-paying occupation. Records also do not include the number of hours worked; instead, they reflect the number of quarters that workers were employed.
- There is a need for standardization across states around data elements that the National Center for Education Statistics collects. Information varies greatly across states, which can make cross-state comparisons difficult.
- States can reduce gaps and create more value by including more agencies and more data through:
  - Information on postsecondary education wraparound services such as food, housing, and security;
  - K-12 information from social service providers;
  - Information about early childhood;
  - Information from justice system educational programs; and
  - Information about non-credit instruction, such as contracted training for an employer, commercial driver's license programs, emergency medical technician programs, and information technology certifications. There is no standard for these programs, and creating one poses challenges.

### *Lessons Learned*

***The possibilities of training.*** Data training is going well based on richness of existing resources. The MWC leverages the Coleridge Initiative's Applied Data Analytics training program. This program is a project-focused learning approach designed to train government employees and public policy analysts on how to tackle significant policy problems. The approach emphasizes applying modern data analysis tools to their own confidential data. Agency staff is trained through direct use of agency data to answer real, present policy questions and to develop practical tools after the training ends.

***The power of projects.*** Collaborative activities energize participants by showing them how quickly established projects can be pushed to other states, creating momentum. For example, Illinois created an unemployment and reemployment portal over 8-10 months, working with Coleridge staff. Indiana recognized value in the portal and produced an operational version in 2 weeks. Arizona was able to quicky build onto this existing work, creating dashboards in the portal that feed other states.

***The promise of practice.*** States are beginning to see the value of collaboratives, and more are being formed, notably in the South and the East.

## Health

### Champions: Brian Moyer and Kimberly Murnieks, *ACDEB Members*

### Project Description

The use case reviews current and evolving approaches to accessing, linking, and analyzing health data across the federal, state, and local levels with a consideration of how decisionmaking could be enhanced and facilitated. The use case investigates improvements for the current evidence-building ecosystem and weighs the possibilities of a National Secure Data Service (NSDS) by exploring recent advances in the National Center for Health Statistics (NCHS) National Vital Statistics System (NVSS) modernization efforts.

### *Rationale*

This use case highlights key aspects of the Evidence Act's vision for evidence-based decisionmaking. In addition, the use case advances the recommendations of the full Committee as well as of the Governance, Transparency, and Accountability and Government Data for Evidence Building subcommittees, as outlined in the Year 1 report.

More specifically, the use case explores the following:

- The value of health data access, linking, and analysis for evidence building and how this can be enhanced through better two-way communication and collaboration among the federal, state, and local levels.
- The importance of data standards, consistency, and interoperability among federal, state, and local governments to support the value proposition.
- The impact of the interpersonal side of the collaborative process, including the role of communities of practice, training, and resource sharing.

### *Importance and Value Proposition*

The COVID-19 pandemic has highlighted the need for more timely, accurate, and reliable health statistics. Decisionmakers are demanding better data to answer questions like:

- What are current death counts, causes of death, and the level of "excess deaths"?
- How long is the average person expected to live, and how has the pandemic impacted this?
- How do death rates and causes of death vary by state? County? Race and Hispanic origin?
- What is the impact of COVID-19 infection during pregnancy on health services utilization and maternal and infant outcomes?

More broadly, researchers, program administrators, and policymakers are looking for ways to connect health statistics to other sources of information to yield better evidence for decisionmaking.

## National Center for Health Statistics National Vital Statistics System Modernization

### *Background*

The NVSS is the oldest and most successful example of inter-governmental data sharing in the public health realm. It is the mechanism by which NCHS collects and disseminates the nation's vital statistics. The NVSS is a multitude of systems that carries about 6.5 million records a year from the local level through states to the national level, and back again.

***Data sources.*** By collaborating with other public and private health partners, NCHS uses a variety of data collection methods to obtain accurate health information. NVSS sources of data include the following:

- Birth and death certificates;
- Patient medical records, including electronic health records;
- Personal interviews (in households and by phone);
- Standardized physical examinations and laboratory tests; and
- Health care facilities and providers.

***Modernization.*** For the past decade, NCHS has been modernizing its National Vital Statistics System (NVSS) to allow for more rapid receipt, coding, review, analysis, and release of these data. In FY 2020, Congress began to appropriate additional resources for data modernization. These efforts involve extensive collaboration between the Centers for Disease Control and Prevention (CDC) and NCHS, the Jurisdiction Vital Records Offices, and other public and private sector partners. The goal of this work is to improve the quality of data and statistics and increase the ability to exchange data at both the federal and jurisdictional levels for improved evidence-based decisionmaking.

A modernized NVSS will transform the nation's vital statistics network to support nearly real-time public health surveillance and make vital records more available for action. The effort will (1) modernize NCHS internal systems to fully support record-level processing; (2) enhance existing data-quality review tools, incorporating data visualization and automated processes to streamline the detection of data-quality challenges; (3) develop and enhance automated coding systems using natural language processing and machine learning; and (4) develop APIs for Fast Healthcare Interoperability Resources (FHIR)-based, bi-directional interoperability between jurisdictions and NCHS. To achieve these outcomes, a focus is on technical assistance to jurisdictions, including new training materials to bolster jurisdictional staff capacity.

The modernization effort will also ensure quality and consistency in death investigations and certifications by developing consistent processes, standards, systems, and training for investigating, certifying, and reporting cause-of-death information. This will contribute to higher-quality vital statistics during and after pandemic responses, including information on race and Hispanic origin, as well as information on industry and occupation critical to disparities research.

This use case focuses on the steps necessary to modernize federal, state, and jurisdictional systems to facilitate the exchange of vital records data among states/jurisdictions, surveillance programs, and CDC/NCHS—a win-win for all players in this space and, perhaps, a model to highlight certain elements that are particularly relevant to the design of the NSDS.

### *Challenges*

Barriers include the following:

***Consistency, data standards, and system interoperability.*** While there has been funding over the last decade to help develop state systems, including work on systems interoperability and data standards, not all jurisdictions provide data to NCHS using common standards.

***Capacity for data providers.*** The technical maturity of jurisdictional partners varies greatly, so NCHS must find ways to support and build capacity for all state agencies.

***Data sharing for timely decisionmaking on both sides of the data stream.*** Successful evidence building involves two-way data sharing between NCHS and states—data must flow "up" to the federal government and then back "down" in a timely way for those data to be useful and used in decisionmaking.

### *Progress*

Ongoing NCHS modernization efforts are structured to address these barriers and challenges.

***Consistency, data standards, and interoperability.*** A significant aspect of NVSS modernization is consistency across data sets, and this work is being advanced by recent CARES Act funding to all 57 jurisdictional partners.

- ***Application Programming Interface (API).*** The NVSS API supports the exchange of mortality data between NCHS and vital records jurisdictions. The API enhances data interoperability through (1) record-level exchange to improve the timeliness of sending data to NCHS and receiving responses from NCHS, (2) automation so that there is less time spent shepherding the process and more time delivering value from data, and (3) increased reliability and robustness by building reliable message delivery into the system architecture. In March 2022, NCHS implemented a fully functional NVSS API and integrated messaging platform for testing, with the rollout of the production API expected by the end of 2022.
- ***FHIR standards.*** The NVSS API is built using the FHIR standard for the electronic exchange of health information. This standard allows data to be reused by multiple parties for multiple purposes. Benefits of using FHIR for vital records exchange include standardization both within vital records and against non-vital records systems, automated validation, developer support, and process improvements. In conjunction with the API testing, NCHS issued the Vital Records Death Reporting FHIR implementation guide that builds on the base FHIR standard to support the capture and exchange of mortality data.

***Capacity for data providers.*** The Vital Statistics Modernization Community of Practice is a virtual forum for sharing ideas, technical tools, resources, and promising practices to improve birth and death data. NCHS welcomes all jurisdictions and partners interested in modernizing the vital records system, at any level of experience, to participate.

Goals of the community of practice (CoP) include the following:

- Create an environment that fosters technical collaboration, knowledge transfer, and sharing of lessons learned;
- Promote the adoption of modern standards for interoperability;
- Identify challenges and work toward solutions;
- Explore opportunities to collaborate with partners;
- Promote best practices;
- Share advances and lessons learned with communities outside of the CoP who have relevant touchpoints;
- Foster innovation; and
- Create "bi-directional" understanding of where participating entities are in terms of modernization efforts.

The Vital Statistics Modernization CoP offerings include community meetings, communications, a SharePoint Knowledge Management site, on-demand resources, and ongoing quarterly virtual testing events. NCHS is also standing up a Community of Practice Steering Committee with representatives from jurisdictions, the National Association for Public Health Statistics and Information Systems, and NCHS. The Steering Committee will assist with developing agendas for monthly community calls and will help ensure the community understands and is effectively supporting all jurisdictions in their modernization efforts.

***Data sharing for timely decisionmaking on both sides of the data stream.*** NCHS continues to improve both the data flowing from states to the federal government and data flowing back to the states to better inform the states' own decisionmaking.

- ***Data flowing "up."*** NCHS is working with vital records offices to speed the collection of birth and death data and to improve the quality of those data.
    - ***Automated communications.*** The NVSS API allows jurisdictional mortality data systems to automate communication with NCHS in a robust and repeatable way. Automation improves the timeliness of the data exchange and reduces the burden on records stakeholders.
    - ***Advanced coding technologies.*** In June 2022, NCHS introduced a new cause of death coding system, the MedCoder, that fuses natural language processing and machine learning techniques with well-established rule-based approaches to automatically code causes of death. This system is much easier to maintain, as it streamlines corrections, provides a systematic process for testing changes, and harmonizes records processing and transmission.

- ***Data flowing "down."*** The agency is releasing more data more quickly to help decision-makers better respond to emergencies and is enhancing access to those data.
  - ***Timeliness.*** Thanks to recent investments and modernization efforts, NVSS releases statistics and special analyses that are timelier and more frequent. For example, since April 2020, NCHS has published provisional death counts weekly with geographic and demographic details. The agency is also working to reduce the reporting lag for provisional drug overdose counts from 6 months to 4 months.
  - ***Enhanced access.*** NCHS continues to expand and improve its tiered access model that includes public-use data files for open data assets, a web-based query system for creating customized tabular data views, and physical enclaves for accessing data in controlled, secure environments. For example, in early 2022, NCHS enhanced access to provisional mortality data via the CDC WONDER portal and is planning to launch similar improvements to provisional birth data later in 2022. In addition, the agency is developing a virtual data enclave to eliminate barriers that exist with physical locations, decrease costs for researchers, and expand access to confidential data.

### *Lessons Learned*

***The value of health data for evidence building and the role of communication and collaboration for enhancing that value.*** America relies on the NVSS to provide timely, high-quality data on births and deaths. NCHS is reinventing the NVSS to make vital information more available for action. When policymakers and citizens have accurate data fast enough, the nation can recognize and track developing health threats—from preterm births to deaths from drug overdoses, flu, and COVID-19—and make better decisions. NVSS modernization projects focus on implementing innovative strategies that reduce the burden on data providers while making vital statistics more useful and available for public health decisionmaking. The success of these projects relies heavily on communication and collaboration. (For more information on the roles of communication and collaboration, see the bullets below.)

***The importance of data standards, consistency, and interoperability.*** FHIR has opened doors to innovation that will strengthen the entire health data ecosystem. By enabling health systems to communicate information using a common framework, these standards help break complex health information into small, reusable parts that can be combined, disassembled, and recombined to meet a variety of needs. FHIR allows resources developed by one organization to be leveraged by another, delivering more "real-time" and automated data feeds and helping organizations make the transition to high-capacity cloud- and web-based technologies.

***The impact of the interpersonal side of collaboration.*** Throughout the years, NCHS has learned that connecting technology begins with connecting people and ideas. Mechanisms to support this connection include communities of practice, training, and resource sharing.

- ***Communities of practice.*** The National Vital Statistics CoP is a shared space for learning and innovation where people working together on the same challenges in the vital records modernization space can talk about their experiences and offer each other technical and practical help. This forum is key to creating efficiencies and accelerating modernization efforts.
- ***Training.*** Guidance comes through in-depth technical assistance, peer-to-peer sharing, live product demonstrations, and regular touchpoints for connection and collaboration.
- ***Resource sharing.*** NCHS supports sharing aimed to "build and reuse" between jurisdictions and partners, make technical solutions open-source and available online, identify common pain points, and help brainstorm solutions that can be implemented across the entire vital records ecosystem.

## Labor Market Activity

### Champions: Julia Lane and Christina Yancey, *ACDEB Members*

### Project Description

The primary focus of the use case is to review current and evolving approaches to accessing, linking, and analyzing labor market activity data and to discuss the lessons learned for the NSDS.

There is tremendous potential to enhance measures of labor market activity both to create new national and local labor market statistics and to support better evaluation and research leading to continuous improvement and program delivery. This use case considers recent advances that have been made using data from the unemployment insurance (UI) system by drawing on two initiatives—the Department of Labor's (DOL's) UI Equity Data Partnerships and state regional data collaboratives.

### *Rationale*

This use case highlights key aspects of the Evidence Act's vision for evidence-based decisionmaking and seeks to explore governance strategies as described by the Evidence Commission. In addition, the use case advances recommendations of the full Committee as well as the Governance, Transparency, and Accountability and Government Data for Evidence Building subcommittees.

More specifically, the use case explores the following:

- ***The value of pilots and training to test different approaches to ensure feasibility and demonstrate value.*** A common feature was using training to identify the potential to combine data across agency lines to create new measures—of equity, of unemployment to reemployment, and of labor demand—to inform program delivery.
- ***The importance of resources to enable data access, linking, and standards development.*** A common feature was investing in people and technologies to develop legal agreements, de-identification approaches, secure environments, linkage methods, and joint determination of practical standards.
- ***How data governance and privacy protections can be institutionalized and operationalized.*** Common features include formal collaboration agreements, transparency for data stewards, and implementation of a Five Safes framework.

### *Importance and Value Proposition*

Labor market unemployment and employment measures trigger policy responses at the state and federal levels. The COVID-19 pandemic has identified several challenges with the current measurement system, particularly at the local level and particularly for historically marginalized communities.

- For example, there are challenges with allocating survey-based estimates to the level of the individual states. As BLS notes, since the onset of the coronavirus pandemic, they "have been examining the inputs to the employment and unemployment models for outliers and implementing shifts in real-time, where appropriate, based on statistical evaluation of the movements in the inputs." These level shifts help to preserve movements in the published estimates that the models otherwise would have discounted; however, these adjustments also distorted data sets on which governors rely to make timely policy decisions. (For more information, see a recent press release from Michigan.)
- Likewise, the Government Accountability Office (GAO) has identified substantial measurement challenges in capturing UI claims activity and recommends that states, and the Department of Labor's Employment and Training Administration (DOL/ETA), move to analyze claimants.
- A recent article by Erica Groshen, former Commissioner of the Bureau of Labor Statistics (BLS), highlights the need for better measures of labor market statistics to analyze the impacts of the COVID-19 pandemic in relation to racial inequities.
- Furthermore, statistical agencies can share microdata for statistical use only. This means, for example, that while BLS has a wealth of knowledge and platforms that currently interact with administrative data systems, there are barriers to sharing the individual-level records for evaluation purposes.

While there are known challenges with existing official statistics as well as significant programmatic challenges to transforming administrative data into evidence, there are essential roles that federal, state, local, and private partners can play in overcoming these challenges.

- Federal statistical agencies could continue to improve data quality and quality control of official statistics, ensuring the ongoing availability of high-quality national benchmarks. Agencies could also explore better ways to provide administrative data for evaluation purposes, including creating secure repositories built on existing platforms with robust data quality checks and secure data linking.
- There are many ways in which constructive engagement with states could help solve these issues. For example, state regional collaboratives have worked together to develop timely and actionable data, as shown by the Illinois presentation at the Committee's Coleridge Initiative site visit (see ACDEB's Year 1 report).
- In addition, nontraditional data sources could meet the demand for better data for evidence building. For example, the Federal Reserve has used private sector data to study the evidence around the implementation of a major rescue plan (the Paycheck Protection Program) and, on that basis, recommended substantial changes in how data are collected and used for evidence.
- Finally, the role of researchers is potentially massive. The recent Nobel Laureate, David Card, pointed out the challenges associated with measures developed in the 1930s.

DOL's Equity Data Partnerships and state regional data collaboratives provide timely and relevant lenses through which to explore the value proposition for leveraging labor market activity data for evidence building.

## Model 1. UI Equity Data Partnerships

### *Background*

Beginning in 2022, DOL's Evaluation Officer and Chief Data Officer engaged in Unemployment Insurance (UI) Equity Data Partnerships with at least five states as part of a broader response to the COVID-19 pandemic with funds from the American Rescue Plan (ARP) Act of 2021. As one piece of the government response to the effects of COVID-19 on the economy and the employment situation for millions of workers, in August 2021, DOL launched a $260 million grant program opportunity to states with the specific aim of promoting equitable access to unemployment compensation programs. Under this program, states can voluntarily enter data-sharing partnerships with DOL to improve the utility and application of administrative data for analysis and to improve program administration.

Specific goals of this partnership include:

- Helping states better understand how different populations access UI benefits and identify potential barriers that those populations may encounter to receiving timely UI benefits,
- Improving the quality of data and fit for purpose, and
- Identifying opportunities to improve measurement, data collection, and data analysis.

Beyond the benefits for the data-sharing partners, this program offers helpful information for governance considerations for the NSDS. Aspects of this partnership that are compelling for potential future scaling and consideration for the NSDS include the following:

- Demonstrating the power and type of mutually beneficial partnerships between different levels of government,
- Highlighting an approach that is project-based and short-term in funding with longer-term implications,
- Signaling a distinct goal and financing from the federal government with customization and choice for state partners,
- Emphasizing the need to create value quickly while embedding capacity improvements into the project for longer-term value,
- Analyzing administrative data with an explicit purpose of appreciating what data and populations are missing from the existing data asset (i.e., continuous improvement in the quality of the data collected and sufficiency of the data for the intended purpose), and
- Improving equity in UI access and timely receipt of benefits by exploring new measures.

### *Challenges*

UI Data Equity Partnerships leverage existing data-sharing mechanisms, which reflect current legal interpretations that unemployment compensation data are overseen by each state. As such, state UI administrators determine how and when the data are shared based on their interpretation of state laws. All states provide the Bureau of Labor Statistics with summary-level statistics from the UI system; however, only a few states provide individual wage records. Note that the Census Bureau's Longitudinal Employer-Household Dynamics program currently receives most (but not all) states' individual wage records through a memorandum of understanding (MOU) executed individually with each state.

### *Progress*

UI Data Equity Partnerships offer an example of a federal-state data-sharing effort to improve the utility of a high-value national data asset, the unemployment compensation system data. The partnership aims to reduce inequities for eligible individuals in accessing UI benefits in a timely manner. More broadly, the program highlights a powerful model for scaling this type of work, including:

- Supporting interlinking goals to maximize short-term insights from data with longer-term continuous improvement objectives to improve administrative data quality as well as its sufficiency and completeness for intended purposes,
- Developing financing models that include specified goals and purposes with the flexibility to support piloting and experimentation, and
- Embedding sustainability, technical assistance, and subject matter expert support into data projects.

***Data sources.*** The work plans for each of the partner states are still evolving and will likely involve different data portfolios. At a minimum, each state will transfer individual-level data on all individuals applying for and receiving unemployment insurance benefits from 2018 through early 2022. Subsequent linkages are to be determined.

- State: Unemployment compensation microdata from 2018–2022
- Federal: Community Population Survey, American Community Survey, and other publicly available statistical data
- Others (TBD)

***Five Safes framework.*** This project involves the transfer of microdata, including sensitive PII from state systems to DOL using federal security protocols and technology. The arrangements are detailed through individual MOUs with each state and include specific provisions to ensure safety.

***Data products.*** Examples of possible data products include:

- Equity in UI access: before and during the COVID-19 pandemic
- Improving equity in UI access and timely receipt of benefits—exploring new measures

### *Lessons Learned*

***The value of pilots.*** One component of the government response to the COVID-19 pandemic and its effects on the economy and the employment situation for millions of workers is to examine equity issues in UI. In 2022, DOL and the state partners will assess the demographic distribution of UI payments, which will aid federal and state efforts to develop, pilot, and test equity measures in UI receipts. These measures support the broader unemployment compensation system's improvements in eliminating administrative barriers to applying for benefits; reducing state workload backlogs; improving the timeliness of unemployment compensation payments to eligible individuals; and ensuring equity in fraud prevention, detection, and recovery activities.

***The importance of technical assistance.*** States benefit from the partnership through access to an expert team convened at the national level. The team will conduct customized, advanced analysis during the active phase of the partnership and will provide technical assistance to support future state-directed data analyses in order to help improve UI policies and operations, especially as they relate to inequities.

***The importance of resources.*** DOL launched the $260 million grant opportunity to promote equitable access to unemployment compensation programs. Under the program, states can voluntarily enter data-sharing partnerships with DOL to improve the utility and application of administrative data for analysis and to improve program administration. States engaged in the Equity Data Partnership are supported by additional resources through DOL's direct analytics activities.

## Model 2. State Regional Collaboratives

### *Background*

Initiated in 2018, a group of states, mainly in the Midwest, formed a regional data collaborative to facilitate interstate collaboration on data, define a state-led data analytics infrastructure, build production-level technical capability, address privacy concerns, establish a professional development curriculum, develop a process for collective use of data for research and evaluation, and inform and shape the national evidence strategy.

### *Progress and Challenges*

For more information on the Midwest Collaborative (MWC) and examples of data sources, data products, successes, challenges, and next steps, see the "Education and Workforce" use case above.

### *Lessons Learned*

***The value of pilots.*** In March 2020, state agencies faced an immediate need to provide an effective, data-based response to the COVID-19 pandemic, and this need has been unabated since. Many states, inundated with millions of UI claims, did not have the capacity to translate the claims data (on transactions) to the need felt by claimants (on individuals). Fortunately, the March 2020 MWC meeting provided the basis for action. The MWC moved swiftly to develop a pilot unemployment to reemployment portal to inform policymakers—in essence, responding to an urgent GAO COVID imperative to establish analysis based on claimants and cohorts even before the imperative was issued (see above). The structure of the portal highlights weekly (timely), county-based (local), and actionable information on UI claimant composition and transitions.

Why was access to these data so critical? The need for state and local data was never greater, yet survey data were neither granular enough at the local level for subpopulations nor timely enough. Existing data sets only captured point-in-time data, not the experiences of people over time. Local workforce boards were faced with devising effective interventions for worker populations in a rapidly changing situation. The sheer volume was overwhelming; initial claims for unemployment per 1,000 population increased 16-fold from the March 2020 convening of the Midwest Collaborative to April 2020.

There were significant differential impacts by race as well. The inflow and concentration of Black UI claimants during the pandemic highlight local data patterns that suggest the need for strategic intervention. Data showed that white populations represented 60 percent of all claimants prior to the COVID-19 restrictions, and Black populations only comprised 20 percent. The disparate racial impact of the crises is such that 14 months later, the percentage distribution is 49.5 percent (white populations) and 33.3 percent (Black populations). Remediation strategies require understanding the demographic, industry-related, and occupational composition of those who are unemployed to better align resource allocation with local needs.

***The importance of technical assistance.*** The Midwest Collaborative, in conjunction with the Coleridge Initiative, established Applied Data Analytics training classes as an innovation sandbox that uses modular, active learning techniques to train participants on the use of complex data. Both agency staff and university researchers work together to address agency questions.

This approach has been successful because: (1) it develops teams of practitioners who can demonstrate the value of new types of data for solving real-world, practical problems, and (2) it creates a pipeline of new prototype products for stakeholders. State governments build on and enhance each other's products and create robust regional and national feedback loops. DOL's ETA supported training classes on unemployment to reemployment outcomes for over 100 state agency staff in more than 25 states and led to the establishment of related portals in close to a dozen states.

***The role of governance and protections.*** A key element to the success of these regional state collaboratives is the governance structure. The Midwest Collaborative has four components in its leadership structure: a policy council, a data stewardship board, an administering organization, and a platform organization. For more information, see the "Education and Workforce" use case above.

Participating states deposit data on a common platform with a shared security boundary, strong data stewardship, collaborative tools, and analytic capabilities. State representatives from the policy council and the stewardship board serve on the Executive Committee that exercises final approval on all policy recommendations and project approvals.

***The importance of resources.*** The Coleridge Initiative stood up a secure platform using the Five Safes framework with an investment of $2.5 million from the U.S. Census Bureau to inform the decisionmaking of the Evidence Commission. The Coleridge Initiative established training programs and collaboratives with about $10 million in funding over 4 years from private foundations.

## Environmental Quality and Human Health

### Champions: Richard Allen, *ACDEB Member*

### Project Description

The goal of this use case is to explore approaches for better accessing, linking, and analyzing data on environmental quality and public health. This includes investigating improvements to the current evidence-building ecosystem and the opportunities that a National Secure Data Service may provide.

The primary focus of the use case is per- and polyfluoroalkyl substance (PFAS) contamination of drinking water. There is great interest in understanding the human health effects of such PFAS exposure to inform potential policies and programs.

### *Rationale*

This use case highlights key aspects of the Evidence Act's vision for evidence-based decisionmaking. In addition, the use case advances the recommendations of the full Committee and of the Governance, Transparency, and Accountability and Government Data for Evidence Building subcommittees, as outlined in the Year 1 report.

More specifically, the use case reviews the following:

- The value proposition for combining environmental and health data,
- The current and emergent state of play, including potential data sources and barriers to environmental quality-health data linkages, and
- Lessons learned, including opportunities for the NSDS.

## PFAS Contamination in Drinking Water

### *Background*

Per- and polyfluoroalkyl substances (PFAS) are a large class of chemicals manufactured since the 1940s and used in a variety of manufacturing and consumer applications as well as in flame retardants. PFAS break down very slowly and are often referred to as "forever chemicals." Because of the slow breakdown of PFAS, they can build up in people, animals, and the environment over time.

As noted in a 2015 research paper, PFAS exposure is considered ubiquitous in the United States, detectable in blood serum samples of 70 percent of the U.S. population. For adults, contaminated food and water are exposure pathways of concern. Fetuses can be exposed to PFAS in utero and infants through breast milk or through contaminated water used in formula. As a result, drinking water is likely to be a key pathway of concern for communities that live in areas with high PFAS contamination.

Current scientific research suggests that exposure to high levels of certain PFAS may lead to adverse health outcomes. Human exposure to PFAS has been associated with negative effects on the immune, endocrine, metabolic, and reproductive systems. It is also believed that PFAS can increase risk for certain types of cancer; however, the evidence supporting these associations varies by both outcome and specific PFAS examined.

Researchers, government officials, and the public have a limited understanding of the full range of toxicological effects across a broad array of human systems and age groups. This is due in part to the diversity of unique chemicals in the PFAS group, which number in the thousands. Research has targeted the toxicological effects of only a small set of these chemicals. More research is needed to understand the effects of different levels of exposure, exposure at different ages, including in utero, and the effects of low-level exposure over time.

Academic researchers have had some success in this space by linking environmental quality data with restricted-use health data to study the health effects of PFAS exposure. For example, in one project, researchers compared locations with contamination where water filtration systems had been installed to locations that had never been contaminated in order to identify potential impacts on birthweights. The researchers found that before the filtration intervention, newborns in contaminated locations were more likely to have low birthweights; however, there was no statistical difference in birthweight for exposed newborns after filtration systems were installed. This research brought together multiple data elements across government bodies, including restricted-use health data.

The Environmental Protection Agency (EPA) recognizes the value in building on this work to advance the public's understanding of plausibly causal relationships between human health outcomes and PFAS in a potentially large source of exposure (drinking water).

### *Progress*

This is a potential project, so progress is reflected through the identification of the value proposition; possible data sources; and barriers to data access, linking, and analysis (see below).

***Value proposition.*** Research that quantifies the health impacts of exposure to environmental hazards is critical to informing the design of public health policies and programs. Incorporating EPA environmental monitoring data to restricted-use health information could help the following:

- Understudied health endpoints,
- Understudied PFAS,
- Pharmacokinetic modeling across exposure pathways,
- Relative source contribution analysis,
- Monetization of health effects,
- Integrated Risk Information System assets,
- Inform the design of policies to address active and legacy PFAS contamination,
- Improve retrospective analysis of policies and programs, and
- Environmental justice research on differential exposure and risk.

