July 10, 2024

THE AMERICAN ECONOMIC ASSOCIATION

Committee on Economic Statistics: www.aeaweb.org/about-aea/committees/economic-statistics
Committee on Government Relations: https://www.aeaweb.org/about-aea/committees/government-relations

The Evolving Federal Data Ecosystem: New Opportunities for Economic Research?

Summary of an AEAStat/CGR Working Session held January 5, 2024, at the
2024 Allied Social Science Association meetings in San Antonio, TX

Background: The federal data ecosystem is currently undergoing a period of rapid change brought on by technological advances and institutional changes. One key development has been the enactment of the Foundations of Evidence-Based Policymaking Act in January 2019. The "Evidence Act" calls for greatly increased use of federal government survey and administrative data to build evidence on the effectiveness of government programs and policies. This in turn has required major innovations in ways that confidential government data assets are accessed and used. Changes already in place include the launch of a one-stop web portal for exploring and applying to use restricted micro data from federal statistical agencies; the establishment of a demonstration project for a National Secure Data Service (NSDS); and the creation of the Center of Excellence in the Census Bureau to help evaluate key government programs authorized by legislation such as the Inflation Reduction Act and CHIPS for America Act. These and other advances are providing new opportunities to link government data sets with each other, as well as to proprietary, state and local government, and other types of data. Many other changes are in the works.

Objective: The AEA's Committee on Economic Statistics (AEAStat) and Committee on Government Relations (CGR) organized a working session at the 2024 ASSA meetings on the evolving federal data ecosystem and its implications for economic research. The session brought together 40 economists and other experts from academia, statistical agencies, other federal agencies, and research organizations for a facilitated discussion of four central questions, listed below. This document summarizes some of the key issues that arose in discussion, organized by groups of questions that framed the meeting. *It should not be inferred that all participants agreed on points that were raised*.

A word of caution: Information contained in this document is current as of July 10, 2024, but, because the federal data ecosystem is in a period of rapid evolution, specific facts given in this document could be different going forward.

Overview of the Evolving Data Landscape

Traditionally economic researchers have gained access to confidential federal statistical data sources by submitting an application to the agency that collected and/or housed them. Approved projects have been carried out in secure computing environments like the Federal Statistical Research Data Centers

AEA Contact: Martha Starr, PhD, Washington, DC, Representative of the AEA and Director of Government Relations.
martha.starr@aeapubs.org. 703-283-0054.

(FSRDCs) and the Bureau of Labor Statistics' Virtual Data Enclave (BLS VDE), and analyses, findings, and results have had to undergo disclosure review by the agency before they could be shared.

The new ecosystem is largely building on existing ways of accessing and using confidential data, with the intentions of making it easier for people to find data assets available for research and apply to use them, and expanding their use beyond small communities of established users at well-resourced universities. Five notable changes already underway are as follows:

ResearchDataGov.gov and the Standard Application Process (SAP)

- A major recent innovation is ResearchDataGov.org, a "one-stop" web portal for locating and requesting access to restricted microdata from federal statistical agencies.[1] As of July 2024, 1,502 data sets are indexed on the site; many more data sets are expected to be added going forward. For each data collection indexed in the portal, "metadata" are included that describe the data's scope, coverage, and methodology, and information is given as to how to apply to access it.
- Applications to use confidential data sets are now submitted via a Standard Application Process (SAP) that is shared across agencies. While individual agencies remain responsible for approving projects that access their data, now only one application needs to be submitted for projects intending to link data across multiple agencies, which is intended to facilitate research taking advantage of complementarities across data sources.
- To reduce uncertainty and increase transparency, agencies have committed to timelines for reviewing projects, including making final accept/reject decisions within 12 weeks or 24 weeks for projects using data from multiple agencies.[2] Applications can now be tracked online, and rejections can be appealed.