***Data sources.*** EPA has identified potential data sources to advance these efforts, including the following:

- Drinking water monitoring (federal and state); PFAS production and use sites; discharge monitoring; spills and response locations; federal sites; soil, water, and tissue samples; and water intake locations. For example, EPA could supply drinking water data from its 3rd Unregulated Contaminant Monitoring Rule on the presence of unregulated contaminants in drinking water, which includes data for six PFAS chemicals.
- CDC could be a federal partner and provide relevant health data from its National Health and Nutrition Examination Survey pertaining to blood serum-based PFAS measurements.
- In addition, the National Institutes of Health (NIH) Influences on Child Health Outcomes program provides information on 50,000 children, including blood serum measurements.
- Further linking CDC data assets such as natality statistics for birthweight and other health outcomes could be used to build evidence on the human health impacts of PFAS through the drinking water pathway.
- States have data with finer temporal and spatial resolution and, in some cases, lower detection limits to complement the EPA data. For example, states and localities may have more frequent monitoring data and, potentially, data for pollutants not available at the federal level. State health departments also often have detailed hospitalization and patient health records.

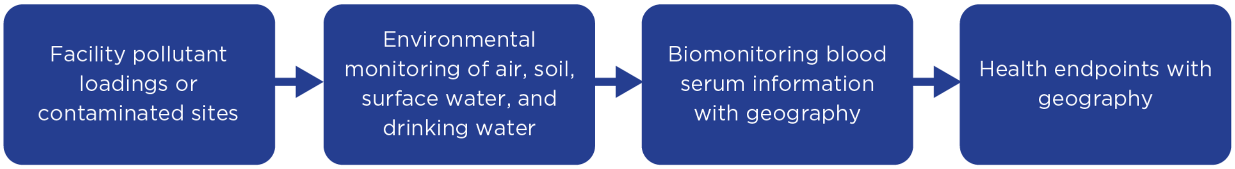
***Plans and priorities.*** This use case seeks to bring together at least two significant federal data assets with better controls around individual identification and with more granular state data to advance understanding of the national landscape of PFAS toxicology and to inform evidence-based policy decisions. EPA would also support building stronger research coalitions with relevant academic and other external research groups to get a stronger multiplier effect.

### *Challenges*

Barriers to environmental quality-health data linkages include the following:

- Expanding to new hazards and more powerful study designs could require more complicated data linkages covering the pollutant lifecycle (see Diagram S1).
- EPA researchers must invest resources and time to request the data, and CDC and NIH must review and manage each research protocol.
- The requirement to use physical Research Data Centers may include travel costs to the location and an onsite usage fee for each visit.
- There are substantial data management fees.
- The time between application submission and data access poses a challenge to the development of evidence on policy-relevant time horizons.
- Any changes to the analytic plan require revision and approval of a new data use agreement.
- Some local data may not be public, thereby requiring data-sharing agreements even if personally identifiable information or confidential business information are not involved.

Diagram S1. Pollutant Lifecycle


Facility pollutant loadings or contaminated sites
Environmental monitoring of air, soil, surface water, and drinking water
Biomonitoring blood serum information with geography
Health endpoints with geography

### *Lessons Learned*

***The potential of the NSDS.*** The NSDS could enable easier and timelier data linkages between environmental quality information and restricted health information with geographic identifiers. These linkages are often essential for high-quality causal research exploiting natural experiments that arise from quasi-randomized fluctuations in exposures across individuals. An NSDS could also help federal agencies partner with states. Supporting vertical integration of federal and state information as well as horizontal integration of information across federal agencies could lead to tangible public health effects including:

- More timely, actionable, and policy-oriented research,
- Policies that are more responsive to local conditions, and
- More effective local health or hazard interventions.

***Powering environmental health research.*** PFAS in drinking water is an important topic of study, but it is only one example of the many pressing questions on the potential health implications of exposure to these substances. However, advancements that facilitate data linkages to support this topic could be used to explore other important questions, such as the impact of inhalation exposure to PFAS as well as dermal exposure through soil contamination, which disproportionately impacts young children. Further, environmental topics not related to PFAS could also benefit from the data infrastructure of an NSDS in similar ways. Additional topics of interest that may equally benefit include the effects of underground storage tanks on noncommunity drinking water systems and the cognitive impacts of pesticide exposure. Each of these projects would expand knowledge on the human health impacts of environmental quality and lay the groundwork for better policies and interventions.

***Advancing the science of exposure to environmental hazards.*** The focus of this use case is the direct health implications of exposure to environmental hazards, but these health effects have consequences for other dimensions of human welfare that can be difficult to quantify without restricted-use individual data. For example, air pollution exposure has effects on cognitive performance with potential implications for labor productivity, occupational choice, educational attainment and expenditures, and mental health. An NSDS that reduces the barriers to linking restricted-use data assets could benefit this related work. Researchers could leverage restricted-use Census Longitudinal Employer-Household Dynamics data or Bureau of Labor Statistics working conditions data in tandem with environmental quality data to provide greater clarity on the intersection of environmental and economic burdens.

## Data Inventories and Metadata

### Champion: Elisabeth Kovacs, *ACDEB Member*

### Project Description

The use case reviews current and evolving approaches to creating data inventories and managing metadata while considering how such tools could enhance and facilitate evidence-based decision-making at federal, state, and local levels. This includes investigating improvements for the current evidence-building ecosystem and weighing the possibilities of a National Secure Data Service.

The primary model of focus for the use case is the U.S. Chamber of Commerce Jobs and Employment Data Exchange (JEDx).

### *Rationale*

More specifically, the use case explores the following:

- The value of data inventories and metadata in support of data access, linking, and analysis for data providers and users;
- The current and emergent state of play, including potential data sources and challenges; and
- Lessons learned, including opportunities for the NSDS.

### *Background*

JEDx is a data standards-based approach for how employers can produce enhanced and more timely data on both jobs and employment. JEDx is a unique opportunity to modernize America's workforce through a national public-private partnership and data trust.

One of the public's most significant knowledge gaps is accurate, timely, and trusted data for the dynamic U.S. economy. As the United States emerges from a historic economic downturn and seeks to put Americans back to work, there is a critical need for: (1) improved labor market information, and (2) enhanced employment data for evidence-based policymaking and the administration of government programs (for example, unemployment insurance, or UI).

Through JEDx, the Chamber of Commerce Foundation has assembled a unique coalition of state and national public and private sector partners, stakeholders, and leaders that stands ready to close this gap on a national scale. JEDx will begin testing data and use cases as early as 2022 with the goal of forming a public-private data trust by 2024.

The vision of JEDx is (1) to streamline and improve how employers report data to government agencies, (2) to produce better longitudinal data about jobs and employment to power new workforce analytics while protecting privacy, and (3) to empower Americans with data and trusted records to verify their work history as well as their eligibility for government benefits.

### Progress

***Value proposition.*** The JEDx program has the potential to reduce costs and create higher value for stakeholders. For more information on the benefits for workers and learners, education and workforce partners, government agencies, and employers, see ACDEB's Year 1 report.

***Data sources.*** The initial focus of the project is on data collection, aimed at improving federal and state unemployment insurance reporting processes. Businesses are required to report similar, but not always identical, data to multiple and distinct agencies for different purposes. The hypothesis is that if agencies could align on a standard set of data to serve their purposes, then payroll processing companies could integrate those standards into payroll systems for both large and small businesses leading to more efficient reporting and higher quality data.

***Plans and priorities.*** Recognizing that getting 50 states on board at once would be a significant challenge, JEDx is testing this standards-based approach with seven states (Arkansas, Colorado, Kentucky, Texas, California, Florida, and New Jersey) to develop common definitions and to begin building the system architecture. The pilot focuses on the unemployment insurance program, given the high visibility of these data during the COVID-19 pandemic and generally high value as economic data. These data provide a good test case because of the lack of standardization for definitions as fundamental as what a "job" is and the systems involved across states. Today, employers and data processors report to dozens of different systems, each coded differently, so there are great potential efficiencies and savings from consistency across jurisdictions. In addition, there are current unemployment insurance system modernization efforts and investments underway that align well with the goals of the project.

The initial JEDx efforts are as much about the process as the actual results. The project seeks to identify compelling use cases, to target leaders and advocates in states to participate (public sector and coalitions, private-public), and to develop priorities for data uses and system architecture elements. The goal is to stand up an ambitious project that achieves consistency across the seven interested states and their programs and lays the groundwork for a larger constituency around data system improvement.

As such, JEDx is currently in the coalition-building phase. Project leaders are working toward consensus on the specific data elements, focused initially on a small set of data elements where they can find consistent definitions. From there, the project will consider other important variables (for example, occupation, demographics, and equity and inclusion).

### Challenges

Key challenges include the following:

- Businesses are often not interested in sharing more data. It is difficult to get buy-in, especially from smaller employers who self-report payroll and human resources records to government agencies.
- The unemployment insurance systems that are used as the basis for current data collection are not designed to capture more data easily.
- Standardizing and enhancing data collection is difficult. Currently, there is limited access to and use of high-quality job benchmark data and dynamic open competency and skills frameworks. Additionally, there is lack of systems interoperability and data sharing for more effective feedback and advanced data analytics.

### Lessons Learned

***The power of partnerships.*** JEDx is exploring how states, technology partners, and employers can improve data and evidence for decisionmaking. This includes implementing standards and sharing employment data, potentially through a public-private data trust and shared services. Data sharing under this partnership would not come from mandates but from agreements among peers—this is a new take on an old problem.

***The importance of data standards.*** JEDx is built on a standardized format for how data are organized, so they can be shared, compared, and discovered. Data standards allow for data to be organized and compared in citizens' daily lives.

***If a first you don't succeed, try, and try again.*** JEDx is positioned to re-evaluate, re-define, and develop a comprehensive set of end-to-end use cases to align leading public-private practices that have the potential to achieve results and potential solutions for both parties as it relates to capacity, time and resource commitments, and the available data and technology standards.

# 2. ACDEB Virtual Site Visit Summaries

As part of the information-gathering process during Year 2, ACDEB sponsored two Committee-wide virtual site visits—one focused on America's DataHub Consortium (ADC) and another on the Standard Application Process (SAP). Members raised many questions on these topics during and after the January 2022 ACDEB meeting. These field trips provided another opportunity for members to engage with relevant outside experts.

Table S2 provides an overview of each site visit. The Committee would like to thank all speakers and support staff who made these events a reality. This section provides summaries of the site visits. The information shared during these visits does not reflect the views of the full Committee. In addition, the summaries do not reflect changes to the models and examples that may have occurred after the dates of the site visits.

**Table S2. Virtual Site Visits Overview**

| Virtual Site Visit | Host(s), Speaker(s), and Support Staff |
|---|---|
| **America's DataHub Consortium**<br>February 18, 2022 | Speakers: Vipin Arora (NCSES), Keith Boyea (NSF), Dolly Pelto (ATI), John Finamore (NCSES) |
| **Standard Application Process**<br>March 3, 2022 | Speakers: Vipin Arora (NCSES), John Eltinge, John Finamore (NCSES), Alex Marten (EPA), Spiro Stefanou (USDA-ERS) |

| | |
|---|---|
| ATI | Advanced Technology International |
| EPA | Environmental Protection Agency |
| ERS | Economic Research Service |
| NCSES | National Center for Science and Engineering Statistics |
| NSF | National Science Foundation |
| USDA | United States Department of Agriculture |

# America’s DataHub Consortium (ADC)

The ADC site visit focused on a series of targeted discussion questions, with time at the end for open Q&A. This section presents several key takeaways from the site visit, as well as the discussion questions and related responses.

## Summary

### Focus Topic 1: What is America’s DataHub Consortium?

**What is a consortium?**

Broadly, a consortium is an association of entities that work collaboratively to do things better, faster, and more innovatively than any entity could do individually.

Specifically, the ADC offers a flexible acquisition path that allows for the delivery of solutions.

**What is ADC’s purpose?**

The ADC is an enabling mechanism for evidence building that supports and enhances the National Center for Science and Engineering Statistics (NCSES) and the National Science Foundation (NSF) missions and adds to existing data infrastructure and the broader data ecosystem.

**What are the roles of various ADC entities?**

Here are descriptions of the key components of the consortium:

- ***Members.*** Organizations (academic, private firms, nonprofits) who do the work of finding solutions to pressing challenges.
- ***Projects.*** Any issues or challenges that need solutions.
- ***Government Project Management Office (PMO).*** Government entity (NCSES) that oversees operations from the government side.
- ***Consortium management firm.*** Entity (currently, Advanced Technology International) that provides the infrastructure and sets the stage for the members doing the work. A good analogy is that the entity builds the “stadium,” and consortium members are the “players” on the field.

**What are the phases of ADC’s work? What are some examples of sponsored projects? What sorts of questions is the consortium hoping to address?**

The progress of the ADC is phased to allow for revision and modification, leveraging lessons learned from successes and challenges. For transparency, these considerations will be made public. Each phase features different priority considerations (overarching emphases like technology or legal requirements), focus areas (areas to learn about through projects being done, not dependent on the subject matter), and project requestors. Here is more information on each phase:

- ***Phase 1.*** NCSES and NSF are requesting projects through the ADC and prioritizing technology with a focus on data linkage. For this phase, the ADC issued the first request for solutions and will make multiple awards and announce them to the public.
- ***Phase 2.*** The ADC would like other agencies (state and federal) to request projects; the goal is to incrementally grow the consortium through joint projects with NCSES or NSF or by other agencies using the ADC themselves.
- ***Phases 3 and 4.*** Individuals could request projects, and the ADC would focus on aspects like governance, sustainability, and infrastructure. This is another area where ACDEB could provide insights.

### Is the ADC for America or only for ADC members? Can individual academics join? Could you give an example of how a private citizen could use ADC?

Anyone could join the ADC, and anyone could create project proposals; this includes individual academics and private citizens. As far as serving the public more broadly, the Committee could provide guidance around the larger organizational structure needed to get from Phase 1 to Phase 4 (where questions from private citizens could be answered). There are issues around sustainability and resources; the ADC will have clear criteria and evaluators for assessing projects to fund and problems to solve, and there will be an advisory board to help guide this process.

One analogy for the ADC may be as an "Uber" for evidence building, matching decisionmakers with questions to groups who can help them find solutions. Of course, there are bureaucratic caveats like resources and sustainability. Another challenge is getting new organizations involved that are otherwise not working with NSF. The relational contracting approach available under the "Other Arrangements" Authority reduces bureaucratic demands and allows for greater participation.

## Focus Topic 2: What are the key implications of NSF's "Other Arrangements" Authority?

### Is this authority mainly a contracting vehicle or something else?

There is a contracting component with this authority, but it is much more than that. Traditional contracting requires services to be defined upfront, which ignores the fundamental nature of long-term service contracts. Alternatively, this authority allows for relational contracting, which emphasizes the need to establish a working relationship between the parties that attracts nontraditional organizations, fosters ad hoc specification development, allows for joint budget management, and provides flexibility to change performance priorities. The benefits of relational contracting are impossible to achieve using traditional methods of contracting.

### How does this authority help define the structure of the ADC (both now and in the future)? What are the lessons learned that could apply to the structural options for the NSDS?

The "Other Arrangements" Authority allows for many structures; the hallmarks of this model are flexibility, prioritization of relationships, and the ability to reach nontraditional organizations.

Future characteristics of the ADC could include the following:

- Establish R&D proposal evaluation criteria that encourage and prioritize projects with data sets that can be shared inside and outside the ADC.
- Foster collaboration within the ADC through outreach, marketing, and education efforts.
- Make ADC membership more valuable than just allowing access to NSF research funding; membership would enable access to cutting-edge data sets and methodologies unavailable elsewhere, provide a mechanism for state and local government investment, and allow individual researcher access.

The ADC is an approach that allows for scalability and sustainability, and the consortium will continue to consider these pieces moving forward.

### Focus Topic 3: How does the ADC connect to other parts of the evidence ecosystem?

#### How do other parts of the statistical system tie into this process? What about state and local governments? Could state and local decisionmakers get help answering their questions even if they do not have a funding mechanism or resources to bring to the table?

The ADC is a part of a larger ecosystem that extends to federal agencies as well as state and local governments. As such, the consortium seeks partnerships with the statistical system, other parts of the federal government, and state and local governments—as consortium members, funding sponsors, and engaged stakeholders.

The ADC wants states to participate in all phases and in a variety of roles:

- State and local collaborators can choose to be funding sponsors for projects proposed to the consortium.
- There are options for states looking for funding for their projects; state leaders could use the ADC to help fund a project of their own, a state could partner with NCSES on a project, or a state could partner with another federal agency.
- State and local partners could also provide insights into the process, as there is great value when as many people as possible contribute to a given conversation.

#### How does America's DataHub connect with NSDS? A pilot to inform, a building block, or another organization NSDS could coordinate with?

This is another area where ACDEB could weigh in. The ADC could be a pilot and serve as a good place to experiment with different ways to approach a project.

#### How does this fit in with the legislation being considered about a pilot NSDS?

The ADC can be considered a demonstration project for the NSDS or could support a demonstration project for the data service. There are different options depending on what the Committee prioritizes; the flexibility, speed, and scale of the ADC offer opportunities to experiment.

## Focus Topic 4: What about privacy and confidentiality?

### Will the DataHub include any information that will require privacy or intellectual property protection? How is the consortium approaching security and privacy? What about considerations for Tier 1 governance and Tier 3 information and information systems?

The ADC is not a warehouse and uses the existing data ecosystem and infrastructure available to complete projects. These projects could use information that requires privacy protection, but the consortium is not housing data—in that sense, the ADC is a service.

The ADC explores different issues to help build data for evidence-building purposes. Built into that is the need to understand security and privacy issues, the presumption of accessibility, implementation of section 3582 of the Evidence Act, etc.

"Evidence building" is a term of today, but NSF has been working to make data accessible and secure for a long time. NCSES's CIPSEA authority makes it possible to securely provide access to data. In addition, NCSES has several existing platforms to provide secure access (e.g., being an active participant in the Federal Statistical Research Data Center (FSRDC) system, a secure access data facility, and a sponsor center).

## Open Q&A

### What about replication studies? What would that look like?

The consortium will need to develop criteria for how to define an "influential study." At the individual study level, the ADC could do a series of these replication studies. The critical part will be deciding how to choose specific studies and the ability to access the infrastructure and data that are part of those studies. This is one place ACDEB could play a role (in developing criteria and setting priorities) and where CIPSEA status comes in (for access to data through statistical agencies likely needed for these studies).

Two concrete examples are the following:

- Consider how to compare results from one area to another. For example, researchers cannot access raw Census Bureau data but have public-use files and are starting to answer questions in the ADC using those data. The ADC would want to understand how those procedures would look if using raw data; someone working in an FSRDC with those authorities could replicate the findings.
- It may be good to have a sustainability period where data stays in the system for a certain amount of time for certain projects to be reconducted. If there is documentation for a project, could another user leverage that information to produce the same results? The ADC would need to consider how long data would be kept in order to reproduce results for verification, given that the ADC is not a data warehouse.

### At full scale, how many projects could the ADC handle?

Potential is only limited by the capacity of members. This model feeds itself because more customers create more money, which brings in more members, and so on.

### Does the ADC currently have access to CIPSEA-protected confidential data?

The project determines the application for data access. Federal data infrastructure is in place, and any project that wants to use federal resources will apply for access. Eventually, this would be through the Standard Application Process, but right now, a potential user must submit individual requests for access with each agency.

# Standard Application Process (SAP)

The SAP site visit focused on a series of targeted discussion questions, with time at the end for open Q&A. This section presents several key takeaways from the site visit, as well as the discussion questions and related responses.

## Key Takeaways

- OMB and the ICSP agree with the ACDEB Year 1 report's dual emphasis on (1) Evidence Act implementation and (2) envisioning an NSDS; they see these items as interdependent. To progress in these areas, OMB is working to implement Title III of the Evidence Act (i.e., CIPSEA 2018); this includes thinking through problems and challenges and seeking advice from the Committee.
- It is important to put the SAP in a broader context; CIPSEA 2018 highlights the philosophy and service aspects of an "NSDS 1.0."
    - As a "philosophy," the SAP is a customer-centric one-stop shop for access to data for evidence building.
    - As a "service," the SAP is being built by and within the statistical system today because that is the legal requirement. Looking to the future, the ICSP is proposing that the SAP not only be a place for accessing statistical data but also for administrative data.
- There are several SAP workstreams around which the ICSP plans to engage ACDEB's subcommittees in an iterative and ongoing conversation: policy, technical development, technical assistance, and stakeholder engagement.

## Summary

### Focus Topic 1: The SAP and the Evidence Act, John Finamore

**Under the Evidence Act, what are the requirements for building the SAP, and who's responsible?**

Section 3583 of the Evidence Act requires OMB to establish the SAP; OMB works with the ICSP on the implementation of that requirement. The law requires each statistical agency or unit to establish an identical application process—not just a form but criteria for determining outcomes related to applications, timeframes for prompt determination, an appeals process in the event of an adverse determination, and reporting requirements to ensure full transparency. This is a service for the public as it relates to restricted data and requesting access to data, so transparency is extremely important.

**How does this connect with other requirements of the Evidence Act, specifically, the Committee's role?**

The SAP is a service that enables the discovery of data for evidence-building purposes. This process is built by and within the statistical system, but the long-term goal is to offer a place where individuals can discover data that are statistical and administrative. The Evidence Act created ACDEB to advise OMB on implementation of Title III, and the ICSP views the SAP as a building block of the NSDS—or "NSDS 1.0."

## Focus Topic 2: SAP Policy, Alex Marten

### What is the relationship between the data inventories for the SAP and data.gov?

The SAP data inventory is one component of the SAP policy. Implementation of the data inventory is not explicitly required by statute, but it became clear that there is a need for such an inventory within the SAP. There needed to be a way to signal what is available through the SAP—to make data *discoverable*. From the user's perspective, going to the portal and seeing a blank field ("what data asset would you like?") didn't seem ideal.

Data.gov has data assets for which people aren't going to be able to apply for access through the SAP. Additionally, some data fields for metadata specific to the SAP won't be available through data.gov (e.g., information on requirements for being credentialed for a data set or U.S. citizenship).

Agencies cannot afford to have redundancies in the system or to try to build the same thing twice, so there is a concerted effort to ensure that whatever is in the SAP data inventory is an expansion of what is in the broader data catalog and is not a duplication. The goal is for the SAP data inventory to be an extension of data.gov (and whatever the next generation of data.gov is), so agencies are maintaining extra fields in their metadata catalogs to support the SAP inventory that are not submitted to data.gov.

### Are there formatting or syntax differences between the SAP inventory and data.gov?

The goal is not to introduce new steps that add burden to agencies, as resources are a concern. The SAP is a work in progress, and the ICSP has been talking to Chief Data Officers (CDOs) about inventories of the future so that different inventories can evolve in parallel.

### Will the SAP include data from state and local governments? Would it be a one-stop shop in that way?

This is not a focus for this initial phase, as the emphasis is on meeting the statutory requirements; however, the ICSP is designing the SAP for flexibility and the ability to grow and evolve.

### Common data standards across levels of government could help make more data available for evidence building. Are there lessons learned from the SAP experience that might be informative for what kinds of standards ACDEB might want to recommend?

The ICSP is in close contact with CDOs (and the CDO Council) as they're trying to respond to their statutory requirements under Title III of the Evidence Act. The groups are sharing experiences as things evolve, so there aren't dual standards. While the ICSP is not yet coordinating outside the federal government, this should remain on the radar.

Discovery is at the heart of what the SAP is striving to become—the concept here is "discovery" metadata. Metadata can mean different things, but the SAP's focus is on data that would help users figure out what data are available that might meet their needs. That helps with standardization by providing a standardized way of doing discovery.

The goal is to provide information that researchers can use to identify data that meet their needs for evidence building or research. There is a connection between discovery and metadata and synergies around communication, coordination, and efficiencies within and across agencies.

It would be helpful to look at existing standards rather than creating new ones. The statistical system has for a long time participated in creating standards that are common for statistical data, so current efforts don't have to start from scratch.

From the agency perspective, the ICSP would strongly support a recommendation that most metadata reside on agency sites and that systems (e.g., the SAP or the NSDS) direct users to those sites. This speaks to the need for standardization so that users have a seamless experience. In addition, it is costly for agencies to maintain multiple versions of their metadata across multiple sites—minimizing duplication requires uniformity.

**The data.gov experience isn't particularly user friendly or organized as well as it could be. What are the lessons learned that could be used to inform other efforts?**

The current data.gov is not ideal; however, conversations with the CDO Council on inventory and metadata work are very forward-looking. CDOs are thinking about the data.gov of the future. The SAP efforts are aligned with this vision but purposefully separated, respectful of legal requirements, and focused on the user experience of accessing statistical data.

Again, the SAP data inventory is not required statutorily but is being built because of the need to make something accessible and usable for the customer. From the beginning of the SAP and throughout its development, the view has been of a customer-focused service built through user interaction and user experience. So, the process has included a pilot test, talking directly to potential and actual applicants, and gathering feedback to make sure the discovery process works for them.

**Does the SAP data inventory include data that users don't need permission to access?**

The inventory reflects restricted-use data for which users must apply to be given access, as well as a listing of related public-use files to make users aware that there are other data available that may meet their needs without going through the review process. Metadata listings are provided for the restricted assets only.

### Focus Topic 3: Transparency in Reporting, Spiro Stefanou

**How will the SAP help agencies assess the return on investment in data resources and where future investments would be the most valuable? Would this look like a promise to report on data uses (i.e., papers written and published, policy applications, etc.)?**

Transparency in reporting is a critical element for the success of the SAP. The SAP's focus on transparency shows up in the discovery process of restricted-use data, not of outputs. When comparing existing methods of reporting, citations are most common but quite difficult to implement. For example, the Bureau of Economic Analysis and IRS have research paper programs, and FSRDCs feature working papers. This type of reporting isn't within the scope of the SAP as it is currently set up. In the future, the SAP could automate citations by requiring a tag or Digital Object Identifier that could be tracked. This is part of the evolution of transparency in reporting.

## Focus Topic 4: Stakeholder Engagement, Vipin Arora

### What does "expanded" stakeholder engagement mean?

Lots of time and effort across the ICSP has gone into thinking about stakeholder engagement and how to move it forward—the communication aspects sometimes dwarf the technical aspects. There are several streams for stakeholder engagement:

- ***Public website and e-blasts.*** To be as transparent as possible, the ICSP is working on a public website that gives people information on development and goals; this is coupled with e-blasts, but it takes time to build a stakeholder list.
- ***Engagements with outside groups.*** Presentations like this one, or talking to CDOs or agencies, are very helpful. The ICSP would like to engage more with state and local governments.
- ***User testing.*** User testing has been important, and the ICSP hopes to continue and broaden these efforts in the next development phase.

The SAP provides a way to think about the user experience—what's working and what isn't on the portal and what could be available within a couple of clicks to an agency's website. So, it's important to learn from the experience of people trying to use the portal to build something more robust. For example, agencies want to be able to provide more in-depth technical assistance to users but are concerned that they could be overwhelmed with customer requests. Considerations like these impact future policies and service levels of the SAP and could align with ACDEB's recommendations around a data concierge.

### What role can states play in the development of SAP policies?

ACDEB, especially the state and local representatives on the Committee, can advise OMB and the ICSP on what a roadmap would look like to incorporate the states.

### How is the ICSP reaching out to academics across the country, especially those from historically Black colleges and universities, to make them aware of the SAP?

The ICSP is targeting associations, as researchers are so dispersed. ACDEB could provide advice on which groups to focus on or other outreach mechanisms.

### Is researchdatagov.org the site that the pilot was performed on? How far from "final" is the version that is publicly available now?

The pilot (Phase 1) used this website and focused specifically on assets available through the FSRDC system (not through single agencies). The SAP Phase 2 covers all federal statistical assets and all secure access facilities, including individual agencies and FSRDCs. The website is not the same but will be transitioning to Phase 2. Currently, metadata inventory includes 900 assets, and this will be built out incrementally. Changes to both the metadata inventory and the application process are coming in Phase 2.

It is important to keep in mind that the focus of Phase 2 is discovery and application. To build beyond this, the access regulation is in some ways the next "car in the train" in terms of how these pieces fit together. This regulation will create a standardized way of classifying data sensitivity and creating access tiers that correspond to sensitivity levels. Often when stakeholders talk about the SAP, they skip over the approval part to get to access (or the outcome of the process).

### What ensures that all federal agencies participate this the SAP?

The legal requirement is for all recognized statistical agencies and units to participate. In Phase 1, this applies to the 13 principal statistical agencies and three additional units. Once the policy is finalized, it will be a requirement.

### Timeframes for applications approvals seem long (3 months, 6 months). Is this because it is a preliminary estimate, or is this a staffing issue?

There is a timeframe written into the policy, but it can be updated over time as agencies realize the efficiencies of having a standardized process. As a starting point, these timeframes are based on an assessment of how long it typically takes to get through the agency approval process. This reflects current realities and practical limitations of the user experience, but, in the future, agencies will push to work under a timeframe that is more ambitious.

This is also a staffing issue, so it would be helpful to see a recommendation from ACDEB to properly staff agencies and dedicate more resources to support new uses of data for evidence building in order to carry out these new responsibilities under the Evidence Act.

### If a user requests data from six different agencies, will it take 24 weeks, or 72 weeks (12 times 6)?

The 24-week timeframe applies; however, agencies can get extensions for timeframes for complex cases, and timelines aren't applicable to state agencies since OMB cannot enforce policies at the state level.

### Each new application seems to be viewed as something completely different than before, which requires in-depth analysis even though projects are often variations of each other. Could the federal government adopt standards so the SAP does not have to do an in-depth review each time? The NSDS should help cut through bureaucracy, so how does the Evidence Commission's vision translate into the Evidence Act and to where things stand now?

Currently, the SAP is moving from Phase 1 to Phase 2, and there will be many more phases. The ICSP is starting with the legal requirements, which is a big step to get all 16 main statistical agencies and units to do this core function (of providing access to confidential data) in a new way.

The 12-week timeframe for application approval is a max, and the process could move faster. The SAP has two components to help with efficiency—approving the person and approving the project. The person could be approved in a way that allows them to be involved in multiple projects without having to redo paperwork, and the project could be approved so that more than one person can work on it. The project management office can also help with efficiency in the future.

Standardizing the application process reduces redundancies, and the statute did not change the fact that the ultimate responsibility and liability for ensuring confidentiality remains with the statistical agencies who are providing access to data. These agencies still have an obligation to review and approve applications; however, the SAP helps ensure that standardization within approval criteria, which allows for reciprocity and will help reduce burdens for agencies and applicants over time.

It is important to keep in mind that CIPSEA 2018 is like a jigsaw puzzle; one piece is the application process, another piece is about data sharing, and another piece is about standardization. The goal is to create a system that interlocks these three pieces.

### Focus Topic 5: The SAP as the Foundation for an NSDS, Vipin Arora

#### How could the SAP fit into the broader structure of a possible NSDS? It's a possible "front door," right?

The Evidence Act requires OMB to establish the SAP that must be adopted by all statistical agencies as their sole means of accepting applications. A fitting analogy is that the SAP is the single "front door" to gaining secure access to the statistical system's confidential data assets.

The current version of the SAP is the initial way that OMB and the statistical agencies are complying with the statutory requirements. Like any system, the SAP will continue to evolve. For example, there is an expectation that nonstatistical agencies will use the SAP as well. If an NSDS is funded, the SAP would transition and expand to fit the roles of this new entity. Either way, the SAP is the federal government's first step toward a truly standardized process and represents a practical step from which to build in whatever direction comes next.

The SAP is being designed from policy, implementation, and governance perspectives—all viewed with the recognition that while the data ecosystem is evolving, OMB and the ICSP must move forward in a practical way to meet this statutory requirement in the short term. The ICSP is asking for ACDEB's advice on how it sees the SAP fitting into an NSDS and whether that relationship would have implications for how to develop these initial, practical phases.

#### What are some other questions on the SAP that ACDEB could provide advice on to OMB and the ICSP?

The ICSP identified the following areas where it would be useful to get advice from the Committee:

##### *Policy*

- What role would the Committee like to see state and local governments play in the development of policies for the SAP?
- OMB is currently asking for feedback from the public on a preliminary plan for implementing the SAP through a Federal Register Notice (FRN). How would ACDEB address OMB's request for feedback on the FRN, specifically regarding metadata standards, application windows, application evaluation, appeals process, and public reporting?

- What would be the best vehicle for maintaining reporting transparency? Is there a preferred method other than citations?

### *Parallel activities in the federal statistical system*

- What initiatives currently conducted by the federal statistical system can inform the Committee's work?
- What role would ACDEB like to see the SAP play in the broader structure of a possible NSDS?

### *Stakeholder engagement*

- What recommended strategies should the ICSP implement to expand stakeholder engagement (in addition to current efforts?
- What should the playbook look like for engaging with nontraditional user groups and related communities?
- The ICSP recognizes that there are many stakeholders who want a seat at the table. How can the ICSP gather broad stakeholder feedback while respecting the inherent prioritization that comes with working with both federal and public stakeholders?

# 3. OMB/ICSP Workstream Reports

At the January 2022 ACDEB meeting, OMB and the ICSP provided the overview for iterative engagement with the Committee. These "workstreams" offered the Committee the opportunity (1) to provide feedback on federal statistical system initiatives that were already in progress and (2) to use these existing efforts to inform the Committee's work. These two pieces support the Committee's charge to advise OMB on CIPSEA (Title III of the Evidence Act)—both on the "here and now" and on the target vision for the National Secure Data Service.

On any given workstream, there may have been a series of discussions with the primary subcommittee (or members thereof) and other members, as appropriate. As each conversation evolved, members provided preliminary input through discussion, answers to targeted questions, written comments, or other informal mechanisms. This initial input filtered through to the findings and recommendations presented in ACDEB's Year 2 report.

Table S3 presents the OMB/ICSP workstreams, provides a brief description of each topic, lists the "assigned" ACDEB subcommittee(s), and notes the ICSP leaders for each workstream. The Committee would like to thank OMB and the ICSP for this engagement. This section provides summaries of the OMB/ICSP sessions with ACDEB's subcommittees, including key takeaways. The information shared during these sessions does not reflect the views of the full Committee. In addition, the summaries do not reflect changes that may have occurred after the meeting dates.

### Table S3. OMB/ICSP Workstreams

| Workstream | Description | ACDEB subcommittee | OMB/ICSP leaders |
|---|---|---|---|
| Access and Confidentiality Initiative: Regulation | • Elicit feedback on goals, principles, and frameworks.<br>• Build shared understanding of challenges.<br>• Identify innovative ways to meet the goals and overcome challenges.<br>• Includes data sensitivity levels that are also a part of the new Zero Trust policy development in which the CDO Council is engaged. | Legislation and Regulations | Lead Champion: Spiro Stefanou<br>Supporting: Shelly Martinez |
| Access and Confidentiality Initiative: Methods Coordination | • Gather input and feedback on priorities to build out Data Protection Toolkit.<br>• Validate what ACDEB is looking for in case study narratives/pilot goals that cover value-driven projects and privacy-preserving technologies.<br>• Provide feedback and validation on methods coordination activities and how an NSDS could address those activities. | Technical Infrastructure | Lead Champion: Barry Johnson<br>Supporting: Michael Hawes |
| Stakeholder Engagement: Access and Confidentiality and Standard Application Process Initiatives | • Solicit advice on the essential elements of a successful engagement plan, especially how to effectively engage nonfederal users.<br>• Ask for feedback on key stakeholder messaging.<br>• Inform ACDEB on stakeholder engagement work and gather consensus on messaging that highlights data access, linkage, and data protection methods.<br>• Share best practices on agency collaboration.<br>• Gather suggestions for continued synchronized messaging. | Other Services and Capacity-Building Opportunities<br>Supporting: Government Data for Evidence Building | Lead Champion: Bill Beach<br>Supporting (Access and Confidentiality): Guadalupe Cerritos<br>Supporting (SAP): Vipin Arora |
| Standard Application Process Initiative: Technical Development | • Gather input on the technical requirements to develop and implement the process required under Section 3582.<br>• Discuss initial technical requirements for the SAP, including the vision for metadata and how tiered access requirements relate to SAP implementation. | Technical Infrastructure | Lead Champion: Barry Johnson<br>Supporting: Shelly Martinez |
| Standard Application Process Initiative: Governance | • Seek advice on the governance structure that the ICSP is looking to put into place prior to the launch of Phase 2 implementation, consistent with the draft policy proposal (FRN). | Governance, Transparency, and Accountability | Lead Champion: Brian Moyer<br>Supporting: Alex Marten, Spiro Stefanou |
| Standard Application Process: Technical Assistance | • Gather feedback on how to incorporate more user support features in future releases. | Other Services and Capacity-Building Opportunities | Lead Champion: Barry Johnson<br>Supporting: Shelly Martinez |
| Other Items | Other items upon which advice could be useful and for which OMB/ICSP can provide topics/questions if ACDEB is interested in engaging:<br>• Responsibilities of Statistical Agencies (aka Trust) regulation<br>• Chief Statistician Priorities<br>• FY 23 Appropriations<br>• Standards Setting | Legislation and Regulations<br>Legislation and Regulations<br>Government Data for Evidence Building<br>Governance, Transparency, and Accountability | Supporting: Dominic Mancini, Shelly Martinez, Robert Sivinski |

# Access and Confidentiality: Regulation

## Primary ACDEB Subcommittee: Legislation and Regulations

### *Key Takeaways*

- A guiding principle of the regulations is that more data should be made available, not less.
- Another guiding principle is that disclosure risk is on a continuum and is not binary.
    - The idea of middle tiers of access helps create flexibility to move beyond a binary approach where disclosure is either risky or not.
    - Synthetic data and secure multiparty computation are great examples of middle tiers of access. In addition, query systems should also be part of the conversation.
- The ICSP is looking to ACDEB for feedback on the following:
    - Where are the federal statistical system's greatest opportunities moving forward?
    - What are opportunities to implement shared responsibilities between data users and federal agencies to ensure confidentiality while efficiently using resources?

### *Summary*

#### What are the goals and guiding principles of the regulation?

In keeping with CIPSEA 2018, OMB must promulgate a regulation to safely and securely expand access to the data assets of statistical agencies and units while protecting such assets from inappropriate access and use. This regulation must include standards to assess data assets in terms of sensitivity level, corresponding level of accessibility, and whether less sensitive versions can be created. The standards will be designed to improve access and will contain requirements to conduct re-identification risk assessments and make processes transparent and easy to understand. In a sense, the regulation will implicitly set out a risk management framework, covering risks such as disclosure risk, re-identification risk, and reputation risk.

The general guiding principles include the following:

- More data should become available, not less.
- Disclosure risk is on a continuum and is not binary.
- Not all data are equally sensitive.
- Want to be consistent with other efforts, addressing both current and future data.
- There is shared responsibility—agencies and users share responsibility for protecting and not disclosing data.
- Protect good faith actors.
- Emphasis on linked data, more data coming into the statistical system, and creative activity around nontraditional data sets.

- Consider multiple audiences in determining sensitivity.
- Conduct risk assessments for re-identification risks.
- Improve access to rich data in administrative records.
- Cost-effectiveness.
- Consider agency-specific privacy laws.

OMB believes these principles push some groundbreaking ideas and reflect a commitment to advance major goals through the drafting of the regulation. For example, most agencies today operate in a binary mindset where data are either accessible or not, which drives risk-averse behavior. Viewing disclosure risk as a continuum is a considerable step toward looking for acceptable levels of risk and freeing agencies of the situation where the law is interpreted as requiring risk to be zero. The idea of shared responsibility reinforces this principle. If there is a way for data users to share the responsibility not to re-identify those data, that could further buffer against the need to have zero risk. Additionally, the principle of protection for good faith actors encourages creative thinking, so making that protection explicit is essential. OMB/ICSP would like to explore models for shared responsibility and welcomes ACDEB's suggestions for how to accomplish this. In addition, OMB seeks input on the implications for reputation risk and trust.

The goal is for the eventual regulation to reflect the guiding principles. In the end, the public should be able to reconstruct these principles through the text of the regulation—the principles should be clear and consistent. To achieve this, OMB and the ICSP are seeking feedback from the Committee on the goals, principles, and frameworks. The Committee can inform the following questions:

- Where are the federal statistical system's greatest opportunities moving forward?
- What are opportunities to implement shared responsibility between data users and federal agencies to ensure confidentiality while efficiently using resources?

### How will the application of a risk framework and the concept of tiered access help achieve the desired goals?

- Rather than categorizing risk into tiers, statistical agencies currently consider individual requests and evaluate what can be done to make the disclosure risk acceptable. One critique of this practice is that agencies miss cumulative risk effects when looking at data sets one at a time. Instead, agencies could use a composite approach that incorporates current requests and considers what might be released in the future. Part of the regulation is taking this broader approach that moves agencies to apply a common framework instead of considering individual releases. To do this, the ICSP must understand how to make risk analysis more systematic and how risks can be quantified and communicated.
- Agencies must move beyond a binary mindset where data either live in the open or are protected in an enclave. Using middle access tiers creates flexibility and moves the system beyond a binary approach. Synthetic data, secure multiparty computation, and query systems are examples of potential middle access tiers.

- A risk continuum is good in principle but is a complex concept to apply in the context of the public domain. For restricted access, agencies can analyze the chance that users will try to access more secure levels of data than those to which they have been granted access. Data use agreements can also include clauses that restrict reconstruction and re-identification. However, when applied to public data releases, it is more difficult to quantify the risk.
- OMB and the ICSP are thinking about the technical dimensions of risk disclosure mitigation. They are considering how to quantify risk and determine levels of tolerable risk. The regulation will set a policy framework around those kinds of decisions. OMB and the ICSP are thinking of risk in different parts—for access and output. For access, it is important to identify different levels of data sensitivity to align with different tiers of access; for example, data sensitivity “level 1” goes into access tier “level 1,” etc. For output, there will be different tiers for assessing risks, as well. For example, if an output is in a restricted tier, then appropriate disclosure risk assessments must be performed on it. The specific methodology will depend on the tier, as there may be different dimensions on which to apply different risk analyses.
- There must be policies to align different sensitivity, access, and output tiers and to help users determine which tiers best meet their needs. While it takes effort for statistical agencies, the data must exist in different tiers with different degrees of access. In addition, it is feasible that a single data set could exist in multiple tiers. The law refers to data assets rather than single variables, which means that a data asset could move from tier to tier by removing and synthesizing select variables, allowing for different degrees of access. For example, a user could go to a public site and get 5 percent of their data needs met, then use synthetic data with a validation server to get 80 percent more, and then only go to an enclave for the remaining work.
- OMB/ICSP is developing risk assessments with two components: the probability of an adverse event and the cost of adverse event happening. Even if there is certainty of such an adverse event happening, agencies can still assess the impact of that risk and determine if it is acceptable to bear. These types of robust risk assessment frameworks exist in other contexts—they are just new in this context. Ultimately, these risk frameworks will help the federal statistical system retain the utility of the data while moving beyond the notion that disclosure risk must be as close to zero as possible.

# Access and Confidentiality: Methods Coordination

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- Data releases inherently come with disclosure risk requiring a consistent approach to determining the acceptable level of risk and how to manage it.
- The Data Protection Toolkit is endorsed by the Federal Committee on Statistical Methodology (FCSM) and, as such, avoids commercial solutions.
- Shared responsibility can help mitigate the risk of data breaches.
- Privacy-preserving technologies (PPTs) may be able to be deployed in ways that reduce the movement of data while retaining the ability to learn from those data.

### *Summary*

#### What can be done to mitigate disclosure risk?

*Any* release of data carries risk of disclosure, so effectively releasing anything requires that you know what the risk is and determine if that level of risk is acceptable. Even tools like differential privacy can only reduce disclosure risk, but they cannot eliminate all risk.

There are two sources of disclosure risk:

1. Access to the personally identifiable information (PII) itself increases the risk that data can be hacked, and
2. The richness of the linked data also increases the disclosure risk and the likelihood that re-identification could happen.

There are many barriers to identifying and managing disclosure risk. These include:

- ***Legal.*** Historically, laws determine no risk as being acceptable.
- ***Standards.*** There is no governmentwide framework to help agencies decide how much risk is acceptable.
- ***Transparency.*** In the past, disclosure risk assessments have been subjective.
- ***Human capital.*** Agencies have tackled these issues on their own with limited collaborations.

A variety of solutions are available to help address those barriers and better manage disclosure risk. The statutory framework provided by CIPSEA 2018 lays the groundwork for identifying disclosure risk and assessing acceptable risk levels systematically. To build off that groundwork and create a common framework across government, OMB is currently developing a regulation on safely and securely expanding access to federal data assets while protecting such assets from inappropriate access and use. Key features of this regulation include frameworks for sensitivity levels, access tiers, and output controls. The common framework for tiered access will help guide federal agencies in better leveraging tiered access solutions, especially the “middle” tiers between open data (on the one end) and confidential microdata (on the other).

It would be beneficial to create a mechanism that allows agencies to piggyback on what has been done and better share solutions. As discussed in ACDEB's Year 1 report, these activities could fit within the coordination and capacity-building roles of the NSDS.

Emerging technologies are also promising. As noted in ACDEB's Year 1 report, the R&D function of the NSDS could advance work in this area. Synthetic data, query systems, secure multiparty computation (SMC), and differential privacy can all help manage disclosure risk, but none of these options offer off-the-shelf solutions today.

Agencies can improve risk assessments by reviewing existing approaches. The forthcoming OMB regulation described above will help agencies carry out objective risk assessments. Disclosure risk assessments give a starting point, so agencies must do "hacks" of their own products. But tools like re-identification risk studies can be resource intensive. For example, a recent study of re-identification attacks on the 2010 Census required 30 large nodes on Amazon Web Services for approximately 3 days at the cost of $20,000. Furthermore, making these studies a part of the decisionmaking process requires infrastructure, such as privacy staff, review boards, technology staff, and survey staff, as well as a potential need for hosting data, possibly at the NSDS, to facilitate these sorts of attacks.

### Can the Data Protection Toolkit provide agencies with useful support?

FCSM's Data Protection Toolkit provides a forum for resource and information sharing across agencies, so they do not have to reinvent the wheel when they want to improve on existing methods or explore emerging methods. The toolkit promotes collaboration on basic infrastructure needs and can help build toward the philosophy of the NSDS.

Initial components of the Data Protection Toolkit have been completed and are available today. However, effectively fulfilling its ideal role of embedding interagency collaboration on data protection and access challenges requires much more work. Completed and planned components of the Data Protection Toolkit include the following:

- Completed components:
  - Initial site architecture;
  - Inventory of 200+ tools, resources, and templates;
  - Primer on confidentiality and secure data access;
  - Commonly used Statistical Disclosure Limitation (SDL) techniques;
  - Tiered access mechanisms; and
  - Training modules.
- Planned components and features:
  - Best practice recommendations and recommended tools,
  - Emerging SDL methods,
  - Performing risk assessments,
  - Content curation by audience, and
  - Collaboration features.

### Are there ways that shared responsibility can be created for data breaches to help lessen their likelihood?

Policies around shared responsibilities for protecting and not disclosing data could be outlined in the forthcoming OMB regulation described above, and ACDEB could recommend that any attempt to disclose the data would be subject to serious financial penalties.

### How can privacy-preserving technologies (PPTs) be deployed?

The xD group within Census is conducting projects exploring whether researchers can answer questions using data they never "see." This group is testing a suite of PPTs to create data networks, reduce data sharing burden, and increase privacy federally, domestically, and internationally.

The technologies the group is testing have been around for a long time but have never been combined. Their initial assumptions include that:

- The suite of techniques can be combined to produce results that prove researchers can answer questions of a network of data without seeing it.
- Selecting an open-source technology stack will enable more partnership opportunities, enable easier technology analysis through transparency, and enable deployment at a much lower cost than bespoke or acquired solutions.
- They can start from zero trust and build trust over time by thoroughly vetting the technology solution and infrastructure and by increasing data sensitivity and volume as trust continues to build.

A pilot project with the United Nations (UN) includes five countries and is focused on imports and exports. The UN functions as a network node while each country operates as a domain node. The feasibility project uses open data intentionally. The project tests the highest level of security measures, using the Azure cloud, to see if that works. Since it is open-source, once users are cloud-ready, it becomes easier as they can pick it up without paying for software. The group is in the process of selecting an open-source solution, with the initial assessment determining that PySyft has packaged the most to offer.

There could be multiple network nodes. These nodes could govern different studies—it will start to be more like an APRAnet, where researchers can opt into different private data networks. The NSDS may be able to serve as a network node, with state and local users seeing value in being domain nodes to get reports without complicated data sharing and data moving.

Within the next few years, Census Bureau researchers expect a massive uptick in the maturity and use of PPTs. The idea is that it ultimately requires less movement of data, instead becoming a more federated and connected world where data owners stay data owners, thus increasing privacy.

# Stakeholder Engagement

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- There may be a perception gap about the effectiveness and value of current engagement and data partnerships between federal agencies and state and local governments.
- State and local government capacity is exceptionally varied; building an NSDS that is truly useful, usable, and used by all will require substantial investments in human capital.
- Engaging with the ICSP and seeing the Standard Application Process as a "one-stop shop" could be an avenue from which to grow stakeholder engagement.
- Engagement, messaging, and multi-channel two-way communications are key to future success.

### *Summary*

#### What are effective strategies for obtaining input from a diverse set of stakeholders?

Engagement, messaging, and multi-channel two-way communications are keys to future success. Federal agencies often start with a minimum viable product and need to conduct outreach to identify this product using few resources and short execution timelines. Using a human-centered approach may be important to creating two-way communication that informs the value proposition from stakeholders' perspectives. Personas, initially developed by ACDEB, can be enhanced by the NSDS as a tool to inform product development across federal agencies. The federal government must be diligent not to let the SAP become a barrier to entry but rather to build the tools to help agencies identify the relevant personas and data efficiently.

Currently, federal agencies are focused on traditional users and current researchers they serve who are already comfortable coming to a virtual or physical enclave and doing research with microdata. Guidance is written to facilitate tiered access that helps serve nontraditional users. The guidance provides different levels of security with the idea that the SAP can point people quickly to something they can access today. This tool would be particularly useful if it were integrated with metadata so users could start the process by searching for data that would meet their needs. It would then send users to the SAP, which would steer them to channels where they can access the data needed for their purpose or inform them if they cannot access the data or the data do not exist.

Engaging with the ICSP and seeing the SAP as a "one-stop shop" could be an avenue from which to grow stakeholder engagement. To help facilitate this for diverse stakeholders, the goal is that metadata systems will get to a place where they have a similar enough look and feel across agencies.

### How can the federal government engage state and local governments to continue building robust partnerships?

The current perception of state and local governments seems to provide a less favorable characterization of the relationship between themselves and federal agencies than what the agencies believe. Particularly, state and local governments feel that communication primarily goes in a one-way direction where they are told what to do, or, when they are being consulted, it is on policies that have already been decided. State and local representatives are looking for a seat at the table when decisions about policies and actions are being made so that they have a more prominent voice in deciding what data are collected, how they are collected, how they are disseminated, and the standards and technologies used for data collection.

More can be done with state and local governments to cement two-way communication—this is a relationship-building opportunity. There is a high degree of heterogeneity across state and local governments concerning money, talent, and support from governor's offices, etc. Federal agencies need to build more support and grow local capabilities where they can.

For example, BLS has begun a demonstration project to engage 15 Labor Market Indicator (LMI) shops around the country (they selected 10 with good capabilities and 5 with limited capabilities) to see if they can create more robust data sets that answer questions at local and federal levels about the demographics of labor markets. An aspiration is to develop a broad definition, beyond the scope of just CIPSEA, of what data could go into an NSDS and be validated and usable for creating tools like dashboards. In addition, a goal of the project is to address clear shortcomings that have developed over time in working with local governments. BLS will report back on the demonstration project in the near future.

### What role does capacity building play in supporting state and local governments?

State and local government capacity is exceptionally varied; building an NSDS that is truly useful, usable, and used by all will require substantial investments in human capital. The NSDS could, in a constructive, polite, and discreet partnership, identify states and local entities that would benefit from assistance and scope out the best means of support. While good models exist, some places have little capacity to engage, so the federal government must be thoughtful and measured with expectations of human capital and capabilities available at the state and local levels. Resources may not be as deep as federal agencies think, as tight budgets have led to tough sacrifices from state and local governments. A stark example of weakness in the data system was demonstrated in 2020 and 2021 unemployment insurance deployment across governmental entities.

### Are there use cases out there of effective data collaboration across state governments?

The Midwest Collaborative is an effort among 8 or 9 states that came together based on the realization that there are research questions and opportunities that would be better addressed and studied through a collaborative effort. The governance structure allows states to maintain a level of autonomy over how data are utilized while also coming together to find value proposition in collaboration.

Through collaboration with the Coleridge Initiative and the National Association of State Workforce Agencies (NASWA), the states have stood up a framework for governance that allows them to set priorities as well as to use a platform where data sets can be shared, allowing research across state borders. To get started, a few states worked through NASWA's LMI Committee and state administrators who sit on multiple committees in order to identify overlaps in issues and opportunities for joint research. There was an understanding that the workforce has changed quite a bit (gig work, movement of workers, etc.), and the states wanted to better understand movement across geographic borders (going to school in one state and working in another or working in multiple states). The Coleridge Initiative's Administrative Data Research Facility provides a platform that helps collaborators obtain a more holistic and regional perspective.

There is a real effort to broaden the data sets available for engagement in a collaborative setting. The Midwest Collaborative started with labor market information, but topics including equity; education to employment; and children, family, and social services offer areas for overlap and collaboration with agency stakeholders. States' levels of involvement have depended on what the states are interested in and what data sets they're willing to put into the collaborative.

# Standard Application Process: Governance

## Primary ACDEB Subcommittee: Governance, Transparency, and Accountability

### *Key Takeaways*

- Developing the Standard Application Process (SAP) is an important and necessary effort.
- The initial scope should be narrow and focused on access for researchers to statistical data produced by statistical agencies.
- At this stage, it is unclear how applicable the SAP will be for access to data from federal programmatic agencies and from state and local entities.

### *Summary*

#### What are the key attributes of governance for the Standard Application Process?

CIPSEA 2018 requires an SAP that all CIPSEA agencies are statutorily required to use. The mandate is much broader than just the technical application, covering review, timelines, public reporting, and an appeals process. The key principle is that all aspects of the process must be standardized; standardization is the goal, and the ICSP is working toward a common framework to achieve it.

The SAP policy workstream will set the governance framework, working toward a single online portal that will serve as the "front door" for users to get access to any confidential data from a recognized statistical agency or unit. This includes both statistical and administrative data in covered agencies and units.

To ensure that users understand what is available to them and how it can be used, it is important for the SAP to begin with data discovery. There will be a focus on "discovery" metadata with links to full metadata from agencies.

The SAP framework focuses on standardizing elements of the application while allowing agencies to include customized requests for additional information where required by laws governing specific data sets. The SAP will also include a common set of review criteria, including confirmation of statistical use, specified need for confidential data, feasibility for access to data, and statistical disclosure limitation techniques. There are different risks for different data sets and thus different levels of screening needed for applicants.

Once a user applies through the portal, the SAP will track communications and facilitate a dialog with the applicant. Agencies can ask for more information or corrections from users along the way through the SAP, and, at the end of the process, the application will be approved or rejected. The process will require standardized timeframes to track how long an action is with an agency before it moves to the next step.

### What are some potential concerns as the SAP is developed and on which ACDEB can provide input?

The Committee can help the ICSP think about what the SAP should and can be today, what the vision is for the future, and how to connect the dots between the current and target states. ACDEB can provide scenarios that will highlight requirements for the process, particularly from the perspective of state and local governments.

Additionally, the ICSP would like input on how the regulation should drive the process from the "front door" portal to the larger SAP sitting behind it. For example, ACDEB can provide ideas for outlining access rules, data sensitivity guidelines, and a tiered access framework. It will be important to take the right approach to tiered access and to try to mitigate unintended consequences, especially when looking at the limitations of state and local data compared to CIPSEA agencies' data sets.