National Secure Data Service (NSDS)

- The Evidence Act foresaw the creation of a new federal shared service called the National Secure Data Service (NSDS). Currently housed in the National Science Foundation's National Center for Science and Engineering Statistics (NCSES), a 5-year NSDS demonstration project is currently piloting development of a variety of methods and services needed to support more intensive research use of confidential federal data assets for evidence building purposes, including new data-linkage methods, privacy-preserving technologies, and shared-services platforms needed to manage secure multisite research access to confidential data.
- The NSDS aims to solve challenges that stand in the way of unlocking the value of confidential data assets for evidence-building purposes. As such, it can be viewed as a philosophy (producing value for the American public by facilitating evidence building), a service (providing coordination and capacity-building services for data users, data providers, and related communities of practice), and a

[1] As of July 2024, ResearchDataGov.gov indexed data sets from the Bureau of Economic Analysis (13 data sets), Bureau of Justice Statistics (734), Bureau of Labor Statistics (21), Bureau of Transportation Statistics (6), Census Bureau (36), Federal Reserve Board Microeconomic Surveys Unit (2), IRS Statistics of Income Division (4), National Center for Education Statistics (342), National Center for Health Statistics (45), National Center for Science and Engineering Statistics (6), SAMHSA Center for Behavioral Health Statistics and Quality (1), USDA Economic Research Service (5), USDA National Agricultural Statistics Service (18).

[2] There can also be "revise and resubmit decisions," where researchers get feedback with respect to items in the proposal which could result in its approval if they can be revised. Agency-specific data on application outcomes can be accessed on "Agency Info and Metrics" pages at ResearchDataGov.gov.

place (a legally recognized entity that functions within the larger federal ecosystem, with hardware, software, and administrative infrastructure and capacity that allow it to meet its mission). The NSDS may do the problem-solving needed to make a project possible, with implementation then carried out through an FSRDC or agency-specific secure data access facility (e.g., the BLS VDE).

- One part of the NSDS demonstration project, conducted by the Inter-university Consortium for Social and Political Research (ICPSR) at the University of Michigan, is examining how access to federal confidential data assets via the FSRDCs can be equitably expanded beyond the traditional base of "R1" research universities. Multiple other efforts are underway as part of the NSDS demonstration project and are listed on the NSDS Demonstration Project information webpage.

Census Bureau's FSRDCs and Center of Excellence

- The Census Bureau has had longstanding, extensive involvement in providing research access to confidential federal data assets via the FSRDCs. From the first use of Special Sworn Status (SSS) at Census Bureau headquarters in Suitland, MD, in 1982 to the network of 33 centers in place today, the FSRDC system has housed research projects on everything from job creation and destruction by U.S. businesses to intergenerational inequality. As of September 2023, there were 572 active research projects in the FSRDC system, of which about 375 used Census data.
- The Census Bureau has also established a new Center of Excellence that will work with other federal government agencies to help monitor and evaluate their programs, for example, by linking Census Bureau data to administrative records from agency programs. As a cutting-edge example of a project of this kind, done jointly by the Census Bureau's Center for Economic Studies, the Treasury Department, and the Internal Revenue Service, Census matched administrative records on recipients of Economic Impact Payments distributed in the COVID-19 pandemic with census, survey, and administrative data on individuals and households, in order to examine the extent to which there were disparities in the distribution of payments with respect to race/ethnicity and other household characteristics.

Modernizing Vital Statistics

- The National Center for Health Statistics is undertaking a large-scale, highly innovative project intended modernize the U.S. vital statistics system. Among other things, the project is working to:
    - Develop standards for interoperable electronic data exchange to facilitate standardized, automated collection of vital-records information by state agencies and other jurisdictions responsible for collecting it and transmitting it to NCHS.
    - Shift to natural language processing methods for coding cause-of-death-information.
    - Improve and expand automated transmission of birth information by hospitals to state vital-records agencies.
    - Facilitate linkage of records of fetal deaths with information as to causes.
    - Develop consistent standards for medical examiners to use in recording opioid-related deaths.
    - Improve the timeliness and accuracy of death reporting in natural disasters.
- The project is expected to substantially improve public-health authorities' access to granular, real-time data on vital events in their own jurisdictions and facilitate research on factors that advance or impede improvements in population health.