The SAP should not be a hindrance; instead, the goal is to reduce barriers and streamline the process without increasing disclosure risks. Section 3582 of Title III of the Evidence Act includes requirements for tiered access, and the SAP will work with the envisioned tiered access framework. This could allow for a less involved process depending on the access requested.

The ICSP is also looking for ways to provide data concierge-type services that connect users to data. Currently, the process does not include this type of service; as a start, the SAP will clearly outline agency points of contact so users can connect with them to ask about data.

The ICSP recognizes that a single portal will not solve all problems, but the goal is that it will be an improvement over the current state, where users must rely on distinct processes for each data set when requesting data from multiple agencies. The aim is not to produce a policy that specifies every detail to which the federal government is bound. Instead, the SAP will be a significant step forward that meets statutory requirements while leaving room to grow based on use cases connected to the NSDS and influenced by the recommendations of ACDEB.

### What are the main components of the SAP governance structure?

There are two main pieces to the SAP governance approach—a governance body and an advisory committee. For each of these pieces, the ICSP recognizes that there could be a phased approach, starting with a structure that meets the statutory requirements and expanding over time as the SAP grows.

***Governance body.*** This group would operate on behalf of the ICSP by representing statistical agencies and units and assisting OMB with SAP oversight. Dual functions recognize the interdependence between the SAP and other statistical policies and the role of providing advice to OMB on these connections.

ICSP members have drafted a charter that outlines aspects of the governance board, including addressing who will chair the board (Chief Statistician or representative), who will be on the board, and how to nominate members. It is envisioned that there would be three standing subcommittees—on policy, communications, and technology, in line with existing SAP workstreams.

The majority (or super majority) of members on the governance board will be from the ICSP, and the remaining members will come from other stakeholder groups (like federal Chief Data Officers). The SAP is not voluntary for statistical agencies, but it would be good to have members from groups who would use the SAP voluntarily, as it seems logical that other agencies would move into this process over time.

The goal is for the SAP to grow and for the governance body to be transparent and open to stakeholder input while respecting the legal risk and responsibility of the statistical agencies. While academics, researchers, and state and local governments could also be seen as key stakeholders to the SAP, this is a required entry point for CIPSEA data, so there cannot be a board that would take the process in a direction out of line with the requirements around those data.

***Advisory committee.*** This body would advise on policy, procedures, and technology for the SAP and meet regularly with the governance body.

The ICSP has provided initial feedback on the advisory committee. Comments ranged from indicating that the SAP does not need such a body (for example, the governance body could solicit input directly from stakeholders) to suggesting that the advisory committee start with federal employees or be a full-fledged group under the Federal Advisory Committee Act (FACA) with a diverse set of stakeholders.

Initial thoughts from the subcommittee on a potential advisory committee include the following:

- The advisory committee is critical if the SAP grows beyond what is required by law. If the SAP is to be the central intake point for the NSDS, it will be important to have an advisory committee to provide input from the broader community. This could happen in phases, perhaps starting with a body composed of federal employees and adding members from other key stakeholder groups as the SAP expands.
- It is important to think about the incentives for expanding the use and usability of the process. The governance structure should be set up to reward the entity responsible for administering the SAP for bringing in more users—for example, creating a key performance indicator that measures the use of diverse groups like states or programmatic agencies as a signal for usefulness.
- Often, federal advisory committees do not provide easily actionable recommendations. To extract the most value from the assembled expertise on the advisory committee, it is helpful to provide materials to which the group can react. So, the governance board could ask the advisory committee to address specific topics and provide direction, identifying what works well, what does not work well, and lessons learned to move forward. This may be more effective than having members think about the entire process and how it may evolve.
- The advisory committee could provide input into key performance indicators to incentivize expansion and transparency.

### What advice would be most helpful for ACDEB to provide?

It would be helpful for the subcommittee to provide input on the advisory committee, addressing questions like:

- How formal should the structure of the advisory committee be? Could this change over time? For example, is it better start with an advisory group composed of federal employees and expand later to include other key stakeholders? Or to start right away with a FACA committee?
- Where could the SAP go in the future, and what are the implications for the "right" composition of members on the advisory committee?

# Standard Application Process: Technical Assistance

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- The SAP illustrates that the 13 statistical agencies can come together on a common template—a good foundational step. This is an area that warrants more subcommittee discussion.

### *Summary*

### What is the status of the Standard Application Process (SAP)?

The early version of a portal for accessing the SAP is the ResearchDataGov portal that the Inter-university Consortium for Political and Social Research (ICPSR) created for the Census Bureau. Currently, six statistical agencies make limited data available through the portal.

Ultimately, the SAP builds on this foundation. ICPSR, the current contractor, has been working on the portal for the last year and a half. While limited in its initial rollout, the number of data sets will grow over time. The metadata portion of the portal is now available for researchers to explore. This portion helps users discover what data exist and how they can be used. The portal links to the agencies who own data rather than replicating the information in two places.

### What is the SAP design approach?

The design of the portal has been advanced by three working groups of the ICSP, led by an executive steering committee. The SAP policy was issued for public comment. Currently, the ICSP is reviewing the public comments and updating the policy to reflect this feedback.

The ICSP stakeholder engagement working group is developing a communications plan for publicizing the updated portal. The group will also create interactive engagement opportunities to begin once this version rolls out. The ICSP is open to advice on how to contact and interact with a broader user community to inform future iterations.

The current version of the portal is designed to help statistical agencies meet statutory requirements of Title III of the Evidence Act. As such, this version will be used by statistical agencies per the legal requirements, and other participants are being invited to join. Moving forward, the SAP may include administrative data providers.

### How will the SAP work with applicants and agencies?

The portal is a front-end application tool for a process that will be managed at the individual agency level. The Evidence Act did not unify data access rules and requirements across the statistical system, so each agency is still responsible for interpreting its own statutes. However, the application policy establishes a framework for determining who gets to use data and timeframes for how quickly applications will be evaluated.

The portal is designed to track the process and provide a one-stop shop for applicants as their project moves from application to consideration to decision to appeal, etc. The process includes the ability to pause the application to iterate with an applicant, as needed, to clarify the proposal or to address questions. Through the portal, agencies can reach back out to researchers and work with them to try to ensure proposals meet the statutory requirements.

The application portal collects data both on the researcher and on the project:

- The process is not designed to capture detailed personally identifiable information but only to obtain enough information about the researcher to streamline the accreditation process and provide transparency. The ICSP recognizes that there is still room for improvement. For example, security clearance processes vary by agency. Ultimately, the goal is that once an agency accredits a researcher, that person will not have to go through the full process to apply to use another data set in another agency.
- For project information, the goal is to collect enough details so that an agency can determine whether data can be used for the proposed purpose with minimal back and forth with the researcher.

The whole process is overseen by the National Center for Science and Engineering Statistics (NCSES) program management office. NCSES requested and received funding to cover a small staff responsible for holding agencies accountable for meeting timeline markers written into the policy. Each agency has been asked to provide a test group of users to provide feedback on the portal.

Developing the SAP policies and portal has helped the statistical agencies to work as more of a seamless system. This exercise has shown that 13 agencies can come together to agree on a common template and process. The process will need to evolve further, but there has been a great deal of agreement, compromise, and demonstration that the federal statistical system is evolving toward a more cohesive future.

# Other Items: Responsibilities of Statistical Agencies (Trust Regulation)

## Primary ACDEB Subcommittee: Legislation and Regulations

### *Key Takeaways*

- Codifying Statistical Policy Directive #1 into law was one of the Evidence Commission's key recommendations. This regulation is not just about the statistical system; there is the potential to have a bigger conversation about trust in data and what that means for OMB, agency heads, etc.
- ACDEB could suggest actions in the context of the Trust Regulation to bring pieces of the Evidence Act together. For example, if CDOs and statistical agency heads are creating separate data inventories, does that demonstrate transparency that leads to public trust? How could the Trust Regulation encourage better collaboration that would build trust in data?
- For the broader public, it would be helpful to articulate this work in an efficient and accessible form like a news story—for example, clearly describing what political people can and cannot touch and how this supports reliability and trust in data.
- The Committee does not need to wait until the Year 2 report to weigh in on the Trust Regulation; the Legislation and Regulations subcommittee can provide preliminary input to OMB in the shorter term, following ACDEB's standard processes and guidelines.

### *Summary*

#### What is the "Trust Reg" and how does it fit in with the Evidence Act?

The Evidence Act provides for common frameworks around data acquisition, application process, sharing, protections, and access, built on the philosophy that autonomy depends on the ability to produce trustworthy statistics—this is at the core of what the trust regulation will address.

For federal agencies, the statute is clear; for instance, parent agencies need to enable, support, and facilitate the statistical agencies doing their jobs. The goal is to put policies in place to encourage parent agencies and statistical agencies to do what is outlined in the Evidence Act. This is a larger endeavor to translate existing policy directives and expand them into regulatory text.

#### What is the process for developing the regulatory language? And what are the next steps?

Through the ICSP, OMB has sought interagency input into the development of the regulation. Once the regulation goes into formal Executive Order 12866 regulatory review (including interagency review), there will be another opportunity for both parent and statistical agencies to provide more formal input. Additionally, there will be a public comment process after the formal review through a Federal Register Notice.

The subcommittee can provide initial input to OMB that would be influential even before the public comment period. This could include providing answers to targeted questions, preliminary findings framed as written comments, or another informal mechanism.

### What advice can ACDEB provide now?

An overarching question will be: does this regulation advance the system and do what it is intended to do, which is to enhance reliability and trust in the statistical system? Key questions of interest include the following:

- What should the regulation require of existing statistical agencies? What are the optimal ways in which statistical agencies could ensure requirements are met?
- Thinking through the list of responsibilities for statistical agencies, what are some helpful tools (e.g., transparency, reporting, peer review, policies, or standard operating procedures) for addressing each responsibility?
- What about new statistical agencies or units? OMB is the entity that recognizes an agency as being under CIPSEA, so what does an agency need to do to be recognized? What responsibilities does an agency need to fulfill?
- What is the best way to translate the general facilitation required of parent agencies into specific actions?
- This regulation will apply throughout government, so does it need to include information specifically about OMB's role? What about other parts of the government—what can they do to facilitate trust?
- Is this regulation enough? Are there unintended areas where the public may feel that the regulation does not do enough to enhance reliability and trust in the statistical system?
- On transparency, what is the best way to communicate with the public what the federal statistical system, OMB, and other government actors are doing to support trust in government?
- On relevance, what are demonstrations of relevance by a statistical agency? Or are there certain things entities could do that impinge on an agency's ability to be relevant? Who determines relevance? How do other stakeholders (like state and local governments) get involved in deciding what's relevant?

### How does the guidance deal not only with reassuring citizens there is no political interference but also improving the larger public perception of trust?

When thinking about transparency, there is no difference between the appearance of interference and actual interference. The existing policy directives are about predictability and transparency. For example, there is already a pre-approved calendar for data releases, and deviations from that must be approved. In addition, statistical agency heads are responsible for making sure data remain confidential; they make the decisions about who has access, who is a sworn agent, etc. so that decisions are tied back to the people with legal responsibility for the data. OMB is translating those directives into regulatory text to enhance reliability and trust in the system.

The policy directives outline four responsibilities of statistical agencies that stretch well beyond freedom from political interference. The first responsibility, for instance, is around relevance and timeliness—one significant reason statistical agencies get "permission" to have independence from political oversight is because they've demonstrated their relevance. It is critical to consider these responsibilities as a set.

**The Evidence Act is about making more information available to the public for decisionmaking, and data sets are being inventories and searched that were not available before. Will the impact of OMB's rules extend trust obligations in terms of data reliability to data sets that weren't covered before?**

To the extent that administrative data are brought into the system, the answer must be "yes." This isn't easy because, for example, statistical agencies link CIPSEA data and non-CIPSEA data. Looking across CIPSEA 2018 (that is, Title III of the Evidence Act), some provisions focus on data and some focus on the agency. This provision focuses on the agency as a whole—it's about the agencies' policies and actions and it covers whatever the agency is doing (whether with CIPSEA data or non-CIPSEA data).

**If the reliability requirements extend to more data sets, will that impact staff requirements? Are new protocols needed to address these requirements?**

This could certainly create resource issues and is an area where the Committee should review the President's FY 2023 Budget request and weigh in with preliminary input and recommendations.

**From the security and privacy perspective, what are the benefits and drawbacks of responsibility ultimately lying with the statistical agency heads? On one hand, there are benefits of the federalism approach to handling confidentiality and privacy issues within each agency; on the other hand, it would be helpful to have some overarching guidance. To what extent should there be such guidance laid out in the regulations, and what advice could ACDEB offer in this area?**

It makes sense to assign responsibility to the head of the agency. However, OMB will provide regulations to create consistency around the consideration of issues with the acknowledgment that decisions may be different across agencies. Tiered access is the obvious approach for this, but translating that concept into the modern world, especially in terms of bringing it to the National Secure Data Service, is at the crux of the issue—ultimately, looking to promote autonomy within a framework that is subject to public scrutiny.

There will be further examination of a different regulation on expanding restricted access to data, which will require OMB to regulate on tiers of data sensitivity, corresponding tiers of access, and types of re-identification risk mitigation. This question is relevant to the Trust Regulation but also to forthcoming conversations around the other regulation. It would be helpful for ACDEB to consider how these regulations fit together. In addition, since regulations are at a high level, OMB will issue guidance beneath them and provide a methods toolkit.

# Other Items: FY 2023 Budget

## Primary ACDEB Subcommittee: Government Data for Evidence Building

### *Key Takeaways*

- There is an emphasis in the budget narrative and throughout the budget on the importance of promoting evidence-based policymaking. Most of the budget discussion related to ACDEB's work is in the "Leveraging Federal Statistics" Analytical Perspectives chapter that emphasizes trust in the federal statistical system.
- Significant budgetary impacts include the following:
  - Investment in the National Center for Health Statistics for developing a virtual data enclave.
  - Additional money for the Statistics of Income Division at the Internal Revenue Service could allow the agency to support more research by outside researchers.
  - Increased funding for the National Center for Science and Engineering Statistics to support the building out of the America's DataHub Consortium (ADC), which may serve as a foundational component for the NSDS.

### *Summary*

#### How does the FY23 Presidential Budget reflect ACDEB priorities?

A consistent priority expressed throughout the budget and supporting documents is the importance of evidence-based policymaking. The following two chapters focus on evidence building and evidence use:

1. Building and Using Evidence to Improve Government Effectiveness
2. Leveraging Federal Statistics to Strengthen Evidence-Based Decision-Making

The most relevant budget initiatives and the reflected vision for ACDEB are covered in the Analytical Perspectives chapter on "Leveraging Federal Statistics to Strengthen Evidence-Based Decision-Making." The chapter lays out a vision that OMB and leaders of the statistical system are developing, which is particularly relevant to Title III of the Evidence Act. This vision reflects OMB's and ICSP's discussions with ACDEB. It covers the core functions envisioned by the Evidence Commission for an NSDS, which were included in CIPSEA 2018 and that have been affirmed by the Committee.

### What does the FY23 Presidential Budget suggest for the future NSDS?

Many of the envisioned capabilities for the NSDS are being developed even before a new entity is brought into existence, as many of these functions are part of the statistical system's mission. OMB and federal statistical agencies are considering the right role for an NSDS inside this evolving ecosystem, as they build a better understanding of these capabilities. A key idea is that the NSDS will live inside the CIPSEA ecosystem, as was suggested in the ACDEB Year 1 report, and that any new entity will complement the expanded missions of the statistical agencies under CIPSEA 2018.

Statistical agencies are trusted intermediaries between data providers and evidence builders, and this vision retains their role as trusted stewards of the nation's most sensitive data. As such, the vision starts with a foundation of trust and recognizes the close alignment between existing statistical missions and the goals of expanding evidence building and regular engagement with stakeholders to identify and develop relevant data sets and products.

### What investments in the FY 2023 Presidential Budget support the emerging vision?

There is alignment between the stated vision for and investments in the work of the Evidence Act. Some investments that reflect these priorities include:

- ***National Center for Health Statistics:*** Investments to build on existing infrastructure around administrative data standardization, acquisition and linkage, and access.
- ***Statistics of Income, IRS:*** Funding to expand staff support for a small researcher access program, which has yielded groundbreaking studies, as a partial solution toward expanding evidence-building capacity.
- ***National Center for Science and Engineering Statistics:*** Investment to expand the components of the Standard Application Process, to conduct early work on the NSDS, and to leverage America's DataHub.

# Policy Implications for the Standard Application Process, the Access and Confidentiality Regulation, and the Presumption of Accessibility

## Primary ACDEB Subcommittee: Legislation and Regulations

### *Summary*

### How do the SAP and access regulations support the realization of the Evidence Act's goal to expand secure access to CIPSEA data assets?

A key theme of the Evidence Act is to ***safely expand access to protected data for evidence building.*** The ICSP has been thinking deeply about what is necessary to realize that big idea and recognizes that effective federal data use requires high-quality governance. The ICSP is working on elements of that governance in constructing the SAP, such as what a governance board should look like and who will be responsible for it.

Ultimately, the SAP is envisioned as a separate but complementary construct to the NSDS, and it is part of the statistical system along with the NSDS. The SAP is looking to be broader than only covering recognized statistical agencies. One question that ACDEB can help address is how to make it attractive to other entities that are not required to participate. While the ICSP does not have the authority to compel participation, the ideal state is one where any federal agency that holds protected data is participating in the system. This is a growing conversation between the SAP and CDOs as they build toward a seamless, discoverable set of inventories and catalogs that include restricted and unrestricted data.

The ICSP is working on putting the mechanisms in place to move forward toward realizing this vision. Those mechanisms include the following:

- Having a transparent process for recognizing statistical agencies and units in terms of identifying trusted intermediaries;
- Defining responsibilities in the trust regulation;
- Empowering trusted intermediaries to access data, actualizing the Presumption of Accessibility;
- Ensuring effective provisions of data to evidence builders in terms of expanding secure access to confidential data assets; and
- Building tools and infrastructure like the SAP for access.

### What do the SAP and Access Regulations look like through the perspective of the Five Safes framework?

ICSP members have found it helpful to think about the different regulations, policies, and tools and how they all fit together as if pieces in a jigsaw puzzle. The Five Safes provide a framework for thinking about that puzzle. Each of the "safes" serve as a different dimension of a comprehensive process for access and confidentiality.

The SAP's project approval process will address safe projects and safe people, while the Access and Confidentiality Regulation will address safe data, safe settings, and safe outputs in a way that encourages tiered access in alignment with data security needs.

The SAP uses the CIPSEA concept of agent status to ensure safe people, and the ICSP welcomes suggestions for how the statistical system can use the SAP policy and legal provisions to manage access. Within the SAP process, project approval is the mechanism for limiting access to safe projects, and this is kept distinct from approving the people doing the project. This distinction could allow for retaining safe status as an individual when applying for subsequent projects. Likewise, a person could have fast-tracked or tiered reviews for evaluating safe projects based on meeting a set of conditions for agency evaluation.

The other safes are going to be addressed by the access regulation. The access regulation is about common frameworks for data sensitivity and aligning sensitivity level to access level. For the moment, the SAP has focused on restricted tiers, but the access regulation will formalize a tiered access structure (as required by the law). For example, microdata that are fully identifiable could be the most sensitive tier, a level up from that could still be restricted but with fewer requirements for trusted access, and a level up from that could be synthetic data with validation or a query tool. At some level up, there might not be a need for a background check or project proposal.

The SAP acts as the "front door," at least for a certain subset of data, and the access regulation creates the structure for tiers of access. The NSDS should fit into this system, broadening data discovery beyond the SAP and in a way that is transparent to users, while operating in the same regulatory structure for a tiered access framework. The whole system living under the same framework will formalize the concept of access existing on a continuum rather than as a binary state.

### What will the regulations do to further the Presumption of Accessibility?

The Presumption of Accessibility for statistical agencies and units provides a powerful piece of law that the regulation is meant to realize. The law says that any part of the federal government must give data to agencies upon request with exceptions. It is designed to support a very broad concept encapsulating any purpose of developing evidence. In the future, any statistical agency can acquire data from anywhere else in the government with limited exceptions to support evidence-building activities. The NSDS can play a critical part in this ecosystem, allowing evaluation beyond existing agency data sets by acting as a matchmaker and a facilitator to link those data sets.

One challenge will be ensuring equitable access. For example, there should a be a way to make it such that young researchers are not at a disadvantage to access data, or that institutions without long track records or deep infrastructure, such as small schools and historically Black colleges, are able to access the system on equal footing as better funded and established institutions.

Right now, ICSP members are working on the preamble and framing out the rest of the regulation. They are almost to the point of drafting the regulation and looking at issues related to identifying risks and defining harm. They see part of this as addressing a risk management challenge and are tackling how to start operationalizing it. The ICSP recognizes that risks can evolve over time and best management practices change, so they are thinking how to include periodic reviews of disclosure methods, whether that is an essential feature, and how to conceptualize the word "harm" in terms of drafting the regulation. The ICSP is open to suggestions and advice from ACDEB.

# 4. Subcommittee Guest Speaker Summaries

As part of the information-gathering process during Year 2, ACDEB subcommittees hosted targeted discussions with outside experts. To cross-pollinate ideas from the focus areas, other Committee members were invited to attend these sessions and ask questions from the perspectives of their subcommittees and areas of expertise.

Table S4 provides an overview of each outside expert meeting. The Committee would like to thank all speakers and supporting staff who made these sessions a reality. This section includes summaries of these meetings. The information shared during these sessions does not reflect the views of the full Committee. In addition, the summaries do not reflect changes that may have occurred after the meeting dates.

### Table S4. Outside Expert Meetings Overview

| Organization(s), Topic, and Date | Host Subcommittee | Speaker(s), Planning, and Support |
|---|---|---|
| **Data Quality Campaign: Communication strategies**<br>February 24, 2022 | Other Services and Capacity-Building Opportunities | Rachel Anderson, Jenn Bell-Ellwanger |
| **Midwest Collaborative (MWC), National Association of State Workforce Agencies (NASWA), and Workforce Information Advisory Council (WIAC): Governance insights**<br>March 4, 2022 | Governance, Transparency, and Accountability | George Putnam (MWC and Illinois Department of Employment Security), Yvette Chocolaad (NASWA), Lesley Hirsch (WIAC and New Jersey Department of Labor and Workforce Development) |
| **Results for America: Communication strategies**<br>March 10, 2022 | Other Services and Capacity-Building Opportunities | Cheryl Burnett, Zachary Coile, Nichole Dunn |
| **Urban Institute: Synthetic Data and Validation Servers**<br>April 8, 2022 | Technical Infrastructure | Leonard Burman (ACDEB and Urban Institute), Graham MacDonald |
| **Datavant: COVID-19 Research Database**<br>April 8, 2022 | Technical Infrastructure | Claire Cravero, Jake Plummer |
| **Jobs and Employment Data Exchange: Data dictionary**<br>April 20, 2022 | Government Data for Evidence Building | Kenneth Poole (Center for Regional Economic Competitiveness) |
| **Privacy threats and re-identification risks**<br>April 21, 2022 | Technical Infrastructure | Claire Bowen (Urban Institute) |
| **Federal Statistical Research Data Centers Panel: Decisionmaking, infrastructure, and technical assistance**<br>April 21, 2022 | Other Services and Capacity-Building Opportunities | Mary Campbell (Texas A&M), Barbara Downs (Census Bureau), Cathy Fitch (University of Minnesota), Maggie Levenstein (University of Michigan), Amy O'Hara (ACDEB and Georgetown University) |
| **State Wage Interchange System**<br>May 4, 2022 | Government Data for Evidence Building | Greg Wilson (Department of Labor), John (Jay) LeMaster (Department of Education) |
| **Inter-university Consortium for Political and Social Research: Technical assistance**<br>May 5, 2022 | Other Services and Capacity-Building Opportunities | Maggie Levenstein |
| **Privacy-preserving solutions for the future and risk evaluations**<br>May 6, 2022 | Technical Infrastructure | Wade Shen (Actuate Innovation) |
| **Opportunity Insights**<br>May 6, 2022 | Technical Infrastructure | John Friedman (Brown University) |
| **National Institutes of Health (NIH) Library Data Services: Technical assistance**<br>May 19, 2022 | Other Services and Capacity-Building Opportunities | John Doyle |
| **NIH National Center for Advancing Translational Sciences National COVID Cohort Collaborative**<br>June 3, 2022 | Technical Infrastructure | Kenneth Gersing, Sam Michael, Leonie Misquitta |

# Data Quality Campaign

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- The Data Quality Campaign provided examples of specific artifacts that they used in the education sector to conduct advocacy efforts at the federal, state, and local levels.
- The work of the Data Quality Campaign has illustrated that it is crucial to be transparent, earn trust, and protect privacy. These qualities are interlinked and foundational to working with student educational data, of course, and must be kept at the forefront to inform ACDEB's work around structuring the NSDS.
- Strong infographics and other communications assets go a long way for public engagement and transparency.
- Measuring what matters (i.e., clearly articulating the questions to answer before collecting data) and building state and local capacity could improve data quality.
- Remember that data are about people.

### *Summary*

#### What is the Data Quality Campaign (DQC) and what does it do?

The DQC is a national, nonpartisan nonprofit focused exclusively on changing the role of data in education and in the full spectrum of sectors and services that serve young people and students. The DQC works primarily at the state level but also has experience with federal and local sectors and is focused on the role that information, data, and connected systems can play in serving young people—in service of students' needs rather than in service of the system itself. A key question is: "What does it look like for individuals to have data work for them?"

#### What are the DQC's policy priorities, and how do they help build strong data systems and a strong culture of data use?

- ***Policy Priority 1: Measure what matters.*** Be clear about the questions a data system seeks to answer, starting with the big policy questions and working backward to the data and data linkages that can answer those questions.
- ***Policy Priority 2: Make data use possible.*** Data don't just live in a system but become an actionable tool, often using data at the federal level to inform decisionmaking and improve policy alignment across federal, state, and local levels. This also involves building capacity within agencies focused on the human aspect of data capacity (ensuring people working with young people have the skills to use data effectively and ethically).

- ***Policy Priority 3: Be transparent and earn trust.*** People won't use data they don't trust; it's hard to get people on board on the backend, so it is important to get buy-in from the start. This could be done by providing useful information to the public about how young people's needs are being met, which could be a "quick win" (e.g., meet a need or answer a question) to garner public support. It is important to communicate clearly how federal data are safeguarded and used to help students, rather than waiting for someone to bring up privacy concerns that then control the narrative of what it means to use data.
- ***Policy Priority 4: Guarantee access to data while also protecting privacy (the double-sided coin).*** Ensure authorized people have role-based access to the types of data they need to do their jobs. It is important to see access and privacy as interconnected rather than opposing.

### How are trust and effective communication related? What are some communication mechanisms the DQC uses to bolster trust?

The discussion around data isn't objective. Public conversation recognizes that data are about people, and the very existence of data reflects decisions made by people—what data are collected, how data are collected, who is represented in data, and who made these decisions.

The DQC created a consumer's guide to data that features principles to guide building trust in data, both for data users and producers. The three main areas of focus are the following:

- ***Context.*** Help consumers make sense of the numbers. Data are a product of decisions, rather than being inherently "right" or "wrong." It is important to communicate measurements, definitions, and what data can and cannot explain.
- ***Proximity.*** Consider the source and voice of the person sharing the information. The messenger matters, so it is valuable to share data tools through trusted and "local" voices, and to think about the intended user and where intermediaries can be helpful for sharing information.
- ***Framing.*** Think about ways that words and asset framing build trust. Use asset framing (or defining people by their aspirations before their challenges) instead of deficit framing (or describing people by their problems). Asset framing uses language about both equity and accuracy.

### Are there infographics or other resources that have resonated with people who aren't already on board with linking and using data?

Even great data efforts sometimes fall apart because they aren't grounded in the immediate needs that people on the ground are having. The DQC finds it helpful to start with identifying the burden that educators are facing, then provide them with a data system or data tool as a solution.

### Does the DQC have any materials targeting privacy skeptics? Or privacy-first infographics?

The DQC materials focus on privacy by design, as it is often not helpful to take a mythbusting approach. Even repeating a privacy concern to bust it is still repeating that concern. The DQC suggests reviewing Actionable Intelligence for Social Policy's framework on privacy by design as a way to build in protections and resources from the start.

### Does the DQC have suggestions for communicating to real people what the privacy budget is?

One message that resonates is taking privacy out of a vacuum. There are enormous risks associated with not acting and not using every resource to support young people. Risk exists no matter what; the focus is on mitigating privacy risks because the alternative is not helping young people.

### Is the DQC specifically contracted by organizations for engagements? What are typical deliverables? How has that been successful?

The DQC is foundation-founded and philanthropically funded, with a focus on national best-practice work as well as working with states and districts. The DQC serves as a thought partner to work through challenges and provide resources. The DQC can also provide help in thinking through legislative responses and about data culture in a state.

### How often does the DQC bring in data from other sectors outside of education (e.g., education and workforce or education and justice)? Could the DQC provide materials or resources on interagency collaborations within a state?

This is primarily an approach that works through specific sectors, but lessons are generally applicable (e.g., work with foster care system and education provides a decent model for looking at other sectors). The DQC has a roadmap for bringing together different systems.

### Are there educational resources that are designed to help policymakers and other decisionmakers understand how to use data. This is not about communicating with the public but answering the question: "How do I, as a state agency or federal agency decisionmaker, start to use data more effectively in my policy role?"

The DQC has resources that help answer this question and recommends talking to Results for America. (For information on the ACDEB session with Results for America, see below).

# Midwest Collaborative, NASWA, and WIAC

## Primary ACDEB Subcommittee: Governance, Transparency, and Accountability

### *Key Takeaways*

- There is an urgent need for timely, locally relevant data and evidence that can be used to respond to the changes in the COVID-19 pandemic economy, particularly for low-income earners and workers, at-risk youth, marginalized populations, immigrants, and formerly incarcerated individuals. New information can inform policies about investments in education and training, student debt, as well as health, welfare, and corrections programs. Federal, philanthropic, and state partnerships have led to new projects, products, and practices for evidence building. The results inform decisionmaking by many state departments of labor, education, and human services.
- Regional state data collaboratives are creating networks with a national reach across agency and state lines. They are working in partnership with each other and regional universities to produce data products that policymakers, practitioners, and citizens can use to answer questions critical to society. They are using a FedRAMP-authorized cloud environment to store, access, and share data to produce value for state and local decisionmakers and to ensure that evidence is reproducible and robust, and that there is equitable analysis for diverse and disparate groups.
- There is a variety of robust governance structures, including federally funded research and development centers (FFRDCs) and public-private partnerships.

### *Summary*

George Putnam, Director of Labor Market Information, Illinois Department of Employment Security

### What is the Midwest Collaborative and what is its governance model?

The Midwest Collaborative (MWC) is a coalition of state workforce and education agencies working in partnership with the Coleridge Initiative and regional university partners to design a system that enables individual states to answer critical questions that are relevant to societal well-being. Key elements to MWC governance include principles, structure, project approval process, and the trigger for implementation of permanent governance.

- ***Principles.*** Shared commitments that guide collective decisionmaking include state autonomy; agency oversight; documented value; project rigor; transparency; adherence to all applicable federal and state legal and compliance requirements; continuous process improvement; minimized burden on states, agencies, and researchers; and ensuring ability to scale and innovate.

  Adherence to federal and state legal compliance is the bedrock all states share before data can be hosted on a common platform; in developing these principles, states were clear that products need to have documented value to the states.

- ***Structure.*** The MWC governance structure includes the following components:
  - The MWC Executive Committee determines final approval on all policy recommendations and project proposals.
  - The MWC Council has the role of the policymaking body for the collaborative.
  - The MWC Data Stewards Board provides technical advice for the collaborative.
  - The Coleridge Initiative is the platform organization, providing and supporting the Administrative Data Research Facility (the common platform used for data ingestion, data documentation, data analytic tools, and data stewardship).
  - The National Association of State Workforce Agencies (NASWA) is the administering organization, serving in an advisory and consulting role. States need support in day-to-day operations for a functional governance structure, which is the larger role NASWA plays.
- ***Process.*** The project approval process includes the following two tiers:
  - ***Tier 1:*** A streamlined review for state-led projects that is only necessary where collaborative resources are needed or where the project is associated with the collaborative by name; fast-tracking collaboration among states with a low bar to make work across collaboratives easier.
  - ***Tier 2:*** A formal request for proposal process for external projects, which must address priority state topics. This tier ensures strong oversight, control, and direction of work that is not led by participating states.
- ***Transition to Permanent Governance.*** This involves formal MOUs with the states, one with the Coleridge Initiative and one with NASWA, on how states will work together on a project, establishing collaborative value added to the process. Once three states in a collaborative have executed both MOUs, the collaborative moves from an interim governance structure to a permanent one.

### What are the different levels of state participation?

State levels of participation form a continuum. For each level—exploratory members, participating member states, contributing member states, and full member states—there are different commitments and benefits.

### How do projects move to products and then into practice to support evidence-based decisionmaking?

Collaboratives have been defining the process. For example, the MWC started a project in response to COVID-19, then developed the unemployment to reemployment portal, and now member states are figuring out how to move from that product to a practice. The MWC is working with local workforce boards to implement this product as part of their practice of decisionmaking. The states had to demonstrate that the product provided actionable information: the value proposition ties closely to the level of information provided—the more granular the data, the more useful the product. The Coleridge Initiative's National Convening in March 2022 focused on the process of regional collaboratives moving from projects to products to practice.

### Yvette Chocolaad, Workforce Policy and Research Director, NASWA

#### How did NASWA become involved with the Coleridge Initiative and the Midwest Collaborative?

NASWA's interest in the Midwest Collaborative started within two groups—the Labor Market Information committee, focused on research and data, and the Employment and Training committee, on the program side. The committees expressed the following:

- A strong desire to leverage more state agency administrative data to create more timely data insights for policy, practice, and use; and
- Different frustrations with accessing and using administrative data, difficulties executing data-sharing agreements within and across states, lack of staff capacity and tools, and the need for training and talent.

States have a long history of acting as "laboratories of democracy," and while they may not have the resources to complete formal program evaluations, they are sitting on administrative data and can use that to generate insights for better decisionmaking. So, the question became how to help NASWA members move these goals forward.

In 2016 and 2017, NASWA received funding to document states' needs, developed a survey, collected member input, and published a report. This report looks at challenges, needs for data, and what questions governors, legislators, and agencies are asking, covering topics like understanding local labor markets, citizens served, impacts, programs, and services. The group also completed a related survey and report on the COVID-19 pandemic.

#### What role does NASWA play in the MWC governance structure?

NASWA's surveys and related reports laid the groundwork for the association to apply for and receive its first-ever grant to become the administrative entity for the MWC.

### Lesley Hirsch, Assistant Commissioner, Research and Information, New Jersey Department of Labor and Workforce Development

#### What is the Workforce Information and Advisory Council (WIAC)?

WIAC is a federal advisory council of workforce and labor market information experts representing national, state, and local data users and producers. The purpose of the Council is to evaluate and make recommendations to improve national workforce and labor market information systems, particularly focused on sharing information on innovative approaches, new technologies, and data to inform employment skills training and workforce decisionmaking and policy.

### What is WIAC's interest in ACDEB's work, and does WIAC have specific recommendations for the Committee's next steps?

Building on previous recommendations and reports around data sharing and the Evidence Act, WIAC's Data Sharing subcommittee reviewed ACDEB's Year 1 report and laid out several design considerations for the National Secure Data Service (NSDS) with slight variations on and different areas of emphasis than the priorities in the ACDEB report. In addition to the principles outlined by the Data Foundation and embraced in ACDEB's Year 1 report, WIAC also prioritized the following:

- The Federally Funded Research and Development Center (FFRDC) model, whose features include data security, legal authority, stable funding, and public-private partnership, as well as being interdisciplinary, intergovernmental, more independent, not bound by federal budget priorities, not bound by the federal production schedule, conducive to innovation and rapid turnaround, and able to attract and compensate statisticians, data scientists, and engineers, as needed.
- A linkage service, not a clearinghouse or data warehouse.
- The ability for owners to maintain control of their own data.
- Research capacity building.

Based on these priorities, the WIAC subcommittee outlined recommendations that support immediate action for an NSDS by assigning staff, articulating key principles (see above), and publicly committing to the Department of Labor's engagement with the newly established NSDS. As part of this engagement, the subcommittee recommends that the Employment and Trade Administration be engaged in planning and governing the NSDS as policy organizations and policymaking agencies (the "learning" agencies) need to be part of the governing body.

### How will it work to have states be responsible for their own data? How will a central collective help?

States retaining rights over the use of their data is essential because state mandate governs whether a stated purpose fits with requirements for data use. This is not a matter of states saying "no" but a matter of providing value for the states (e.g., states are less interested in supporting doctoral researchers than in creating actionable products).

So, the role for a centralized NSDS is to prevent redundancy and the need for expenditure. One part of this is sharing costs, but it is also about sharing expertise and technology.

### There are not many privacy experts out there, even if each state wanted to hire an expert and could afford to do so. How could the FFRDC model help with this?

FFRDCs attract people, and partnerships between federal and state levels create space to do work for the public good. For these reasons, FFRDCs are set up as public-private partnerships.

### There is interest in not having NSDS be a data warehouse, but would data from linkages be available to answer questions about reproducibility after the fact?

For reproducibility, it's also important for the NSDS to maintain the code used for analysis. Perhaps the states could maintain the analytic extracts for the specific analyses after the analysis is over.

# Results for America

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- Results for America (RFA) offered a series of concrete recommendations for developing an effective communications and education strategy for an NSDS.
- The RFA recommendations to develop key communication artifacts, such as a "History of the NSDS," a "what we do" document, and FAQs, are helpful to consider as part of the NSDS communications and education approach.
- It will be important to continue to identify key audiences for the Year 2 report (and beyond) and to take advantage of the networks of organizations like RFA to share key messages around the NSDS.
- RFA is focused on building capacity of state and local government leaders around data and evidence—a natural partner for the work of the Committee and (future data service). These connections were apparent in RFA's presentation.

### *Summary*

#### What is the mission of Results for America (RFA) and how does this connect to the priorities of the federal government?

RFA's mission is to make investing in what works the "new normal." RFA sees that the Biden-Harris Administration's commitment to evidence-based policymaking is helping make this "new normal" a reality.

To meet its mission, RFA helps governments across all levels implement evidence-based data practices by setting standards for data and evidence use at each level of government (local, state, and federal). RFA agrees with the Evidence Commission's recommendation that the United States needs a National Secure Data Service, and RFA has long been a champion of the NSDS effort.

#### What are RFA's recommendations for a comprehensive communication and education strategy for the NSDS as described in ACDEB's Year 1 report?

RFA has seven recommendations on the NSDS communications strategy. The theory behind these recommendations is that it's best for NSDS to tell its own story—any of the documents that get produced as part of this communications process will be what other people use when talking about the NSDS, so those documents should be the basis for communications. RFA's communications recommendations are outlined here:

- ***Recommendation 1: Define the NSDS.*** Make sure the public understands an NSDS at a very basic level, including a set of brief documents that build toward what could be a website and a set of materials. These basic materials could include a clear and brief mission statement, "History of NSDS, "What We Do" (and why), "Who We Are" (noting key agencies and other partners), and "FAQs" (addressing key questions and concerns).
- ***Recommendation 2: Identify key audiences.*** Start with the list in the Year 1 report and create a more robust, comprehensive list. For efficiency, focus on organizations and leaders who can be liaisons.
- ***Recommendation 3: Create a website.*** To build trust and transparency, launch and maintain an online and public presence; evaluation.gov is a good model for transparency. Also, consider resources and staffing to maintain the site.
- ***Recommendation 4: Build a communications plan.*** Brainstorm key elements needed for a comprehensive strategy; this will require a dedicated leader or coordinator. Think about a "big splash" versus a "soft launch." A soft launch may work if things are still in-process, as it allows for a release of select materials and provides a mechanism for feedback from stakeholders to refine efforts; the downside is a smaller initial impression and less attention.
- ***Recommendation 5: NSDS roadshow.*** Use a series of virtual and in-person events as an effort to get the word out and engage stakeholders on the value proposition. Events should include all levels of government, data providers, researchers, and the public.
- ***Recommendation 6: Case studies.*** Highlight successful strategies around data sharing, linkage, and analysis to help explain how the NSDS will create better policymaking, accelerate progress, and show impacts of efforts.
- ***Recommendation 7: Enlist partners.*** It is important to build networks of partners to share information with target groups so that story can be heard broadly. Many organizations and leaders would be interested in participating, including RFA sharing with its own network.

### How could an NSDS address people who are skeptical about privacy concerns?

An FAQ can be a good place to address concerns head-on; good faith critics can be valuable to meet with in order to listen to concerns. The Doar and Gibbs report addresses this—focus on how concern has been addressed satisfactorily, so it is clear how benefits outweigh risks.

### On building a list of potential users, could imagine list being extremely long and diverse—does RFA have thoughts on listservs?

A stakeholder list could be long, so the focus might be on an organization's membership (United States Conference of Mayors, National Association of Counties, National League of Cities, as examples); listservs could work. RFA tries to be proactive but via trusted relationships and umbrella ambassadors so that the communications coordinator doesn't have to do all the work.

### How does RFA approach passive versus active updates, posting information to a website regularly as opposed to blast emails, for example?

Blast emails and social media posts can be very brief and can focus on driving traffic back to a website that contains more detailed information.

### How to pick targets for the roadshow? Fear is limiting communications to select groups versus maximizing opportunities.

RFA provided the following suggestions:

- Pick organizations that are aligned most closely with the key target audience and have the biggest reach; start with those and then decide how to prioritize outreach to other audiences.
- Think about risk tolerance; some level of risk is good if the conversation isn't drowned out by critics. It is important to choose good faith partners and get their input.
- Transparency is important (again, FAQ could be helpful); be clear about what are the challenges, why are they worth overcoming, and why the whole effort is designed to do just that.
- Consider what decision needs to be made, who impacts that, how do target audiences align with that strategy, and who has the most reach with the target audience.
- Remember that one set of people doesn't have to do all the roadshows. If there are enough ambassadors, they can multiply the impact (e.g., different ambassadors speaking to different groups at different conferences).

### There seem to be two phases to communications—first is to Congress and how to address their potential objections and second is to broader audiences. How can use cases help characterize the benefits and describe how they outweigh the risks?

Can anticipate concern that this will be an enormous, new entity, so it's helpful that the NSDS is "a place, a service, and a philosophy"—this is an outgrowth of work happening in federal government. It is encouraging to note that RFA is a bipartisan organization, finding members of both parties to support these activities. In addition, it's easy to see the risks but much harder to understand benefits; this is a good focus to keep in mind as ACDEB moves forward (helping people visualize benefits as easily as they already visualize challenges). Focus on "making it simple enough that the person you spoke to could easily tell the next person they talk to what you shared" (accessible framing that's easy to digest and repeat).

# Synthetic Data and Validation Servers

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- Synthetic data and validation servers offer a promising set of complementary technologies that can be leveraged to provide statistically valid estimates for targeted research purposes that protect the privacy of the records in the underlying data set. These technologies can be advanced through additional research and investment.
- Challenges remain with these technologies, including the following:
    - ***Privacy and utility tradeoff.*** If the data are too "good," the synthetic data set will "leak" information. In addition, a synthetic data set only generates valid estimates for problems it is designed to handle.
    - ***Privacy standard.*** For example, differential privacy is mathematically very neat but provides an unrealistic threat model that assumes an attacker has nearly infinite resources (data to attach with, computing power, time, etc.).
    - ***Privacy budget.*** Users cannot have unlimited access to the synthetic data set. For example, if a researcher publishes thousands of estimates, someone could learn too much about the underlying data set. Possible solutions include tiered access and allowing researchers to do more queries on the validation server but only publish certain information.
    - ***Technological advances.*** Researchers need a user interface and different ways to complete projects with synthetic data and validation servers. Further, researchers must explore opportunities to run complex statistical programs at scale.
- The NSDS could do a lot to move the technology forward, including investing in more open-source tools and training.

### *Summary*

#### What are synthetic data and validation servers, and how are they being used today?

Synthetic data are generated by drawing random values from empirical distributions for sensitive data without using actual records. This provides more robust and systematic privacy protection than traditional Statistical Disclosure Control methods. Instead of taking a survey or administrative data set and protecting against disclosure risk using ad hoc methods, researchers generate random values that are designed to look like tax returns (or surveys), simulating the empirical distribution of the sensitive data. The resulting file does not include any actual tax returns or survey responses and "looks" enough like the underlying data set that the synthetic data can be used for specific research purposes.

Synthetic data can be paired with validation servers to increase data quality, enhance privacy and confidentiality, and allow for testing and debugging of code. A validation server runs statistical programs on a data set and modifies the returned statistics to protect privacy. For example, the validation server may add a random error to a statistic, with the error variance and the standard error of the estimate increasing with the disclosure risk associated with publishing the raw statistic. So, the higher the disclosure risk, the larger the error variance and the standard error of the returned data set.

The blurred estimates produced by the validation server are still statistically valid (and can be published) but are less precise than those derived directly from the underlying data, providing privacy-preserving results. The validation server provides information about and enforces the "privacy budget" of released results for each researcher and across all users. Researchers can use the validation server to ensure they do not release too much information, protecting every record in the data set.

Alternatively, a verification server can be used to provide information about the quality of statistical inference derived from the synthetic data. For example, the verification server might report whether inferences about sign (statistical significance) derived from the synthetic data are consistent with the underlying data set.

Today, some federal government agencies are using synthetic data and even a rudimentary validation server. For example, the Census Bureau has produced a synthetic American Community Survey file. In addition, the IRS has created a validation server for Statistics of Income (SOI) data. Currently, the validation server is subject to manual review; however, as SOI researchers learn more about managing disclosure risks, the process could be automated. The technology is also being used by nongovernmental researchers, such as at Cornell.

### What are current challenges in the use of synthetic data and validation servers?

A key challenge for synthetic data is finding the right balance between privacy and utility. In addition, synthetic data will generally only provide valid estimates for applications they are designed to handle, i.e., where they can match means, medians, etc. Synthetic data struggle, for example, to capture shifts like skewed reporting of data around tax rate changes. Currently, developing complex synthetic data sets can be a slow, labor-intensive, and expensive process. Researchers are trying to change that, but more investment in open-source tools and training would be helpful. The NSDS can serve as a driver for these investments and advances.

For validation servers, a fundamental issue is defining an appropriate privacy standard and then developing and implementing privacy protections consistent with that standard. Researchers must be able to measure and allocate the privacy budget, understand how the budget works, and have an easy-to-use interface to track and interact with the budget. The validation server needs to balance the information it provides. Every analysis leaks information on the underlying data set. For example, if researchers receive information on statistical significance, this takes away from the privacy budget. With unlimited access, a data set would eventually have to be shut down to prevent unacceptable disclosure risk. This can be avoided or postponed by managing user privacy budgets.

### What are possible solutions to these problems?

Validation servers could be set up as part of a tiered access system where users would be vetted to:

- Ensure a legitimate research purpose for accessing the data, and
- Avoid publishing or otherwise disseminating intermediate findings.

These constraints would allow researchers to conserve the privacy budget, so there can be more users and more queries before the privacy budget is exhausted.

While tiered access assumes that users will not try to hack the data, a system could also be designed to prevent attacks in two ways:

- Allow users two privacy budgets; a larger one for intermediate queries that will not be published and a more modest one for the published results, and
- As a failsafe, limit the number of queries.

### What role can NSDS play to drive this forward?

There is a lot more work to do to make validation servers and synthetic data consistently useful tools. While many researchers are working on developing these technologies, their work is still in the early development stages. Many of the ongoing efforts are focused on proprietary solutions in the private sector. Foundations like Ventures, the National Science Foundation, and the Alfred P. Sloan Foundation are supporting public, peer-reviewed research in this area, but there is much room for additional resources and investments. Thus, there may be a role for the NSDS in "turbo-charging" the effort by encouraging more researchers to contribute to this work and by aiding users with varying levels of technical expertise.

# COVID-19 Research Database

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- ***Background.*** The COVID-19 Research Database was developed in 6 weeks, as industry leaders came together on a pro-bono basis to create a new tool to help address the challenges of the pandemic.
- ***Requirements.*** To be successful, the project required various data and technology partners to provide different inputs, resources, and support, including enough data to be significant and to cover enough people, a way to join data in a privacy-preserving manner at the patient level, technology to host the data sets, flexibility to add data sets over time, ability to scale quickly, and an appropriate governance structure. Out of 700 possible data partners, 13 joined the effort.
- ***Contents.*** The COVID-19 database includes 87 billion records covering more than 300 million unique individuals from 13 private data sources and more than 90 public data sources. Using privacy-preserving methods, researchers have joined over 72 million health records, representing over 6 million COVID-19 patients, more than 268,000 COVID deaths, and over 3.5 million vaccine doses.
- ***Users.*** Database users are a diverse group, representing government organizations, non-profits, and labs; medical centers and schools of public health; state governments; and other countries. Over 300 study proposals have been submitted for review with more than 200 projects ongoing and more than 500 researchers.
- ***Privacy and confidentiality.*** The database leverages privacy-preserving record linkage (PPRL) to protect privacy and satisfy regulatory and scientific requirements. The first step in PPRL is de-identification and tokenization; once this process is complete, researchers cannot undo it to re-identify the underlying individuals. The data are then determined to be de-identified based on a HIPAA expert determination standard. The system applies a linking algorithm to match records across data sets. While there are different privacy and governance requirements for various data sets, the data reflect the most conservative standard.
- ***Challenges and opportunities*** include the following:
  - Workspace is limited, so only so many researchers can access the system.
  - When researchers apply to the database, they cannot see the contents to help decide whether the data sets and variables will meet their needs.
  - Certification layers could create fiction for researchers as this process is not fully automated, and there is a time lag for approval.
  - Certifying a common data schema requires a lot of remediation and redaction; this has limited the studies that have been possible.
  - There are challenges to managing privacy-utility tradeoffs with PPRL—different use cases require different levels of precision.

### *Summary*

### What is the COVID-19 Research Database and what problems did it solve?

Data fragmentation hinders health care research, as each institution only has a piece of the information on the patient journey. The pandemic has exacerbated the need for an open research platform with enough data and enough subjects to conduct robust research while preserving privacy. To meet this need, researchers require technology that is flexible and scalable and a legitimate governance structure that ensures compliance.

The COVID-19 Research Database provides a tool to help policymakers and researchers better understand and address the challenges of the COVID-19 pandemic. A coalition of industry leaders, including technical and data partners, came together to develop the database in just 6 weeks. Different organizations play different roles, all pro bono, including the following:

- Datavant onboards data sources, links data sets, and provides technical assistance and researcher onboarding.
- Mirador Analytics certifies individuals and linked data sets as de-identified.
- Medidata manages refreshes and loads, data provisioning, and the analytics environment.
- Snowflake provides cloud hosting services.
- Health Care Cost Institute manages data governance and researcher engagement.

### What drives the success of the autonomous database?

The project's success is driven by governance and research management and by technical infrastructure.

***Governance and research management.*** Governance includes researcher administration, researcher support, data governance, and data privacy and security. The governance process has two separate review portions—a privacy review and a scientific review. For the privacy review, the statistical certification is not automated—someone must look at the data and make the determination. Likewise, the scientific review portion relies on a certifying partner to make determinations, like a traditional peer-review process. While the official requirement for scientific review varies depending on the data set, the posture is to use the most conservative framework when making determinations and carrying out linkages.

Users, who must be noncommercial, are required to sign up as part of audit trail even to access the data dictionary. Users cannot access a snapshot of the contents and then figure out which variables they want to combine. Instead, researchers must work through the entire approval process before touching any data. Once approved, users can access the data dictionary and forums for support.

It is important to keep in mind that it takes time to link new data sources. Common data schema make access quicker, but the biggest problem with that approach is remediation and redactions that limit the number and types of studies that can be done.

***Technical infrastructure.*** Technical Infrastructure includes data source and onboarding, data linking and preparation, data ingestion and regular data loads, data provisioning, and managing the secure analytics environment. A critical component of data linking and preparation is PPRL, which brings data together without compromising patient privacy. PPRL de-identifies individuals through tokenization based on a Health Insurance Portability and Accountability Act (HIPAA) expert determination standard. Once a token, or patient key, is given, researchers cannot identify the underlying person. The same person could be different tokens in different data sets; however, de-identified matching across data sets is still possible.

# Jobs and Employment Data Exchange (JEDx)

## Primary ACDEB Subcommittee: Government Data for Evidence Building

### *Key Takeaways*

- JEDx is an initiative of the U.S. Chamber of Commerce Foundation aimed at standardizing employment information that businesses are required to report to a variety of federal, state, and local governments. The basic idea is that standardized employment data would be reported by human resources (HR) processing first to a single portal and then would be distributed to the various governmental units that need the data.
- Key benefits include the following:
    - Reducing the burden on businesses,
    - Providing enhanced payroll and HR reporting that would add to data currently being collected and that could be used for better, high-frequency measures of the performance of the economy in more localized areas, and
    - Standardizing occupation and credential data to allow workers to use their own data and advance their careers.
- Key challenges include the following:
    - Difficulty in getting buy-in from key players, especially smaller employers who self-report payroll and HR records to government agencies, and
    - Making clear how difficult it is to standardize and enhance data collection across the various entities involved in most data collection efforts.
- Next steps include the following:
    - Early in the project—starting with a few test states and focusing on reporting of unemployment insurance data.
    - This is part of a larger effort by the U.S. Chamber to use technology to increase the efficiency of data reporting by U.S. businesses.

### *Summary*

#### What is JEDx, and what does it hope to achieve?

There is a lot of discussion of the value in enhancing wage record data; however, businesses are often not interested in sharing more data, and the unemployment insurance systems that are used as the basis for current data are not designed to easily capture more data. Through the JEDx initiative, the U.S. Chamber of Commerce Foundation is partnering with states and payroll processors to standardize and improve data at the point of collection. The goal is to develop a standards-based approach to sharing and using data consistently to address an array of jobs and employment issues.

In addition to providing better insight into economic conditions at a more localized level, the hope is that the standardized data will improve job descriptions in a way that helps job seekers better understand jobs posted, improve job posting quality and consistency to support better data analysis, and enhance job searching capabilities.

### What is the initial approach to building consensus and realizing improvement?

The initial focus of the project is on data collection, aimed at improving federal and state unemployment insurance reporting processes. Businesses are required to report similar, but not always identical, data to multiple distinct agencies for different purposes. The hypothesis is that if agencies could align on a standard set of data to serve their purposes, then payroll processing companies could integrate those standards into payroll systems for both large and small businesses leading to more efficient reporting and higher quality data.

Recognizing that getting 50 states on board at once would be challenging, JEDx is testing this standards-based approach with seven states (Arkansas, Colorado, Kentucky, Texas, California, Florida, and New Jersey) to develop common definitions and to begin building the system architecture. The pilot focuses on the unemployment insurance program, given the high visibility of these data during the pandemic and generally high value as economic data. These data provide a good test case because of the lack of standardization for definitions as fundamental as what a "job" is, as well as the systems involved across states. Today, employers and data processors report to dozens of systems, each coded differently, so there is great potential for efficiencies and savings from increasing consistency across jurisdictions. In addition, there are current unemployment insurance system modernization efforts and investments underway that align well with the goals of the project.

### What are the key milestones and lessons learned along the way for JEDx?

The initial JEDx efforts are as much about the process as the actual results. The project seeks to identify compelling use cases, to target leaders and advocates in states to participate (public sector and public-private coalitions), and to develop priorities for data use and system architecture elements. The goal is to stand up an ambitious project that achieves consistency across the seven interested states and their programs and lays the groundwork for a larger constituency around data system improvement.

As such, JEDx is currently in the coalition-building process. Project leaders are working toward consensus on the specific data elements, focused initially on a small set of data elements where they can find consistent definitions. From there, the project will evaluate other important variables (for example, on occupation, demographics, and equity and inclusion).