- As issues of collecting, standardizing, and processing data from multiple jurisdictions are central to the modernization project, this initiative is an excellent example of the type of challenges an NSDS is designed to address. To help inform a future NSDS, an effort within the NSDS demonstration project related to the National Vital Statistics System Modernization is exploring new opportunities for the use of interoperable health data to support timely research and public health surveillance.

State and Multistate Data

- The Coleridge Initiative has been integrally involved in developing "Multi-State Data Collaboratives" – coalitions of state workforce, education, human services, and other agencies that work together and with university partners to compile de-identified micro data stored in a common platform, with the aim of opening up new possibilities for conducting policy-oriented applied research. At present 25 states actively participate in at least one multistate collaborative. Collaboratives are currently set up in the east, south, and mid-west. Coleridge has many efforts underway to increase participation in existing collaboratives and encourage creation of new ones, so continued expansion going forward can be expected.
- An integral part of this work is providing Applied Data Analytics (ADA) training for government agency staff, which is key for realizing the full value of newly available data for real-time, policy-making purposes. As of 2023, more than 1,000 people from 317 organizations from 46 different states have benefited from this opportunity.

Questions discussed at the AEAStat/CGR working lunch:

1. *Input from economists: What are the most important questions about the implementation of the evolving data ecosystem that require input from academic researchers? What is the best way for that input to be solicited and delivered?*
    - People on the supply side of new data-access systems have imperfect information on what researchers most need or want from an improving access system. Consequently, they very much value input from academic researchers as to: projects they'd like to carry out but think may not be feasible, constraints and problems they face getting projects off the ground, questions about methods that could be used to match different data sets, and other issues relating to expanding and improving research use of confidential data.
    - A key message is: PLEASE ASK if there is something you would like to do, and it does not look like you can do so. Such input is valuable for architects and stewards as they work to build a new data ecosystem that addresses unmet needs of expanding communities of users.
    - Conversely, academics would generally be happy to answer questions and provide inputs to architects and stewards considering alternative options as to how to set up the ecosystem and what features to build into it.
    - NCSES has set up an "Idea Bank" via which all stakeholders are encouraged to submit ideas that could inform a future NSDS and help shape the future development of the ecosystem in ways that will maximize its potential for building evidence while keeping access for confidential data assets secure.

2. *Standards for approving research projects: Are projects proposed by academic economists expected to contribute to evidence-based policymaking (and if so, how)? How important is it for projects to be done collaboratively with researchers from data-housing agencies, or is that optional? If academic economists have specific skill sets that evidence-building projects need (e.g., expertise in causal design), how can connections be made? Will researchers participating in the new data ecosystem be expected to deposit programs or linkage code to shared repositories?*

   - Confusion and lack of information about procedures for getting access to confidential government data can hinder expansion of its use beyond existing communities of knowledgeable users. Session participants conveyed that they would value clear, high-level explanations of criteria used by different agencies to evaluate applications to use confidential data assets. It was also suggested that organizing sessions at large conferences and holding webinars could be good means of expanding the pool of people making use of new opportunities for economic research.
   - In fact, a key goal of setting up ResearchDataGov.gov (the SAP portal) has been to greatly increase the supply of readily available information about how access to confidential data can be obtained. The site now posts a Users' Guide that aggregates agency-specific information on who can access the agency's data (e.g., U.S. citizens, people affiliated with U.S. higher-education institutions, etc.), how access to the secure computing environment works (onsite vs. virtual), what software is available in that environment, agency-specific requirements for the application process, and who to contact with questions.
   - Researchers should be aware that, even if the Evidence Act aims to expand access to confidential federal data assets for research purposes, agencies continue to apply mandated approval criteria when evaluating research proposals.[3] Approval criteria that are common across agencies include:[4] (a) data can be used for statistical purposes[5] only; (b) use of the data must be consistent with what people from whom it was collected were told about how it would be used; (c) information that could identify individual people or businesses in the data cannot be publicly disclosed; (d) there must be a demonstrated need that cannot be met without access to confidential data; (e) objectives of the research project must be feasible in view of characteristics of the data; and (f) the project poses no risk of reducing public trust.
   - In addition, specific agencies may have additional criteria they are required to use when evaluating research projects. Of particular note for economic research projects:
     - Projects using confidential Census Bureau data must contribute to the agency's mission, and researchers must become sworn agents of the Census Bureau.[6]
     - Projects using Internal Revenue Service tax data must have a tax-administration purpose.[7]
   - It was noted that less experienced researchers often find it very valuable to consult with FSRDC administrators as to how they can develop research proposals that will satisfy agency-specific approval criteria – although it was also pointed out that FSRDC administrators are experiencing

[3] See OMB Memorandum, "Establishment of Standard Application Process Requirements on Recognized Statistical Agencies and Units," December 8, 2022, pp. 15-17.
[4] Ibid.
[5] "The term 'statistical purpose' means the description, estimation, or analysis of the characteristics of groups, without identifying individuals or organizations that comprise such groups." Ibid., p. 6.
[6] Ibid., p.17.
[7] Internal Revenue Service, "Statistics of Income – Joint Statistical Research Program," pp. 1-2.

increased workloads as interest in accessing and using confidential data rises, even as numbers of administrative staff in the network have increased.

- Session participants also noted that it would be great if new infrastructure created ways to connect people conducting similar research projects, so they can exchange information about the approval process, the mechanics of getting projects off the ground, obstacles to clearing disclosure review, and other issues.[8]

3. *Changes in data-access routes: For academic researchers, will new ways of accessing and linking confidential federal data replace or complement structures like FSRDCs and secure data enclaves? Will projects that link survey and administrative data have different disclosure-review standards than projects using survey data only? What is being done to ensure that researchers from a diverse range of institutions can take advantage of new research opportunities?*

    - The FSRDC system has been a phenomenal platform for realizing the value of confidential federal data assets for research purposes. However, because it has been a bespoke system involving careful review and backstopping of each individual research project, it has generally been well suited for projects that can afford to have multiyear time frames and uncertainties with respect to timing and outcomes. These types of features could hinder efforts to expand and diversify the pool of users of confidential data assets. For example, early-career researchers often cannot afford to undertake projects with long time frames and uncertain returns and so may turn to other data sources even if using confidential federal data would provide the most accurate and reliable answers to their research questions.
    - An additional issue with the current system concerns a Catch-22 in research funding. Acquiring research grants usually requires having proof of concept in hand – yet significant amounts of time and money are required to get preliminary results from projects using confidential data. This could be a problem for expanding use of the federal data ecosystem beyond researchers from R1 universities, as researchers elsewhere are less likely to be able to get institutional support for early-stage research projects. Session participants recommended asking research funders to consider creating new pools of seed money that could help support early-stage work of this type.
    - It is also a concern that expanding use of confidential data assets beyond the R1 universities generally requires extra external funding as non-R1 colleges and universities may not be in a position to pay fees to join the FSRDC system and then to support the infrastructure on their campus. Even current FSRDCs do not necessarily have secure and adequate funding. The FSRDC system would also need extra resources to support a higher volume of research projects. Again, arranging novel types of research funding that could support expanded use would be beneficial.
    - Additionally, it could be valuable to consider complementing expansion of the existing system with expansion of automated secure research access to confidential data. One option being explored by projects currently underway is to provide synthetic public-use data files to researchers for use in the development phase of their analyses, after which final-stage analyses could be submitted to secure validation or verification servers that return results cleared

[8] A new FSRDC research repository is expected to be a valuable asset in this respect.

through disclosure review in an automated way.[9] Automating access to data sets in relatively high demand could free up agency resources to handle more complex and novel projects that require project-specific support, and could address at least some part of the needs of new users in lower-cost, more timely ways.