# Privacy Threats and Re-identification Risks

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- Traditionally, there are three main types of privacy and confidentiality risks: identity disclosure, attribute disclosure, and inferential disclosure. Researchers and statisticians have been measuring these risks for decades, mostly on an ad hoc basis.
- In contrast, differential privacy frames the risks by defining a maximum privacy loss that can result from a data publication or statistic under the absolute worst-case scenario, where an attacker has unlimited computing power and other resources.
  - Under this framework, the privacy loss budget is set to adjust the tradeoff between privacy and accuracy. On the one extreme, the accuracy may be so high in the resulting data set that it would make sense to release the original data; on the opposite end, the data are so obscured that they are nearly useless, and it does not make sense to release them at all.
  - Differential privacy composes and computes the total privacy loss from multiple individuals and releases of data publications or statistics.
  - Currently, no standard framework exists for applying these concepts to policymaking. In theory, there's a turning point that helps identify the optimal tradeoff between privacy and accuracy; however, this only applies when looking at specific use cases, and there have not been enough use cases researched to determine if there can be a more generalizable optimization point. For example, even if researchers figure out the optimal tradeoff for census data, this may not apply to health data or tax data. The considerations are nuanced—it's not just about how the data are structure but also about the context of the data and the uses for those data.
- It is difficult to develop a new privacy definition—it took years to come up with differential privacy. The National Secure Data Service should support more research on how to balance traditional approaches and more formal privacy methods. There are possibilities between these two options that could take the framework for differential privacy and relax it to focus more on the data sets themselves.
- The target of what is important to "protect" is often elusive. For example, even though someone may not know an individual's salary from a data set, it may be too personal even to know the average salary of a small, definable set of individuals.

- Education and transparency are keys to increasing the public's understanding of privacy risks and appropriate tradeoffs between privacy and accuracy.
    - Agencies have not clearly communicated their methods to protect privacy and the quality of published data sets using traditional approaches. It is important to be transparent about these methods and new frameworks like differential privacy. This way, the public can weigh the pros and cons of the old and new methods.
    - There needs to be a mechanism that teaches a broad audience a whole new concept around privacy; however, there are not the right incentives in place to do this. One approach could be to build expectations on communications into grant programs. Of course, this incentivizes people already in the field, and there is a need to incentivize more people to work in this area.

### *Summary*

#### Why is re-identification risk so challenging?

Privacy experts work between data users and data stewards to help preserve both the quality of the data released and the privacy of the data subjects before information is released. As a result, what usually comes out is public microdata or summary tables and statistics. Yet privacy and re-identification risk remains a challenge.

In many cases, even when researchers remove personally identifiable information like names, social security numbers, and zip codes, identities can still be determined. Even though the released data may seem safe, researchers do not know what everyone else is releasing that may be used in concert to identify an individual. There are many examples from the private sector where data stewards assumed privacy protections had been taken, yet sensitive information could be recreated from combinations of data and metadata. Examples include using changes in shopping searches to identify pregnancy in Target Corporation data, determining sexual identities based on viewing habits in Netflix, Inc. data, and using location data from smartphones to determine when individuals switched jobs and then identify those individuals through LinkedIn.

#### What has re-identification risk measurement traditionally looked like?

There are three primary traditional types of risk:

1. Identity disclosure risk
2. Attribute disclosure risk
3. Inferential disclosure risk

***Identity disclosure risk.*** This is the most common risk materializing in cases where multiple sources are linked to re-identify data. In one example, Latanya Sweeney, a Harvard researcher, was able to identify individuals' medical procedures and diagnoses by linking voter data with the personal genome project data set based on zip code, gender, and birth date data.

***Attribute disclosure risk.*** This is the ability to identify group characteristics from a data set. For example, using location and movement data during COVID-19 lockdowns, users could identify essential workers and thus groups who may have higher probabilities for long-term effects of COVID-19.

***Inferential disclosure risk.*** This is when information can be inferred from a data set with a high probability. Using the example above, once a group has been identified, insurance companies could infer whether an individual is at higher risk for long-term effects of COVID-19 and use that information to raise health premiums.

These risks have made up the focus of traditional statistical disclosure control, or limitation, statistics. Managing these risks is difficult because it requires data stewards predict people's actions and anticipate the contents of future releases.

### How can differential privacy help measure and mitigate re-identification risk?

Differential privacy provides a distinct way to think about the aggregate risk of subsequent disclosures. First, it defines a maximum privacy loss that can result from a data publication or statistic by considering the absolute worst-case scenario. Differential privacy asks users to think about what other data may exist and what potential, new data that could come later. In addition, it frames the risks from the perspective of someone with unlimited time and computing resources (like a supercomputer) to attack the data.

Second, differential privacy features a privacy loss budget, epsilon, that can be used to adjust the tradeoff between privacy and accuracy. As epsilon approaches infinity, accuracy increases toward that of the original data such that data stewards might as well release the original data. As epsilon approaches zero, accuracy is so low as to make the data relatively useless to the point of there being no point in releasing the data.

Finally, with this framework, data stewards can compute the total potential privacy loss from multiple individuals and the releases of data publications or statistics. By identifying the cost at the component level, data stewards can see what it takes to "break" the privacy loss budget in any given set of conditions.

While this is still an emerging field, identifying the ideal point at which to set epsilon, in theory, is a point where the data utility curve begins to flatten out. In the real world, however, there are not enough case studies yet to identify that optimal point, and that point may vary for different types of data sets and use cases.

### What should be done to help put differential privacy to use as part of the NSDS toolkit?

While theories in this field are peaking, practical application remains minimal. The NSDS should retain flexibility for what emerges as applications evolve. The NSDS can help facilitate research and development of techniques to balance traditional approaches and differential privacy. Additionally, the NSDS can help bridge the communication gap, reaching beyond computer scientists and allowing data users and data owners to understand differential privacy and to become comfortable with this framework as an approach to privacy risk management.

# Federal Statistical Research Data Centers (FSRDCs) Panel

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- Even though FSRDCs serve relatively sophisticated users (e.g., university-based, government, and nonprofit researchers), they provide a significant amount of technical assistance across the entirety of the research process. This includes but is not limited to developing a researchable question, discovering relevant data sets, meeting agency and steward requirements for access, analysis support, and disclosure review.
- There is significant variability in the price structure across FSRDCs for access to and use of their services. This is due largely to factors associated with the host institution, including the level of financial support it can provide and the availability of other local and regional partners to share in the cost of the FSRDC's operation. The high price associated with the use of some FSRDCs likely limits some users from exploring questions of value.
- The FSRDC system represents one effective model for providing qualified users secure access to high-quality data meant to be used for statistical purposes. Achieving that success has taken, and continues to take, significant investments in human, technological, and financial resources.

### *Summary*

#### What is the process for gaining access to FSRDC data for a project?

Obtaining access to an FSRDC is a multi-step process that varies across data-owning agencies. As an example, here is a description of the process for researchers requesting access to Census Bureau data:

- ***Proposal development.*** Researchers usually begin the proposal process by emailing the local FSRDC administrator. The researcher is asked to write up a paragraph or two describing their research question and listing the data sets they may like to use. A meeting is scheduled with the FSRDC administrator to go over how the system works, the research question, whether the data requested are likely to support the research question, what other data might be helpful as complements to what the researcher has in mind or alternative data that might be easier to get, and so on. The FSRDC administrator and director review and comment on multiple versions of the proposal, including Census benefits, and may send a version of the proposal to Census data experts for feedback and advice.
- ***Proposal submission.*** Currently, the FSRDC administrator submits the proposal on the researcher's behalf. This will move to the Standard Application Process when it is available.
- ***Proposal review.*** The Census Bureau reviews the proposal and ensures all relevant approvals are in place (e.g., from co-sponsoring agencies or data providers) before approval.

- ***Security clearance.*** The FSRDC administrator provides the researcher with paperwork to start the Special Sworn Status application. Researchers must complete the background check paperwork, submit fingerprints and a photo, and are subject to an interview to determine suitability.
- ***Project start.*** The researcher finalizes any administrative processes necessary for accessing the FSRDC (e.g., fees and scheduling) and works with the FSRDC administrator to schedule the researcher's badge issuance, orientation, and first project logon.

The approval process typically takes 8 to 12 weeks, and fees vary based on the agencies and FSRDCs involved.

### What do FSRDC projects typically look like?

As of April 2022, FSRDCs are hosting about 420 active projects. About 60 percent of those projects use Census Bureau data, or Census Bureau data plus data from the Bureau of Labor Statistics (BLS) or the National Center for Education Statistics. A third of the projects use health (NCHS or Agency for Healthcare Research and Quality) data, and the rest use only BLS or Bureau of Economic Analysis data. Project durations vary by agency, with the breadth of the research questions ranging from 1-year renewable projects to 5 years. Most projects using Census Bureau data have a 4-to-5-year duration.

Most projects involve academic research. There are only a small number of projects in the FSRDCs that do not involve academics, and these are more likely to be research organizations or think tanks than state agencies. The FSRDCs are very open to hosting state projects if there is a funding mechanism to support them, either through seat fees or participation in local FSRDC consortia. Since academic organizations and Federal Reserve Banks have provided the funding to establish and maintain the FSRDCs, those are the organizations that have traditionally had access.

### What does the FSRDC facility network look like, and how can it be accessed?

Across the growing FSRDC network, there are 281 total terminals, ranging between 5 and 19 per location. The academic institution hosting the FSRDC pays for rent and facilities, while the Census Bureau provides the information technology and security equipment at each location. The 12 branch locations provide funding to the Census Bureau for the IT equipment purchase. The Census Bureau pays for IT equipment at the 20 core locations.

FSRDCs offer virtual access at the same cost as in-lab projects, and the types of services provided by the FSRDCs (except for the seat itself) are all the same. Some data files are restricted to use in the lab or with an administrator present.

### How are FSRDCs staffed?

All FSRDC locations have an executive director (or co-directors) and an administrator, and a few have additional staff. Staff attributes and roles include the following:

- ***Executive director.*** Secures funding and monitors the budget, maintains partnerships with other institutions, advocates for researchers with the Census Bureau and other relevant agencies, supports the FSRDC administrator as an informal supervisor, participates in network activities and coordination, and does outreach to new researchers.
- ***Administrator.*** A Census Bureau employee who is physically located at the FSRDC oversees the virtual and in-person lab and its security, guides researchers through security clearance and orientation to the environment, provides feedback to researchers on project proposals, reviews disclosure requests, and does outreach to new researchers.
- ***Other staff.*** Some FSRDCs employ graduate research assistants to help researchers in the lab with projects, and many FSRDCs employ clerical staff to support the administrator and executive director.

Both administrators and executive directors may provide technical assistance across the project lifecycle from proposal development through project execution. For example, during proposal development they may answer questions about the data, including information about variable and data set availability and other research using the data that is not available through other sources.

Staff is on site at least 20 hours per week with some data restricted to use only when an administrator is present. Because of the support that staff provides helping researchers understand the proposal process and data, it is not anticipated that FSRDC employees could easily be replaced by technology.

### What does FSRDC governance look like?

The FSRDC Program Management Office oversees the FSRDC program and ensures all access and procedures are compliant with relevant regulations and authorities. Agencies determine data access fees, and local FSRDCs established their own fee structure for consortium membership and external projects. The major cost of local FSRDCs is paying a federal employee (FSRDC administrator), with smaller amounts for additional fees for the FSRDC Program Management Office and salary support for the executive director. Terms and costs for federal employees are not set by the local FSRDC.

Statistical agencies set the guidelines for data access and use and approve the use of their data consistent with applicable regulations and statutory requirements. All researchers working at an FSRDC must have Special Sworn Status, which currently requires U.S. residency for 3 of the last 5 years and a favorable background investigation. Statistical agencies may set additional requirements, such as citizenship. All research conducted through the FSRDCs must be statistical in nature, and users may not conduct research for regulatory or enforcement purposes nor be employed in a regulatory or enforcement position.

# State Wage Interchange System

## Primary ACDEB Subcommittee: Government Data for Evidence Building

### *Key Takeaways*

- While the State Wage Interchange System (SWIS) is facilitated and coordinated by the Department of Labor and the Department of Education, it is a voluntary agreement between states in a system that the states approve.
- SWIS allows other state agencies to collect information on participants that move to other states only with approval from all signatories to the agreement. For these purposes, only group information is provided, and all groups must include at least three people. Any evaluation efforts also need agreement from all states, so using data for this purpose is challenging.
- SWIS demonstrates that it is possible to get states to agree to share data but that the agreement may be limited to a very specific purpose and for very limited uses.

### *Summary*

#### What is the State Wage Interstate System?

SWIS is an agreement among states to share unemployment insurance wage record data, so state workforce agencies can track individuals who participated in Workforce Innovation and Opportunity Act (WIOA) activities but then moved to another state. The program is primarily for reporting required outcome measures as part of WIOA and makes use of the Unemployment Insurance Interstate Connection Network system.

The Department of Education and the Department of Labor have established agreements with all 50 states, the District of Columbia, and Puerto Rico. The agreement is voluntary to share wage data and is limited to only interstate wage data. For all states, wage records contain employer and employee names, social security numbers, federal employer identification numbers, state tax identification numbers, associated wages, and North American Industry Classification System codes; some states also include occupation codes. These data are primarily collected for tax purposes and unemployment insurance funding. In SWIS, the wage data are stripped of individually identifiable information and are aggregated in groups of no fewer than three records. Usage of the data are limited since they are considered confidential unemployment compensation information.

### Who is involved in the SWIS agreement?

The parties to the agreement include the state unemployment insurance agency who holds the wage data; the Employment and Training Administration with the Department of Labor who administers the WIOA program; the Office of Career, Technical, and Adult Education; the Office of Special Education and Rehabilitative Services/Rehabilitation Services Administration in the Department of Education; and up to six agencies designated by state governors as being responsible for coordinating or facilitating the assessment of one or more of the state's WIOA core programs, known as Performance Accountability and Customer Information Agency (PACIA).

There are two types of PACIAs—access PACIAs and non-access PACIAs. Access PACIAs, one per state, make requests for wage data from the clearinghouse either for their own purposes or for the benefit of another entity, as permitted under agreement. Non-access PACIAs, up to five per state, receive wage data through their access PACIA.

Only PACIAs and designated contractors or agents can use wage records obtained from SWIS for federal performance reporting and only for named programs in the agreement. There is no national database of wage records, and federal partners do not have access to individual-level data.

### How does the SWIS work?

The process starts when a PACIA makes a request for data and sends a list of social security numbers for which they do not have wage record matches within their state. The clearinghouse looks for a match in its Distributed Data Base Index, which maintains an index of social security numbers reported by participating states up to the last eight quarters. If there is a match, the clearinghouse sends it to a state that has the wage record itself, which queries its database and sends a reply via the clearinghouse. The clearinghouse passes the information back to the requesting PACIA and destroys the file. Federal involvement is limited to providing oversight and ensuring the process occurs per the agreement; federal agencies do not have access to or make use of the data.

Data usage is limited to specified purposes such as:

- Federal performance assessment and reporting for SWIS-approved programs; ultimately made available for public use. Programs use aggregate data and some individual-level data.
- Eligible Training Provider certification and consumer reports, which are available publicly.
- State-mandated performance assessment and reporting, when approved by federal partners.
- Qualified research projects subject to approval of individual SWIS states electing to participate.

### How might SWIS change in the future?

There are currently six proposed amendments in the comment period process. The process includes two comment periods and requires signatures from all members to the SWIS agreement (though that may be changed by one of the current proposed amendments). This makes the process long, and it is difficult to get changes to the agreement approved.

The six current amendments address expanding eligible program uses, formalizing the ability to publish de-identified and masked records in public-use files, expanding wage data access, disclosures, and shelf life, along with changing the amendment process to allow for phased-in approaches that will make it easier and quicker to realize new benefits.

### What can an NSDS learn from the SWIS experience?

To encourage states to voluntarily participate, there needs to be a framework that is explicit about limitations on data usage, access, and reporting. This requires a balancing act between enabling access and maintaining control to ensure everyone understands the rules for use and access, particularly when making individual records available.

Public-use files can serve as a compromise, providing transparency and some form of public access to the information. It is important to understand which data and uses are sensitive, including the ways data may be combined that are not always obvious.

# Inter-university Consortium for Political and Social Research

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- The Inter-university Consortium for Political and Social Research (ICPSR) has a multitude of resources, analyses, data sets, and collections—as do many other organizations that have presented to the Other Services and Capacity-Building Opportunities subcommittee.
- A key recommendation in the technical assistance space is a need for "navigation" assistance—pointing policymakers to data collections and organizations that already exist—not just a concierge for the federal statistical ecosystem. Just navigating the multitude of acronyms is going to be a barrier for state and local policymakers.
- A key goal of the NSDS should be to increase data literacy broadly.

### *Summary*

### What is ICPSR?

ICPSR, based at the University of Michigan, was founded in 1962 by 22 universities to support the sharing of the American National Election Study. Today, it is a consortium of about 800 institutions worldwide, including colleges and universities, statistical agencies, and research think tanks, which focus on social and behavioral science data. The consortium currently holds over 17,000 studies with a quarter million files; about 10 percent of the data sets include restricted data. ICPSR has multiple collections supported by the consortium and topical collections supported by external funders.

While ICPSR seeks to expand its user base, such as with its Open Data Flint project, currently, most users are graduate students and others in academia.

ICPSR is committed to data stewardship by producing, sharing, and preserving high-quality and curated data sets in user-friendly formats. The consortium is committed to data protection with multiple strategies for safe research using sensitive data.

### What technical assistance does ICPSR provide?

Technical assistance provided to data user community includes the following:

### *Computing and Network Services*

- A team who manages security plans and designs and maintains infrastructure. The infrastructure is critical to help users find data, request access to data, manage researcher access accounts, provide training, monitor training, and automate data access control.

- The infrastructure includes tiered access for restricted data with encrypted downloads to safe spaces, virtual data enclaves that can be access from home, physical data enclaves, and secure online data analysis. The tiers have evolved over time, as it has become increasingly important to ensure sensitive data are available to researchers in safe ways as ICPSR better understands re-identification risk; for example, how increasing data linkages increases identifiability.

***Privacy and Security.*** A separate team, led by the Chief Privacy Officer, works with IT and develops templates and revisions to legal agreements. With researchers and institutional sponsors signing agreements for every data set, there must be agreements, lawyers, computer systems to manage agreements, and support for researchers to find institutional sponsors and figure out how to enter into an agreement.

***Project Management and User Support.*** Staff in this area act as the front line for providing direct assistance to data users. They maintain a user support ticketing system and user forums for specific data sets to allow users to talk to each other about data. The team also creates user materials about data sets, trainings, and webinars. Within the virtual and physical enclaves, this group completes disclosure reviews following the guidelines set by the privacy and security group.

***Data Curation.*** The data curation team works on making data more accessible before users even get to it. They enhance data reusability and access through adherence to standards and procedures, including disclosure risk reviews and remediation, creating study documentation, developing standardized metadata, and preparing data for dissemination and online analysis.

***Metadata and Preservation***

- The team reviews the metadata produced by the curators ensuring the data properly use internationally recognized standards that support interoperability. Staff also conducts knowledge transfer through multiple channels, such as the ICPSR Bibliography of Data Related Citations, research spotlights, and instructional modules.
- A key idea is that knowledge transfer is critical for increasing the reach of the data. For the NSDS, this should be incorporated from the beginning.

Currently, the technical assistance team consists of 40 people in data curation, 19 in project management and user support, 4 in privacy and security who work daily with University of Michigan lawyers and contract managers, and 50 in computing and network services. This team can be supported by additional resources within the individual collections.

### How could the NSDS help an organization like ICPSR with its mission?

The NSDS should increase data literacy broadly. ICPSR tries to do this through programs for high schoolers and outreach to community colleges, among other programs. Data literacy is critical as a national strategy, and the NSDS could be an important key to accomplishing this. The NSDS could help get buy-in from state and local partners and serve as a resource for people across the country, not just in academia or in Washington. In this way, the NSDS could be quite transformative.

# Privacy-Preserving Solutions for the Future and Risk Evaluations

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- The speed at which technologies develop is faster than that with which legal agreements allowing for data availability using those technologies can be developed. Experiences with XDATA and Memex show these challenges—even when there is a strong value-add and available technical solutions, it is often not possible to share data for evidence-based decisionmaking.
- The DataSafes program is exploring ways to provide better access to data, regardless of privacy constraints, to support R&D for societal good. Recent work with the Advanced Education Research and Development Fund (AERDF) provides an example of how progress can be made currently. To create a usable data environment, this initiative took 2 years, a 12-person project team, and the drafting of 100+ pages of policy documents that sit on top of legal agreements and policies.
- In the long term, it must be easier to share, link, and analyze data. The law and technology must co-evolve. Technology demonstration, and risk assessment associated with it (like those under DataSafes), should inform policies written by lawmakers.

### *Summary*

#### What are some examples of attempts to combine data sets in a privacy-preserving manner that were hampered by legal challenges?

***XDATA.*** An effort driven out of the White House to make health data available for cancer research. The project was aimed at linking Department of Defense (DOD) health records data for a project with the Veterans Administration (VA). The combination of genetic sequence data from a million veterans and private health data from private healthcare providers into a supercomputing facility would have allowed cancer researchers to perform in-depth analyses.

In practice, providing the data analysis tools to accomplish this was easy; however, after 2½ years of effort, the project never came to fruition, largely because researchers could not get facilities at DOD and VA to share medical records with each other due to security and privacy concerns over the data. Despite data sharing being technically allowed through HIPAA portability, the effort was mired in legal arguments that prevented data linkage and subsequent research that would have supported the shared goal of curing cancer.

***Memex.*** Memex is a program designed to better track human trafficking. The data needed to be shared between state departments, various law enforcement agencies, and open data sets. The project revolved around state and local entities sharing data with the State Department via visa applications and arrest records. Eventually, the data were housed at the National Center for Missing and Exploited Children.

The program built a system that aggregated data sets held by a third-party nonprofit that could hold data shared by state and federal entities. However, it took 4 years to negotiate the relationship, and policymakers had to extend the program beyond its technological development phase to accommodate the legal issues.

### What is DataSafes and how might it help?

The DataSafes program finds ways to provide better access to data, regardless of privacy constraints, to support R&D for societal good. Emerging technologies could enable data processing in private ways with useful results for policymakers and researchers. If a facility existed to enable this in a productive way (e.g., build analyses, conduct analyses, and generate meaningful insights), then the world can change, as decisionmakers can start using data without the delays and legal hurdles faced in the past.

An example of this in action would be the secure environment for R&D in education built by AERDF. Many partnering school districts currently contribute data to the enclave, and they need ways to secure, analyze, and conduct research on these data. While these partners want to build the next generation of technology, they also have to balance the immediate need to deploy a solution today that can accomplish this.

### What can be done to overcome these hurdles, as seen through the AERDF experience?

In the education environment, there are legal frameworks, like the Federal Educational Rights and Privacy Act (FERPA), that allow for data sharing. This law and the related regulations provide cover for people to share data and outline parameters around sharing. Once data owners know there are systems, processes, and regulatory compliance, they are much more willing to share data. However, this serves as the minimum bar for entry to enable organizations to share their data.

There are additional constraints around assessing real-world risks, such as cybersecurity, that are extremely important for figuring out whether data owners are willing to share data. AERDF has partnered with 30 districts. What has allowed this to happen is a legal framework and all sorts of security testing, such as cybersecurity testing and auditing around disclosure control, privacy, etc. It has taken the group a year and a half to gain approval and acceptance from key stakeholders. This has been accomplished by establishing a legal framework for compliance, providing empirical evidence that the systems being deployed have protections associated with them, and coupling it all with procedures and practices associated with how these systems implement risk assessment, auditing, and tracing.

A key to facilitating this has been the mutual agreements among educators and nonprofits around the National Research Data Privacy Agreement. This is a standardized agreement by nonprofits and academics at the University of Pennsylvania that creates a contract between schools and researchers on the use of student data for research purposes, including compliance procedures to help ensure privacy and security (beyond standard cyber procedures, HIPAA, and FERPA).

In addition to the actual legal agreement, another key to success is a corresponding set of documents that translate the legal agreements for the teachers, educators, and administrators working with the data (e.g., what it means to anonymize data, de-identify data, work with that data, and have control processes to release results). For AERDF to achieve this, it took technical infrastructure, as well as 100 pages of policy documents about how to treat student data and statistical disclosure control at a level of detail where states feel data are compliant and protected, requiring a full day of training for real-world users to start working with the system.

### Is this a sustainable approach?

In the long term, it must be easier to share, link, and analyze data than it has been for AERDF. It should not take 2 years, require a team of 12 people, and the drafting of 100+ pages of policy documents that sit on top of legal agreements and legal policy to create a usable data environment. Data owners should be able to easily formalize their policies and turn those into privacy guarantees in a way that allows them to publish data more freely. Users should be able to access data in a larger evidence marketplace.

That is where technology like DataSafes and privacy-enhancing technology will eventually go. Researchers should be able to publish protected data tied to policies that allow for users and environments processing those data to comply with policies in an automated fashion managed by systems that are continuously monitoring the usage of data to ensure they are being protected appropriately.

### What is the next step in enabling the use of privacy-preserving technology?

The root problem is that even if there are technologies that protect privacy, there aren't legal frameworks that permit the use of these technologies for evidence building. The law and technology must co-evolve. Technology demonstration, and risk assessment associated with it, should inform policies written by lawmakers. There needs to be a FERPA 2.0 that takes privacy-preserving technologies into account and a HIPAA 2.0 that includes specificity on what portability means in terms of privacy tradeoffs around it. There must be a systematic co-evolvement. On the technological side, the responsibility is to provide quantifiable risk assessments that inform executable decisions around how to write policies and laws on the legal side.

# Opportunity Insights

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- Legal constraints or interpretations often make combining different government data sets challenging or impossible. Due to the decentralized nature of the federal statistical system, the issues are not just legal or regulatory but include coordination problems handling data from multiple agencies.
- The lack of transparent, standard disclosure review processes complicates the approval process for releasing results without necessarily improving the privacy and confidentiality of the data. Processes at the Census Bureau illustrate the challenges with disclosure review:
  - Over the years, processes have shifted from simpler privacy-protecting approaches to methods that are more formally private but require a review of results without clear expectations of what researchers must do to gain approval.
  - Researchers have recently introduced a set of procedures that improves disclosure avoidance both from a privacy-protection standpoint and a process standpoint.
  - This new approach, however, has not been widely adopted for disclosure review at the Census Bureau; the issue is compounded by different disclosure review methods across institutions.

### *Summary*

#### What are the challenges from a researcher's perspective with combining data sets?

Legal constraints or interpretations often make combining different government data sets challenging or impossible. Due to the decentralized nature of the federal statistical system, these issues are not just legal or regulatory—there are also coordination problems when handling data from multiple agencies.

For example, since its early days, the Biden Administration has been asking questions about student loan policies and racial disparities in student debt. The federal government already has the optimal data to answer these questions (IRS data on household structure, Department of Education data on student loans, Census data on demographics); however, researchers have not been able to combine the data sets because the data-sharing agreement that would have allowed this linkage has expired. Over a year later, it is still not possible to generate the answers to these questions.

The big difference between the United States and other countries, such as Scandinavian countries, is not found in the data that are collected or the legal and regulatory regimes. Instead, other countries do not have data collection scattered across 13 agencies and thus do not have the coordination problems that are found in the United States.

### What are the challenges from a researcher's perspective with releasing the results?

The lack of transparent, standard disclosure review processes complicates the approval process for releasing results without necessarily improving the privacy and confidentiality of the data. For example, at the Census Bureau, there have been three disclosure review "regimes" in recent years.

Historically, disclosure controls included things like not publishing cell sizes that were too small or not publishing exact counts of numbers. While these approaches did not protect privacy and confidentiality well, they offered a smooth process for researchers to follow. More recently, the Bureau has taken a more formal privacy approach that assumes every data point and analytical table is a risk. While this method appears to elevate privacy and confidentiality concerns, it has created a process where researchers must iterate with the disclosure review board to tweak analyses without a clear direction or known goal of what must happen to get the results approved.

The latest development is that researchers have led the charge to streamline the disclosure review process and improve privacy outcomes. For example, in the Opportunity Atlas project, researchers worked with experts inside and outside the Census Bureau to develop an algorithm that is not formally differentially private but that reduces the part that is not down to a single statistic at the state, race, and gender level—a level appropriate for robust analysis. In addition, the broader privacy community produced referee reports that support the privacy and confidentiality of the results.

This new approach lays out a set of procedures that improves disclosure avoidance at the Census Bureau both from a privacy protection standpoint (better for privacy than doing the simpler, more traditional approaches) and process standpoint (provides a clear set of guidelines for gaining approval). Rather than creating data sets that are formally differentially private, this process gets risk down to a level that is measurable, understandable, and comfortable. If a researcher follows this approach, the results can be disclosed.

It is important to note that this new approach to disclosure review has not been widely adopted at the Census Bureau. Most projects are still evaluated using methods that rely on a table-by-table review of the results. In addition, disclosure review methods work differently in each institution.

# Governance Considerations for the NSDS

## Primary ACDEB Subcommittee: Governance, Transparency, and Accountability

### *Key Takeaways*

- A hybrid model, based around the government-owned contractor-operated model, can provide a basis for an NSDS that achieves current goals and is positioned to grow for the future.
- Funding and operational flexibility are needed to ensure agility and the ability to continue to mature and grow.
- Federal and nonfederal stakeholders should have a voice in the oversight and governance of the NSDS.

### *Summary*

#### What is the NSDS in this model?

The core activities of an NSDS include hosting a secure infrastructure where researchers: (1) submit proposed research projects for approval, (2) link and access data for research and analyses, and (3) have research results privacy protected then prepared for public dissemination. Additionally, the NSDS should be a hub for the broader network of the federal statistical ecosystem, playing a role in coordinating the nodes of a collaborative system guided by a coherent governance process.

The NSDS is more than a physical location for linking data securely and is the place that provides a coherent governance process across these activities that may be taking place physically in other enclaves. The data service provides users a way of centrally knowing what research is being done, where they can find it, and how they can connect to other researchers. The NSDS should provide a set of guiding principles for sharing data across the federated statistical system.

#### What is the role of governance in this NSDS model?

Governance and oversight are key to establishing transparency and trust. The NSDS needs to demonstrate to key stakeholders and oversight bodies that it is operating according to broadly accepted principles and practices that are ethical and equitable and that the work being produced is independent and of high quality.

To accommodate both the “here and now” and the future vision, the NSDS must be able to handle core functions and allow for the hub to develop as it matures. The data service must have a scalable governance structure that will accommodate growth and multiple means of funding.

The governance structure should include federal stakeholders, such as the Interagency Council on Statistical Policy, the Chief Data Officer Council, the Evaluation Officer Council, the Chief Information Officers Council, the Federal Privacy Council, and National Center for Science and Engineering Statistics (NCSES) Director, as well as external stakeholders, such as state government representatives, privacy experts, ethics experts, and representatives from the research and evaluation community.

## What are the potential models for an NSDS?

Hart and Potok considered three models for the NSDS—government owned and operated, government owned and contractor operated, and government owned and grantee institution operated.

### *Government Owned and Operated*

A government owned and operated model would be fully staffed by federal employees, such as the Department of Transportation's Volpe Center and the U.S. Department of Energy's National Energy Technology Laboratory. There are two examples of how this could be implemented.

In the first example, the NSDS would have external FACA committees, generally reporting directly to NSF, and a steering committee made of representatives across federal stakeholders. NCSES would then have its own programs but would also need to establish an NSDS program office that would interact with outside client stakeholders—both agencies and researchers.

In the second example, instead of FACA advisory group and steering committee, there would be a combined Board of Directors established as a subcommittee of the National Science Board. This would likely require legislation and would have less leverage over NCSES.

### *Government Owned and Contractor Operated*

This would be like a standard FFRDC model as seen in Department of Energy labs. There would be Steering Committee made up of federal stakeholders. The contractor entity would be directly responsible for executing per the contract and would have a Board of Directors and Science Advisory Board that report to the contractor. The contractor would run operations and reach out for collaborative partnerships. The contractor would be more agile since it is outside of the federal government bureaucracy, which gives more flexibility and creativity in establishing cooperating agreements and partnerships.

### *Government Owned and Grantee Institution Operated*

This model is like the Mathematical Sciences Institutes but substitutes a grantee for a contractor. The entity would still have a Steering Committee, but NCSES would provide a grant to the grantee institution. The government would have less control over the grant execution than with a contractor.

### What is the recommended model for the NSDS?

None of the three “pure” models perfectly fits the desired attributes for the NSDS, so Hart and Potok developed a hybrid model that featured key aspects of each model. The recommended model is a spin on a government-owned and contractor-operated model but incorporates a project approval process that allows data owners to retain control of their data.

The model includes NSF, NCSES, and a Policy Steering Committee forming the core federal oversight. It is operated by a contractor who would maintain a Board of Directors that includes the various stakeholder and outside experts, as well as a Research and Technical Advisory Board consisting of representatives from areas such as the academic community, data science, computer science, and the cutting edge of privacy protection. The entity could be an LLC that is a partnership of multiple contractors, run by an executive director who oversees NSDS operations.

# Linking Confidential Data for Health Research: PFAS Applications

## Primary ACDEB Subcommittee: Government Data for Evidence Building

### *Key Takeaways*

- The Environmental Protection Agency (EPA) has established multiple data layers to help assess per- and polyfluoroalkyl substances (PFAS) at the federal and states levels.
- EPA health researchers have had some success linking restricted-use Centers for Disease Control and Prevention (CDC) and National Institutes of Health (NIH) data sets with PFAS blood serum measurements to gauge health outcomes:
    - The NIH Environmental Influences on Child Health Outcomes program, covering ~50,000 children.
    - CDC's National Exposure Report, a series of ongoing assessments of the U.S. population's exposure to environmental chemicals using biomonitoring.
- Better linkages between a wide variety of EPA environmental monitoring data and restricted-use health data could offer insights horizontally across federal agencies and vertically between the federal and state levels.

### *Summary*

#### How do PFAS exemplify a use case for data linkages?

Since the 1940s, PFAS have been commonly found in homes, businesses, and industry; most people are exposed to PFAS at some point. Given PFAS resistance to decomposition in the environment and humans, there is known or suspected toxicity that could impact a variety of health outcomes. EPA's mission is to protect human health and the environment, using a science-based approach that requires evidence for decisionmaking. To inform regulatory and policy actions, EPA needs a better understanding of how environmental exposure to PFAS impacts public health.

There are a variety of community-level environmental exposure data and individual-level public health data sets that, if combined, could produce rich evidence for decisionmaking. As the federal government moves to expand PFAS regulations over the coming decade, health impact studies based on linked data sets offer immense promise for improving public health policies and programs and for justifying costly interventions. Combining these data sets, however, raises key challenges with privacy, confidentiality, and security.

### What has been EPA's experience linking secure data sets for PFAS research?

To study the effects of PFAS contamination, EPA needs information on the entire lifecycle of the pollutant—not just on biomonitoring and health outcomes but also on the place the pollutant was released and on air, soil, and water migration. EPA has a variety of data sets that provide this information at the state and federal levels, such as data on drinking water monitoring; PFAS production and use sites; surface and wastewater discharge monitoring; spill locations; mapped soil, surface water, and tissue sampling locations; and downstream impacts from water system intake locations.

EPA researchers have had some success linking these data with CDC and NIH data. For example, in one project, researchers compared locations with contamination where water filtration systems had been installed with locations that had never been contaminated in order to identify impacts on birthweight. The researchers found that before the filtration intervention, newborns were more likely to have low birthweight, whereas, after filtration systems were installed, there was no statistical difference in birthweight among unexposed newborns. This research brought together multiple data elements across agencies, including restricted-use data.

There are several reasons conducting research like this has seen limited success, including often requiring a lengthy process. The researcher must craft an application, think through the entire study, and engage CDC and NIH in all steps of the process, all of which requires resources both for the researchers and for the data owners. The researcher often must visit a physical Research Data Center, which raises equity and access issues, along with significant costs even beyond the substantial data management fees. The resulting time between filing an application and finishing the study can be a barrier to the timely development of evidence to inform policies.

### How could an NSDS improve the EPA's use of data for evidence building?

EPA researchers do not have the same experience with data protections that other agencies do, and the agency lacks capacity, resources, and infrastructure. Better linkages between the wide variety of EPA environmental monitoring data and restricted-use health information data could improve insights horizontally across federal agencies as well as vertically between the federal and state levels.

Given the EPA mission around data on the environment and impacts, one promising aspect of the NSDS is its potential to provide the leverage and facilitation to execute data linkages, while maintaining the required levels of security that EPA researchers are not able to do on their own.

There are many promising areas for similar research. For example, while low birthweight has been better studied, other health outcomes have not. Even for those areas that have been studied, many have not been studied for a variety of PFAS, leading to assumptions of similar health effects that have not been validated. Additionally, with access to individual health and exposure records, researchers would be better able to determine how costly treatment is and, thus, provide better monetization of the costs that regulations may be able to address. States also collect data that are sometimes better and higher frequency than those available at the federal level. The NSDS could serve as a bridge between federal and state data to incorporate more local variation into policy research.

In all, the promise of research on enhanced data linkages can inform more effective policies targeting pollution exposure, while also improving retrospective analysis of policies evaluating whether these policies had the anticipated impact. Such linkages could provide more timely, actionable, and policy-oriented research, guiding national policies that are more responsive to local conditions.

# National Institutes of Health Library Data Services

## Primary ACDEB Subcommittee: Other Services and Capacity-Building Opportunities

### *Key Takeaways*

- Highly skilled researchers who are experts in their disciplines may benefit from access to technical assistance managing data and data analysis through the research lifecycle, including locating applicable data, planning new collections, cleaning data, documenting data sets and data elements, conducting and interpreting analyses, and making data publicly available (when appropriate).
- Technical assistance may span a variety of modes, including one-on-one consultations, synchronous group training, asynchronous virtual training, the creation of self-service "job aids," and access to help desk or troubleshooting services.
- Although there is an analytic "stack" that can support work across a variety of fields (e.g., R, Python, Stata, SAS, SPSS, JMP, MATLAB), there are discipline-specific products that are necessary for researchers working in specific fields (e.g., biomedicine or bioinformatics).

### *Summary*

#### What is the NIH Library and who does it serve?

The NIH library provides library, data, and technology services to researchers within the NIH and to an extent the broader Health and Human Services (HHS) community. The NIH Library serves 4,000 full-time researchers, about 2,000 trainees and post-doctorate researchers that come and go during the year, and hundreds of HHS staff members.

The NIH Library operates as a traditional library where people can check out literature, access e-books and e-journals, obtain help with literature reviews, and receive training on how to use abstracts and resources. In addition, staff provide some onsite hardware-based technology services, such as 3D printing, multimedia creation, and virtual reality.

The NIH Library maintains a strong partnership with the National Library of Medicine (NLM), including reciprocal sharing, book borrowing, and strong ties with their data work. NLM is a public library serving the broader public with similar types of references and literature.

#### What services does the NIH Library offer?

The NIH Library provides a collection of data services under three main areas: (1) resources, instruction, and consultations for data analysis; (2) data management; and (3) biostatistics, with a staff of 5 to 7 employees. This staffing level has historically been sufficient to meet demand for their services. All the services are free for NIH staff, rather than being provided through a fee-for-service arrangement.

The data analysis consultation includes assistance with understanding and managing data throughout the research cycle. This can include in-person and remote consultation requests, and a variety of classes from internal resources and external vendors. There are also resources users access directly without staff assistance to help with data and statistical analysis as well as data visualization.

Other examples of NIH Library services include the following:

- Help domain-specific data repositories visualize, process, and clean the data.
- Collaborate with NLM and other federal partners to build data science skills across the workforce.
- Offer instruction through internal and external providers for in-demand skills such as R, Python, SAS, and MATLAB.
- Provide researchers with onsite, and recently virtual, access to high-performance workstations.
- Deliver consultations, classes, and tools for biostatistics, as well as specialized expertise and tools in bioinformatics.
- Provide bibliometric consultations, training, tutorials, and analyses that help to answer questions like: Where is productivity across the work of an institute reflected in publications? What collaborations have been happening? And where are opportunities for new collaborations?

### What are the benefits of NIH's Library data services?

There are many examples of how the data services provide tangible benefit to the community. For example, NIH is instituting new policies next year around data management and sharing that will require researchers to submit data management plans along with applications, conduct follow-up on the plan, manage data and metadata, and make some portion of the data available for sharing. The data services team is helping NIH researchers understand the new policies and abide by them.

# NIH National COVID Cohort Collaborative

## Primary ACDEB Subcommittee: Technical Infrastructure

### *Key Takeaways*

- The National Center for Advancing Translational Sciences (NCATS) uses privacy-protecting record linkage (PPRL) in production for the National COVID Cohort Collaborative (N3C) and is also testing different methods for creating synthetic data.
- NCATS uses a federated dual-factor authentication system administered by a vendor who aggregates identity providers, which is integrated into their technical infrastructure. NCATS staff are testing alternative authentication approaches such as biometrics.
- N3C has different workstreams that handle data acquisition, harmonization, analytics, and quality assurance, engaging in a range of activities related to data curation and standardization
- Recognizing that common data model mapping is critical, NCATS is currently working with different government agencies to create an accessible, dynamic repository of mappings covering the data from the electronic health records through data submission.
- NCATS offers tools, resources, and training with the aim of providing a shared services model, leveraging the federal government's buying power to build a world-class resource that is broadly accessible.

### *Summary*

#### What is the National COVID Cohort Collaborative (N3C)?

The N3C is a partnership, stewarded by NCATS, among NCATS and National Institute of General Medical Sciences research program hubs where collaborators contribute and use COVID-19 clinical data to answer critical research questions to address the pandemic. Combined, it covers almost every state in the country creating a federally funded and operated, unified data network that serves critical research needs.

The project initially started about 5 years ago with the goal of building the infrastructure to support a national network. The COVID-19 pandemic provided a timely use case for combining disparate data using common data models into a unified analytic set. N3C now contains over 15 billion rows of representative data from 74 sites across 48 states.

Since it was not cost-feasible to set up duplicative infrastructures across the country, the value proposition was for NCATS to become the data steward, do data harmonization, and provide network tools for collaborators, and, in exchange, users have access to a harmonized, centralized, secure data sets. This is a partnership between NCATS and Clinical and Translational Science Awards Program National Center for Data to Health (CD2H) with NCATS handling infrastructure ("plumbers") and CD2H handling the science ("artists").

### How does N3C curate and standardize the data?

Enormous effort is put into cleaning up the real-world data and running the harmonization clean-up process every week. The need for curation cannot be underestimated. N3C has different workstreams that help handle data acquisition, harmonization, analytics, and quality assurance. These workstreams engage in a range of activities related to data curation and standardization, including:

- Writing scripts to ingest data from all providers;
- Providing report cards on data submitted, so that providers can harmonize their data sets;
- Developing data visualizations to find anomalies;
- Applying standard value sets, like administrative gender or aspects of meaningful use, like Value Set Authority Center or Logical Observation Identifiers, Names and Codes; and
- Pulling in and cleaning data weekly (16 billion rows in 12 hours).

Certain aspects of data cleaning address known issues. For example, about 50 percent of units are wrong or miscoded, and staff can “rescue” about half of this group. Other problems require a more in-depth interrogation of the data. NCATS recognizes that common data model mapping is critical to maintain harmonized data sets. NCATS is currently working with different government agencies to create an accessible, dynamic repository of mappings covering the data from the electronic health records through data submission.

### How is N3C using privacy and privacy-enhancing technologies?

NCATS is committed to creating a bridge that allows the use of PPRL. N3C has successfully implemented PPRL at 29 of 74 sites and has demonstrated the use of PPRL between enclaves. NCATS is also testing different methods for creating synthetic data in coordination with third-party vendors MDClone and Syntegra. While there has been a lot of scientific validation on the resulting synthetic data sets, there is a lack of validation for security and privacy.

The goal is to release the synthetic data; however, the cost of producing synthetic data is high, potentially including expensive licensing. Additionally, NCATS does not plan to use a validation server and thus strives to produce a very accurate synthetic data set. Furthermore, researchers are concerned with releasing statistics that may present a disclosure risk. In terms of disclosure risk, researchers are focusing on what they can control while acknowledging that is limited in this space where, for example, individuals may “out” their own data when they publicly disclose medical conditions.

### How is N3C governing access and use of the data enclave?

Contributors sign data transfer agreements and send a limited data set that goes through the harmonization process. This process produces three levels of data: (1) synthetic, (2) de-identified, and (3) limited. Then, investigators or institutions who want to use the data sign a data use agreement.

Currently, there are over 2,500 investigators in the enclave including academics, researchers, and scientists. There are rules, codes of conduct, and agreements, but N3C policies are designed to keep the access bar low, at least during the pandemic. When investigators submit their short data use requests, they attest to the codes of conduct and IT training. This then goes to federal staff who provide the minimum level of access for the use case.

NCATS currently relies on a federated authentication system for its whole portfolio, including access to N3C. The authentication system is integrated with other aspects of the technical infrastructure, creating a "virtual research organization." The vendor administering the system is not an identity provider but instead aggregates different identity providers, similar to how login.gov operates. Currently, the authentication method uses dual-factor authentication, but NCATS is testing different providers and methods to vet identity including biometrics.

### What resources does NCATS offer and how does the program ensure that a broad audience can use them?

NCATS provides many free tools and resources to its userbase, like Python, R, and Jupyter notebooks. A wide variety of organizations in the private and public sectors use these resources. Recognizing that many organizations lack the resources and know-how to procure these tools themselves, NCATS is adopting a shared services model that can leverage the buying power of the federal government to build a world-class resource that responds to many users' needs. NCATS looks for "lightweight" solutions that do not overdesign or overbuild in order to give greater flexibility to users and future programs.

NCATS maintains a large machine learning and algorithm repository. These tools are sophisticated and may be challenging to apply, so assigned staff provide technical assistance to users. There are potential problems in the application of machine learning, such as inherent bias, lack of transparency, and lack of reproducibility, so the program provides help for researchers to apply quality machine learning techniques.

In addition to offering these tools and resources, NCATS provides free training conducted by volunteers from the data community. Furthermore, to help ensure equitable access, NCATS assigns dedicated liaisons to self-aggregated domain teams for targeted support. These teams include a data liaison, a trainer for using the enclave, a logic trainer for data elements, and access to the machine learning library. These domain teams are organized around topics like pediatrics and cancer. Common infrastructure across these teams includes office hours and ticketing systems that are integrated with the authentication system.

# Minnesota Labor Market Indicator Data Equity Pilots with BLS

## Primary ACDEB Subcommittee: Government Data for Evidence Building

### *Key Takeaways*

- State and local governments house a lot of high-quality and diverse data, but it is often collected and organized inconsistently and requires considerable cleansing before it can be used in research.
- Both Minnesota pilot projects take advantage of linkages between existing data sets, like the Quarterly Census of Employment and Wages (QCEW) Employer File and the Wage Records Employee Data, which exist in every state.
- The value proposition and framework presented by the pilots are similar and overlap with other identified use cases, such as LEHD, SWIS, JEDx, and the multi-state data collaboratives utilizing the Coleridge Initiative's Administrative Data Research Facility (ADRF) where labor market information is foundational to answering research questions.
- While the pilots offer workforce insights using well-established labor market information, there remain gaps in the data pool for a growing population of workers who are not necessarily attached to traditional employers and industries.
- In the role of data concierge and/or technical assistance, the NSDS should collect and house a searchable inventory of research projects and highlight which data sets that are being used to gain greater visibility as to what types of projects may overlap; the NSDS should leverage commonalities to support better, broader, timelier, more efficient, and more collaborative research efforts.

### *Summary*

#### What was the goal of the Labor Market Information (LMI) data equity pilots in Minnesota?

There has been increasing demand, particularly during the COVID-19 pandemic, for more detailed workforce data to be available for analysis with a strong emphasis on demographic data to conduct studies on equity. BLS is working with colleagues across the Department of Labor to identify specific opportunities to modernize and enhance labor market information for research use. As part of these efforts, BLS worked with several states, including Minnesota, to develop proposals using available administrative data to conduct research pilots.

The goal of the initiative was to conduct demonstration projects that show how states can fill data gaps, providing demographic data to customers and users without creating a new database. The idea was to use only existing databases and administrative records. Minnesota's LMI office proposed two projects that aligned with the initiative's goals to address questions related to equity and demographics of the current workforce.

### What data was available for use in these projects?

The project used two state administration databases, the QCEW and Wage Records Employee Data, combined with Department of Motor Vehicles (DMV) data. The QCEW comes from the Quarterly Contribution Report that each business is required to submit to state agencies. It contains the total number of employees by month as well as total wages and contributions to the unemployment insurance system. The data are augmented by local and county data reports to enhance industry code data.

The Wage Records are also submitted by employers and provide employee-level data for every business. At a minimum, the data include each employee's name, social security number, and wages. Some states include more data elements such as hours worked and occupational codes. By themselves, these records are of limited use for economic analysis. However, when merged with the industry codes and location data in the QCEW the enhance longitudinal database is of more utility.

To make the enhanced datafile even more useful for economic analysis, Minnesota LMI obtained DMV records and linked those with the employee-level data. This added significant richness to the data, including fields like age and gender to enable a more robust equity analysis. BLS can use the Program for Measuring Insured Unemployed Statistics (PROMIS) system and add UI Claimants File data, training program data, and educational statistics to facilitate research and evaluation of those program outcomes.

### What did the projects look at?

Two demonstration projects were done with the Minnesota data. The first project focused on unemployment insurance benefit recipient's reemployment outcomes following the COVID-19 pandemic recession. It tracks workers who lost their jobs and received benefits during the pandemic to see how they fared in the labor market post-recession. By using the Minnesota administrative data merged with DMV records and linked to the PROMIS database they could look at different outcomes across demographics and worker categories to evaluate the impact and emerging equity issues. State administrative records offer the best available tool for conducting this sort of analysis, and this research provided a unique opportunity to demonstrate how those data sets can be used effectively and to study the impact on a large subset of their population.

The second project looked at job mobility over a longer timeframe using the longitudinal wage record data available. They looked at job mobility by demographic groups being able to control for location, age, gender, wage levels, and industry. This served as another demonstration, this time looking not at an acute issue but a longer period, of how linking administrative records can increase the uses of LMI information and produce informative products for data users and policymakers.

### What are the lessons learned from the pilots, and how can they help identify and address challenges of using this data more broadly?

They will be sharing the results of these projects with stakeholders to all states at an upcoming conference, which will be published and discussed broadly to try to increase the application of these approaches in the future. There are still outstanding challenges that they would like to see addressed with future efforts. These efforts require significant resources to review and clean the QCEW and Wage Records. These are not collected for statistical research but that is a viable by-product of their collection, and, as with other administrative data, requires cleaning before it is suitable for those purposes.

A significant outstanding issue is the need for cross-state sharing of wage records to capture data on commuters and worker mobility across state lines. A significant percentage of workers are employed across state lines. In order to conduct comprehensive analysis, the opportunity and incentive for data sharing is required. The other big challenge is protecting the privacy and maintaining confidentiality of individual wage records as those data are merged with other identifying data sets. This is the single most important issue facing users of wage records; convincing legal advisors has been an obstacle within states, let alone for sharing across states.

The key message these projects convey is that this research can be done successfully. They want to encourage others to look at the results of these project and continue to improve on them. Worker job mobility is just an initial example of what can be researched; they can use these data sets to look at multiple job holders, changes in earnings by industry and geography including county data, data on low wage earners by age group, wage and employment differences between gender groups, and many other measures with significant impacts for both economic research and public policy.

# 5. Other Models and Examples

This section contains other models and examples as further references and to provide background knowledge.

## Federal Risk Authorization and Management Program (FedRAMP)

The Federal Risk Authorization and Management Program (FedRAMP), established in 2011, is a governmentwide program that promotes the adoption and use of secure cloud services. FedRAMP's cost-effective and risk-based approach provides a standardized security framework for federal agencies in accordance with Federal Information Security Management Act (FISMA), OMB Circular A-130, and FedRAMP policy. FedRAMP is a public-private partnership that benefits agencies by reducing duplication and inconsistencies while promoting innovative use of information technologies to ensure that standards and processes for security authorizations are transparent governmentwide.

## Five Safes

The Five Safes framework was developed in the early 2000s and has been broadly adopted by various countries and academic institutions, including the U.K.'s Office of National Statistics, Stats New Zealand, Australian Bureau of Statistics, Eurostat, and Coleridge Initiative's Administrative Data Research Facility (ADRF) which was developed in support of the Evidence Commission. This framework provides a standardized approach to mitigating disclosure risk around the following five aspects:

- ***Safe projects.*** Projects are lawful, ethical, viable, valuable, and appropriate to the agency's mission. Users should demonstrate that project goals align with the level of data detail requested, have a valid statistical (not compliance/enforcement) purpose, and provide a public benefit.
- ***Safe people.*** Users are trusted to use the data in an appropriate manner. Users should be appropriately vetted, provided adequate training in data analytics and confidentiality, and agree to all conditions of data use.
- ***Safe settings.*** The access facility or access mode provides data access controls and infrastructure constraints that limit unauthorized use or mistakes. Physical security, cybersecurity, monitoring, and auditing should align with the sensitivity of the data.
- ***Safe data.*** Disclosure risk assessments determine data sensitivity levels, and access tiers are well defined. The data contain sufficient information to support the approved use while minimizing sensitive information (e.g., removal of direct identifiers and the use of synthetic or aggregate data).

- ***Safe outputs.*** Appropriate Statistical Disclosure Limitation methods, including disclosure review requirements prior to the release of results, are applied. The outputs sufficiently protect the confidentiality of subjects in the underlying data. The methods used to protect released data should align with the potential risk to data subjects and should minimize to the extent possible restrictions on future uses of the data.

## Guidance for Implementing National Security Presidential Memorandum 33 (NSPM-33) on National Security Strategy for United States Government-Supported Research and Development

The National Security Presidential Memorandum 33, a 2021 directive, issued information to federal agencies on standardizing disclosure requirements for federal grant applications. The Office of Science and Technology Policy's Guidance for Implementing NSPM-33 provides implementation guidance to federal departments and agencies to make compliance with NSPM-33 as uncomplicated and as transparent as possible, both for federal agencies and for researchers. It provides more detailed advice on disclosure requirements, disclosure standardization, digital persistent identifiers, consequences for violating disclosure requirements, information sharing, and research security programs. The Council on Government Relations additionally released a summary of the guidance that includes comparisons with current NIH and NSF requirements.

## Health Insurance Portability and Accountability Act (HIPAA)

The guidance for Protected Health Information under the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule Section 164.514 has two distinct paths for de-identification of a data set before the data are released. The first is "expert determination" in which experts review and make the determination as to which identifiers are retained or removed. The second is "safe harbor" where a standard set of 18 identifiers that are removed, as well as other variables that alone, or in combination, can be used to identify an individual. While both comply with the law, some residual risk of re-identification always remains.

## National Institute of Standards and Technology Risk Management Framework

The FISMA required the establishment of a tiered framework for information and IT systems based on the risk and magnitude of harm associated with unauthorized access. The established standard for assessing the FISMA level is accomplished with the National Institute of Standards and Technology (NIST) Risk Management Framework (RMF). The RMF is applied to information and IT systems to determine the level of security required based on the confidentially, integrity, and availability of the data and IT infrastructure. The resulting determination is a FISMA risk rating: high, moderate, or low.

## National Artificial Intelligence Research Resource (NAIRR) Task Force Evaluation of Governance Structures for Large-Scale Research Investments and Operations

The NAIRR Task Force evaluated the governance structures of several large-scale research investments and operations. Table S5 presents the results of this evaluation.

### Table S5. Examples of Various Types of Ownership and Administration Entities for Research Resources

| Example | Resource | Organization Designation | Owner/ Administrator | Funding Mechanism | Supporting Agency |
|---|---|---|---|---|---|
| Information Technology Laboratory | Database, software, and funding | Federal Lab: GOGO | Government | Appropriated | NIST |
| Oak Ridge National Laboratory's OLCF | Compute, data, and visualization | Federal Lab: GOGO FFRDC | Government/ contractor | Contract | DOE |
| Vera C. Rubin Observatory (formerly LSST) | Telescope, data | PPP | Consortium | Award (cooperative agreement), contract, private donations | NSF<br>DOE |
| COVID-19 HPC Consortium | Compute time | PPP | Consortium/ resource providers | Donation of resources | DOE<br>OSTP |
| XSEDE | Compute, data, and visualization | Virtual organization | University | Award (cooperative agreement) | NSF |

| | |
|---|---|
| DOE | Department of Energy |
| FFRDC | Federally Funded Research and Development Center |
| GOGO | government-owned government-operated |
| HPC | High Performance Computing |
| LSST | Large Synoptic Survey Telescope |
| NIST | National Institute of Standards and Technology |
| NSF | National Science Foundation |
| OLCF | Oak Ridge Leadership Computing Facility |
| OSTP | Office of Science and Technology Policy |
| PPP | public-private partnership |
| XSEDE | Extreme Science and Engineering Discovery Environment |

## Privacy Loss with Disclosure Penalties

Significant penalties for those who make unauthorized disclosures of confidential data, combined with privacy-preserving technologies and data access rules designed to protect against disclosure, could significantly reduce the likelihood that anyone would even attempt an unauthorized disclosure. Moreover, penalties could allow more and better information to be safely released without increasing the overall risks to privacy.

Privacy-protection measures such as traditional statistical disclosure controls and differential privacy focus on altering data to protect against disclosure. But other complementary policies could also reduce disclosure risk. Penalties on disclosure of confidential information would raise the cost of attempting to infer confidential information. Penalties are most effective when the probability of detection is high. For example, tax compliance studies have repeatedly found that compliance is very high for types of income when the Internal Revenue Service has substantial information about the source of income but quite low when there is no such information.

Assuming the evasion literature is applicable to potential data hackers, imposing legal sanctions including large financial penalties and possible jail time on anyone who uses confidential data for purposes other than the specific purposes identified in a data-sharing agreement would substantially reduce the likelihood of disclosure. One possible model is Internal Revenue Code section 6103, which defines unauthorized disclosure of confidential tax information by government employees and contractors as a felony subject to fines, court costs, and up to 5 years in jail. Those penalties could be extended to anyone who discloses confidential information, including organizations that receive stolen confidential data and disclose those data, and organizations that use computational or statistical techniques to try to re-identify individuals and firms in an anonymized data set. These sanctions could thus serve as a backstop to other disclosure control methods.

With substantial penalties and effective enforcement and prosecution, the legal sanctions could reduce the probability of any disclosure while preserving the integrity of the data for research purposes. The economic decision to disclose confidential data, like the decision to evade taxes, amounts to a calculation that the benefits of disclosure exceed the costs. For data that are anonymized and subject to other statistical disclosure controls, the cost could include the cost of acquiring other data that could allow matching and the computational resources applied to try to undo privacy protections. Applying more stringent privacy protections raises the cost of reverse engineering the confidential data but raising the legal penalties and the probability of detection and prosecution would have the same effect. To be clear, penalties and enforcement would be a complement rather than a replacement for other privacy protections. As noted in a recent Urban Institute paper, there are cases where legal sanctions alone would have limited effect. An optimal mix of policy instruments would include a combination of privacy protection measures, disclosure penalties and enforcement, and cybersecurity protections against hackers that minimize public and private costs.

Finally, there should be an exemption from legal penalties for privacy researchers who conduct research into re-identification techniques in a responsible manner (i.e., without publicly disclosing any personally identifiable information or doing anything else that would cause harm), so that data stewards can learn about vulnerabilities in their data. The Library of Congress defined a "good-faith security research" exemption to the Digital Millennium Copyright Act.

## Risk-Based Approach to Management

Protection of CIPSEA data has been developed according to the principle of disclosure risk, which considers both the probability of an unauthorized disclosure and the expected harm from such a disclosure. The financial cost in managing risk must be proportional to the utility of the information sought balanced with the amount of risk incurred, and a consideration of whether that cost is reasonable.

When applying controls and safeguards to protect data, the financial cost of disclosure protections needs to be managed against realistic assessments of risk. Over-protecting data could mean that less data are being made available or the data are of lesser quality or that another data set is being under-protected in an era of constrained resources. Using available resources appropriately ensures that data are protected, while the maximum amount of data are available and of the highest quality possible within the means of the provider.

Standardizing and making the risk assessment process consistent for all CIPSEA data, including data acquired under the Presumption of Accessibility can assist in finding efficiencies to produce more data and expand access. When releasing data, it is impossible to eliminate all possible risk. Risk is minimized as much as possible given many contextual and often temporal factors when data are released in non-identifiable form. OMB Circular A-130 emphasizes the need to manage privacy and confidentiality risk, which is further explored in NIST Interagency Report NISTIR 8062 to "develop an engineering approach to privacy."

## Theory of Change

The World Bank report, "Impact Evaluation in Practice," outlines a standard approach for describing how an intervention is supposed to deliver the desired results. This approach breaks the intervention into the following five discrete categories:

- ***Inputs.*** Resources at the disposal of the project, including staff and budget.
- ***Activities.*** Actions taken or work performed to convert inputs into outputs.
- ***Outputs.*** The tangible goods and services that the project activities produce; these are directly under the control of the implementing agency.
- ***Outcomes.*** Results likely to be achieved once the beneficiary population uses the project outputs; these are usually achieved in the short to medium term and are usually not directly under the control of the implementing agency.
- ***Final outcomes.*** The final results achieved indicating whether project goals were met. Typically, final outcomes can be influenced by multiple factors and are achieved over a longer period.

## United States Department of Defense Security Impact Levels

The United States Department of Defense (DOD) has classified national security information at three different levels (confidential, secret, and top secret) depending on the severity of damage to national security. In the early 2010s, the DOD created Security Impact Levels (IL) for public, Controlled Unclassified Information (CUI), and secret data to expand the use of commercial cloud computing capabilities, vendor-managed data centers, and other internal capabilities or services. These ILs were later aligned to the DOD Risk Management Framework (RMF) to operationalize the evaluation of data for future use.

The DOD RMF aligns to the NIST RMF used by the rest of the federal government. It was created after the 2013 National Defense Authorization Act required the DOD and the Intelligence Community (IC) to adopt a risk-based approach to modernize security approaches. The NIST RMF was adapted to the DOD RMF to account for the particularities of the IC and the DOD with respect to National Security Systems. The data risk was still based on confidentially, integrity, and availability, and the impact was based on the harm to national security if the data were exposed or compromised by an adversary. IT systems were authorized and controlled based on the IL of the data. This naturally created different tiers of access levels with a variety of access and distribution modes.

## Zero Trust

In early 2022, the Office of Management and Budget released memorandum M-22-09 that sets forth a federal zero trust architecture strategy, requiring that agencies meet specific cybersecurity standards and objectives by the end of fiscal year 2024. The strategy aims to reinforce the government's defenses against increasingly sophisticated and persistent threat campaigns, which target federal technology infrastructure, threatening public safety and privacy, damaging the American economy, and weakening trust in government.

# References

- "Building and Using Evidence to Improve Government Effectiveness," *Fiscal Year 2023 President's Budget, Analytical Perspectives, Chapter 6*, https://www.whitehouse.gov/wp-content/uploads/2022/03/ap_6_evidence_fy2023.pdf.
- "Leveraging Federal Statistics to Strengthen Evidence-Based Decision-Making," *Fiscal Year 2023 President's Budget, Analytical Perspectives of the Budget, Chapter 15*, https://www.whitehouse.gov/wp-content/uploads/2022/03/ap_15_statistics_fy2023.pdf.
- "Presidential Memorandum on United States Government on United States Government —Research and Development National Security Policy (NSPM-33)," January 2021, https://trumpwhitehouse.archives.gov/presidential-actions/presidential-memorandum-united-states-government-supported-research-development-national-security-policy/.
- Access 4 Learning Community, "Student Data Privacy Consortium," accessed October 6, 2022, https://privacy.a4l.org/.
- Actuate Innovation, "DataSafes," accessed October 6, 2022, https://actuateinnovation.org/programs/datasafes.
- Advanced Education Research and Development Fund, accessed October 6, 2022, https://aerdf.org/.
- Advisory Committee on Data for Evidence Building, accessed October 6, 2022, www.bea.gov/evidence.
- Advisory Committee on Data for Evidence Building, "Advisory Committee on Data for Evidence Building: Year 1 Report," *Bureau of Economic Analysis*, October 29, 2021. https://www.bea.gov/system/files/2021-10/acdeb-year-1-report.pdf.
- America's DataHub Consortium, accessed October 6, 2022, https://americasdatahub.org.
- American Rescue Plan, Pub. L. No. 117-2, 2021, https://www.congress.gov/bill/117th-congress/house-bill/1319/text.
- Australian Bureau of Statistics, "Five Safes Framework: Data Confidentiality Guide," August 8, 2021, https://www.abs.gov.au/about/data-services/data-confidentiality-guide/five-safes-framework.

- Autor, David et al., “The $800 Billion Paycheck Protection Program: Where Did the Money Go and Why Did It Go There?”, *National Bureau of Economic Research Working Paper Series*, Working Paper 29669, Cambridge, MA: NBER, 2022, https://www.nber.org/system/files/working_papers/w29669/w29669.pdf.
- Brooks, Sean et al., “NISTIR 8062: An Introduction to Privacy Engineering and Risk Management in Federal Systems,” *National Institute of Standards and Technology, Information Technology Laboratory*, January 2017, https://nvlpubs.nist.gov/nistpubs/ir/2017/NIST.IR.8062.pdf.
- Buhs, Caleb, “Michigan Job Growth Solid, Jobless Rate Edges Down in October,” November 17, 2021, https://www.michigan.gov/dtmb/about/newsroom/all-news/2021/11/17/michigan-job-growth-solid-jobless-rate-edges-down-in-october#:~:text=%2D%2D%20Data%20released%20for%20October,of%20Technology%2C%20Management%20%26%20Budget.
- Bureau of Labor Statistics, “Local Area Unemployment Statistics: Important Information,” June 29, 2022, https://www.bls.gov/lau/launews1.htm#outlier-improvement.
- Bureau of Labor Statistics, “Quarterly Census of Employment and Wages,” September 7, 2022, https://www.bls.gov/cew/.
- Burman, Leonard E., “Penalties for Unauthorized Disclosure and Data Privacy,” 2022, Washington, D.C.: Urban Institute, https://www.taxpolicycenter.org/publications/penalties-unauthorized-disclosure-and-data-privacy.
- Card, David, “Origins of the Unemployment Rate: The Lasting Legacy of Measurement Without Theory,” *The American Economic Review*, Vol. 101, No. 3, Papers and Proceedings of the One Hundred Twenty Third Annual Meeting of the American Economic Association (MAY 2011), pp. 552-557, https://www.jstor.org/stable/29783805#metadata_info_tab_contents.
- Census Bureau, “Federal Statistical Research Data Centers,” August 9, 2022, https://www.census.gov/about/adrm/fsrdc.html.
- Census Bureau, “Longitudinal Employer-Household Dynamics,” accessed October 6, 2022, https://lehd.ces.census.gov/.
- Centers for Disease Control and Prevention, “CDC WONDER,” August 17, 2022, https://wonder.cdc.gov/.
- Chief Data Officer Council, accessed October 6, 2022, https://www.cdo.gov/.
- Coleridge Initiative, accessed October 6, 2022, https://coleridgeinitiative.org/.
- Coleridge Initiative, “Administrative Data Research Facility,” accessed October 6, 2022, https://coleridgeinitiative.org/adrf/.
- Coleridge Initiative, “Applied Data Analytics Training,” accessed October 6, 2022, https://coleridgeinitiative.org/training/.

- Coleridge Initiative, “Midwest Collaborative: Data for Evidence-Based Policy (Spring Convening),” accessed October 6, 2022, https://coleridgeinitiative.org/workshops/workshop-mar2020/.
- Coleridge Initiative, “Multi-State Postsecondary Dashboard, Coleridge Initiative,” accessed October 6, 2022, https://coleridgeinitiative.org/projects-and-research/multi-state-post-secondary-dashboard.
- Coleridge Initiative, “Unemployment to Reemployment Portal,” accessed October 6, 2022, https://coleridgeinitiative.org/projects-and-research/unemployment-to-reemployment-portal/.
- Commission on Evidence-Based Policymaking, “The Promise of Evidence-Based Policymaking: Report of the Commission on Evidence-Based Policymaking,” Bipartisan Policy Center, September 2017, https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2019/03/Full-Report-The-Promise-of-Evidence-Based-Policymaking-Report-of-the-Comission-on-Evidence-based-Policymaking.pdf.
- Coronavirus Aid, Relief, and Economic Security Act, Pub. L. No. 116-136, 2020, https://www.congress.gov/bill/116th-congress/house-bill/748/text.
- COVID-19 Research Database, accessed October 6, 2022, https://covid19researchdatabase.org.
- Council on Government Relations, “Summary of NSTC Guidance for Implementing National Security Presidential Memorandum 33 Disclosure Requirements,” January 11, 2022, https://www.cogr.edu/sites/default/files/V%202%20Jan%2011%202022%20Summary%20of%20NSTC%20Guidance%20for%20Implementing%20National%20Security%20Presidential%20formatted.pdf.
- Cunningham, Jessica, Anna Hui, Julia Lane, and George Putnam, “A Value-Driven Approach to Building Data Infrastructures: The Example of the Midwest Collaborative,” *Harvard Data Science Review*, January 27, 2022, https://hdsr.mitpress.mit.edu/pub/mfhpwpxq/release/2.
- Data.gov, accessed October 6, 2022, https://data.gov.
- Data Quality Campaign, accessed October 6, 2022, https://dataqualitycampaign.org/.
- Department of Defense, “Cloud Computing Security Requirements Guide,” Version 1, Release 3, March 6, 2017, https://rmf.org/wp-content/uploads/2018/05/Cloud_Computing_SRG_v1r3.pdf.
- Department of Defense, “Risk Management Framework for DoD Systems,” *DOD Instruction 8510.01*, July 19, 2022, https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/851001p.pdf?ver=2019-02-26-101520-300.
- Higher Education Act, Ability to Benefit, 34 USC §484(d), accessed October 6, 2022, https://www.ecfr.gov/current/title-34.

- Department of Labor, “Department of Labor Equity Action Plan,” accessed October 6, 2022, https://www.dol.gov/general/equity-action-plan/plan.
- Department of Labor, Employment and Training Administration, “Wage Interchange Systems,” accessed October 6, 2022, https://www.dol.gov/agencies/eta/performance/swis.
- Department of Labor, Employment and Training Administration, “Workforce Performance Results,” accessed October 6, 2022, https://www.dol.gov/agencies/eta/performance/results.
- Desai, Tanvi, Felix Ritchie, and Richard Welpton, “Five Safes: Designing Data Access for Research,” *University of the West of England*, Economics Working Paper Series 1601, accessed October 6, 2022, https://www2.uwe.ac.uk/faculties/BBS/Documents/1601.pdf.
- Eurostat, “European Business Statistics Manual, *Eurostat Statistics Explained*, accessed October 6, 2022, https://ec.europa.eu/eurostat/statistics-explained/index.php?title=European_business_statistics_manual.
- Evaluation Officer Council, September 1, 2021, https://www.evaluation.gov.
- Executive Order No. 12345, “National Security Information,” April 2, 1982, https://www.archives.gov/federal-register/codification/executive-order/12356.html.
- Family Educational Rights and Privacy Act, 20 U.S.C. § 1232g, 34 CFR Part 99, https://www.govinfo.gov/content/pkg/CFR-2017-title34-vol1/pdf/CFR-2017-title34-vol1-sec99-31.pdf.
- Federal Committee on Statistical Methodology, “Data Protection Toolkit Beta,” September 22, 2020, https://nces.ed.gov/fcsm/dpt.
- Federal Information Security Management Act, Pub. L. No. 113-283, 2014, https://www.congress.gov/113/plaws/publ283/PLAW-113publ283.pdf.
- Federal Privacy Council, accessed October 6, 2022, https://www.fpc.gov/.
- FedRAMP, “FedRAMP Marketplace,” accessed October 6, 2022, https://marketplace.fedramp.gov/#!/products.
- Foundations for Evidence-Based Policymaking Act of 2018, Pub. L. No. 115–435, 2019, https://www.congress.gov/bill/115th-congress/house-bill/4174/text.
- Gertler, Paul J. et al., “Impact Evaluation in Practice, Second Edition,” 2013, Washington, D.C.: Inter-American Development Bank and World Bank, https://openknowledge.worldbank.org/handle/10986/25030.
- Government Accountability Office, “COVID-19: Urgent Actions Needed to Better Ensure an Effective Federal Response” (GAO-21-191), November 30, 2020, https://files.gao.gov/reports/GAO-21-191/index.html.

- Groshen, Erica, "Pandemic, Racial Inequities Underscore Need for Better Labor Market Data," *ILR School*, May 5, 2021, https://www.ilr.cornell.edu/work-and-coronavirus/public-policy/pandemic-racial-inequities-underscore-need-better-labor-market-data.
- Grumbling, Emily, Lisa Van Pay, and Morgan Livingston, "Options for Governance, Administration, and Ownership of a National AI Research Resource," STPI, Editor, 2022.
- Hart, Nick and Nancy Potok, "Modernizing U.S. Data Infrastructure: Design Considerations for Implementing a National Secure Data Service to Improve Statistics and Evidence Building," *Data Foundation*, July 2020, https://www.datafoundation.org/modernizing-us-data-infrastructure-2020.
- Health Insurance Portability and Accountability Act, Pub. L. No. 104-191, 1996, https://www.congress.gov/bill/104th-congress/house-bill/3103/text.
- HL7 International, "FHIR (HL7 Fast Healthcare Interoperability Resources," accessed October 6, 2022, http://www.hl7.org/implement/standards/product_brief.cfm?product_id=491.
- Infrastructure and Investment Jobs Act, Pub. L. No. 117-58, 2021, https://www.congress.gov/bill/117th-congress/house-bill/3684.
- Inter-university Consortium of Political and Science Research, accessed October 6, 2022, https://www.icpsr.umich.edu/web/pages/.
- Levine, Suzan G., "Grant Opportunity for Promoting Equitable Access to Unemployment Compensation (UC) Programs," *Employment and Training Administration Advisory System*, August 17, 2021, https://wdr.doleta.gov/directives/attach/UIPL/UIPL_23-21_acc.pdf.
- Lewis, Ryan C., Lauren E. Johns, and John D. Meeker, "Serum Biomarkers of Exposure to Perfluoroalkyl Substances in Relation to Serum Testosterone and Measures of Thyroid Function among Adults and Adolescents from NHANES 2011–2012," *International Journal of Environmental and Public Health*, May 29, 2015, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4483690/.
- National Artificial Intelligence Research Resource Task Force, accessed October 6, 2022, https://www.ai.gov/nairrtf/.
- National Artificial Intelligence Research Resource Task Force, "National AI Research Resource Task Force: Meeting #8," July 25, 2022, https://www.ai.gov/wp-content/uploads/2022/07/NAIRR-TF-Presentations-07252022.pdf.
- National Association of State Workforce Agencies, "Evidence-Building Capacity in State Workforce Agencies, February 1, 2017," https://www.naswa.org/reports/evidence-building-capacity-in-state-workforce-agencies.

- National Association of State Workforce Agencies, "Evidence-Building Capacity in State Workforce Agencies – COVID-19 Pulse Survey ," March 3, 2021, https://www.naswa.org/reports/evidence-building-capacity-in-state-workforce-agencies-covid-19-pulse-survey.
- National Center for Education Statistics, "Statewide Longitudinal Data Systems Grant Program," accessed October 6, 2022, https://nces.ed.gov/programs/slds/.
- National Center for Health Statistics, "Modernizing the National Vital Statistics System," October 1, 2020, https://www.cdc.gov/nchs/nvss/modernization.htm.
- National Center for Health Statistics, "nightingaleproject/Reference-Client-API," August 22, 2022, https://github.com/nightingaleproject/Reference-Client-API.
- National Center for Health Statistics, "nightingaleproject/Reference-NCHS-API," September 30, 2022, https://github.com/nightingaleproject/Reference-NCHS-API.
- National Center for Health Statistics, "Vital Statistics Modernization Community of Practice," October 20, 2021, https://www.cdc.gov/nchs/nvss/modernization/cop.htm.
- National Defense Authorization Act for Fiscal Year 2013, Pub. L. No. 112-239, 2013, https://www.congress.gov/112/plaws/publ239/PLAW-112publ239.pdf.
- National Institutes of Health Library, "Data Services," accessed October 6, 2022, https://www.nihlibrary.nih.gov/services/data.
- National Institutes of Health, National Center for Advancing Translational Science, "National COVID Cohort Collaborative," September 26, 2022, https://ncats.nih.gov/n3c.
- National Institute of Standards and Technology, Computer Science Resource Center, "NIST Risk Management Framework, October 5, 2022, https://csrc.nist.gov/projects/risk-management/about-rmf.
- Office of Management and Budget, Circular No. A-130, https://www.whitehouse.gov/wp-content/uploads/legacy_drupal_files/omb/circulars/A130/a130revised.pdf.
- Office of Management and Budget, "Moving the U.S. Government Toward Zero Trust Cybersecurity Principles," (M-22-09), January 26, 2022, https://www.whitehouse.gov/wp-content/uploads/2022/01/M-22-09.pdf.
- Office of Management and Budget, Office of Information and Regulatory Affairs, "The Interagency Council on Statistical Policy's Recommendation for a Standard Application Process for Requesting Access to Certain Confidential Data," January 2022, https://www.federalregister.gov/documents/2022/01/14/2022-00620/the-interagency-council-on-statistical-policys-recommendation-for-a-standard-application-process-sap.

- Office of Management and Budget, Office of Science and Technology Policy, National Science and Technology Council, "Guidance for Implementing National Security Memorandum 33 (NSPM-33) on National Strategy for United States Government-Supported Research and Development," January 2022, https://www.whitehouse.gov/wp-content/uploads/2022/01/010422-NSPM-33-Implementation-Guidance.pdf.
- Ohio State University, Center for Human Resource Research, "Ohio Longitudinal Data Archive," accessed October 6, 2022, https://chrr.osu.edu/ohio-longitudinal-data-archive.
- Opportunity Atlas, accessed October 6, 2022, https://www.opportunityatlas.org/.
- Potok, Nancy and Nick Hart, "A Blueprint for Implementing the National Secure Data Service: Initial Governance and Administrative Priorities for the National Science Foundation," *Data Foundation*, June 2022, https://www.datafoundation.org/cover-page-a-blueprint-for-implementing-the-national-secure-data-service.
- ResearchDatagov, accessed October 6, 2022, https://www.researchdatagov.org/.
- Results for America, accessed October 6, 2022, results4america.org.
- Stats NZ, "Integrated Data Infrastructure," August 23, 2022, https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure#data-safe.
- Stokes, Peter, "The 'Five Safes'–Data Privacy at ONS," *U.K. Office for National Statistics* (blog), January 27, 2017, https://blog.ons.gov.uk/2017/01/27/the-five-safes-data-privacy-at-ons/.
- https://uscode.house.gov/view.xhtml?path=/prelim@title26&edition=prelim.
- UK Data Service, "What is the Five Safes Framework," *SecureLab*, accessed October 6, 2022, https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework.
- United States Copyright Office, "Section 1201 Rulemaking: Eighth Triennial Proceeding to Determine Exemptions to the Prohibition on Circumvention," October 2021, https://cdn.loc.gov/copyright/1201/2021/2021_Section_1201_Registers_Recommendation.pdf.
- U.S. Chamber of Commerce Foundation, "Jobs and Employment Data Exchange (JEDx)," accessed October 6, 2022, https://www.uschamberfoundation.org/JEDx.
- Waterfield et al., "Reducing Exposure to High Levels of Perfluorinated Compounds in Drinking Water Improves Reproductive Outcomes: Evidence from an Intervention in Minnesota," *Environmental Health* 19, Article number: 42, April 2020, https://ehjournal.biomedcentral.com/articles/10.1186/s12940-020-00591-0.
- Workforce Innovation and Opportunity Act, Pub. L. No. 113–128, 2014, https://www.congress.gov/bill/113th-congress/house-bill/803/text.

# Acronyms

| Acronym | Definition |
|---|---|
| ACDEB | Advisory Committee on Data for Evidence Building |
| ADC | America's DataHub Consortium |
| ADRF | Administrative Data Research Facility |
| AERDF | Advanced Education Research and Development Fund |
| API | Application Programming Interface |
| ARP | American Rescue Plan |
| ATB | Ability to Benefit |
| ATI | Advanced Technology International |
| AWS | Amazon Web Services |
| BLS | Bureau of Labor Statistics |
| CARES | Coronavirus Aid, Relief, and. Economic Security |
| CD2H | National Center for Data to Health |
| CDC | Centers for Disease Control and Prevention |
| CDO | Chief Data Officer |
| CIPSEA | Confidential Information Protection and Statistical Efficiency Act |
| CoP | community of practice |
| CTE | career and technical education |
| CUI | Controlled Unclassified Information |
| DMV | Department of Motor Vehicles |
| DOC | Department of Commerce |
| DOD | Department of Defense |
| DOL | Department of Labor |
| DQC | Data Quality Campaign |
| ED | Department of Education |
| EPA | Environmental Protection Agency |
| ERS | Economic Research Service |
| ETA | Employment and Training Administration |
| FACA | Federal Advisory Committee Act |
| FAQ | frequently asked question |

| Acronym | Definition |
|---|---|
| FedRAMP | Federal Risk and Authorization Management Program |
| FERPA | Family Educational Rights and Privacy Act |
| FFRDC | Federally Funded Research and Development Center |
| FHIR | Fast Healthcare Interoperability Resources |
| FISMA | Federal Information Security Management Act |
| FRN | Federal Register Notice |
| FSA | Federal Student Aid |
| FSRDC | Federal Statistical Research Data Center |
| FY | fiscal year |
| GAO | Government Accountability Office |
| HR | human resources |
| ICPSR | Inter-university Consortium for Political and Social Research |
| ICSP | Interagency Council on Statistical Policy |
| HIPAA | Health Insurance Portability and Accountability Act |
| HHS | Department of Health and Human Services |
| IC | Intelligence Community |
| IIJA | Infrastructure Investment and Jobs Act |
| IL | Impact Levels |
| IRS | Internal Revenue Service |
| JEDx | Jobs and Employment Data Exchange |
| LEHD | Longitudinal Employer-Household Dynamics |
| LMI | Labor Market Indicator |
| MOU | Memorandum of Understanding |
| MWC | Midwest Collaborative |
| N3C | National COVID Cohort Collaborative |
| NAIRR | National Artificial Intelligence Research Resource |

| Acronym | Definition |
|---|---|
| NASWA | National Association of State Workforce Agencies |
| NCATS | National Center for Advancing Translational Sciences |
| NCHS | National Center for Health Statistics |
| NCSES | National Center for Science and Engineering Statistics |
| NIH | National Institutes of Health |
| NIST | National Institute of Standards and Technology |
| NLM | National Library of Medicine |
| NSDS | National Secure Data Service |
| NSF | National Science Foundation |
| NVSS | National Vital Statistics System |
| OLDA | Ohio Longitudinal Data Archive |
| OMB | Office of Management and Budget |
| PACIA | Performance Accountability and Customer Information Agency |
| PFAS | per- and polyfluoroalkyl substances |
| PII | personally identifiable information |
| PMO | Project Management Office |
| PPRL | privacy-preserving record linkage |
| PPT | privacy-preserving technology |
| PROMIS | Program for Measuring Insured Unemployed Statistics |

| Acronym | Definition |
|---|---|
| QCEW | Quarterly Census of Employment and Wages |
| RDC | Research Data Center |
| R&D | research and development |
| RFA | Results for America |
| RMF | Risk Management Framework |
| SAP | Standard Application Process |
| SDL | Statistical Disclosure Limitation |
| SMC | secure multiparty computation |
| SOI | Statistics of Income |
| SLDS | Statewide Longitudinal Data System |
| SWIS | State Wage Interchange System |
| TBD | to be determined |
| UI | unemployment insurance |
| UN | United Nations |
| USDA | United States Department of Agriculture |
| VA | Veterans Administration |
| WIAC | Workforce Information Advisory Council |
| WIOA | Workforce Innovation and Opportunity Act |
| WONDER | Wide-ranging Online Data for Epidemiologic Research |

ACDEB
Advisory Committee on Data
for Evidence Building

# Advisory Committee on Data for Evidence Building:

## Year 2 Report Supplemental Information

[Visit the ACDEB website]