4. *Unresolved questions and expected future directions: What future directions in the evolution of the new federal data ecosystem should research economists be aware of? Are there unanswered questions about its implementation that research economists can help address? How can a constructive ongoing dialogue about these issues be implemented?*

   - What can be done to make it easier for researchers to identify and use best practices for linking data sets? For example, Census and EPA data have been linked dozens of times, yet a specific method stands out as the best practice. Session participants expressed interest in seeing the ecosystem post standardized information on linkage options for specific data sets and programs that show how they have been implemented.
   - What do we do about reproducibility and replication packets? Increasingly journals like those published by the American Economic Association require researchers to deposit data, programs, and documentation in publicly available data archives. While confidential data sets clearly cannot be posted, researchers are still obligated to deposit as much material as they can without violating disclosure terms, and to make provisions that would allow researchers who do gain access to the same data to exactly replicate their work. How is this issue being handled by the statistical agencies and ResearchDataGov.gov? It would be highly valuable if ResearchDataGov.gov could develop consistent, systematic standards for developing and depositing replication packets for research projects using confidential data, including indexing their availability.
   - What about confidential data held by entities other than the federal government – will the evolving ecosystem eventually incorporate confidential data from state and metropolitan governments, businesses, and/or other types of entities willing to make their data available for research purposes?
     - For specific research projects in the FSRDCs and most other restricted enclaves, researchers can always propose bringing other types of confidential data into the secure computing environment for the duration of the project. However, in many cases data brought into the secure enclave are not available to subsequent researchers after the project concludes due to licensing restrictions. Both researchers and statistical agencies can make efforts to ensure that valuable linked data sets are made available to subsequent researchers, with appropriate credit given to those who did the work to link the data and make it available.
     - It was pointed out that state and local governments often welcome increased research use of their data assets for applied-policy purposes, but do not have resources and expertise to set up and manage secure virtual data enclaves. Session participants expressed enthusiasm for exploring arrangements that bring state and local government confidential data into the federal-data ecosystem and setting up ways to link them with federal data.
   - How can interested researchers become involved in efforts by federal, state, and local governments to make better use of newly available data assets to improve the evidence basis

[9] E. Groshen and D. Goroff, "Disclosure avoidance and the 2020 Census: What do researchers need to know?" *Harvard Data Science Review*, June 24, 2022. See also AEAStat, "Implications of New Privacy Protection Methods for Economic Research," June 2023.

for policy? And how can people from federal, state, and local governments find academics who may be interested in conducting research projects using newly compiled data or helping to build data-analysis systems that permit better real-time, granular tracking of policy-related outcomes?

- Session participants expressed interest in creating ways to connect state and local government entities with research needs and academics and other researchers willing and able to help fill them, which would be very valuable for realizing the full value of newly available data for evidence-based policy.
- Forging connections between researchers and tribal governments and community groups would be of especially high value. These groups have high unmet demand for data on their communities' characteristics and needs, yet may not have computing infrastructure and people trained in data analytics to be able to directly make use of new opportunities to fill that demand. Collaborative projects involving nearby institutions could be especially valuable if they also help build local capacity to make use of newly available data to improve community members' lives via evidence-based policies.

Acronym Decoder		
ACDEB	Advisory Committee on Data for Evidence-Based Policy	Committee that was set up under the Evidence Act charged with reviewing, analyzing, and making recommendations on how to promote the use of Federal data for evidence building.
ADC	America's DataHub Consortium	A collaborative entity established by NCSES in 2021 as a vehicle for supporting projects that expand evidence-building capacity, including work relating to the NSDS demonstration project
FSRDCs	Federal Statistical Research Data Centers	Research centers located at universities and Federal Reserve units that work in partnership with the Census Bureau in which approved researchers can carry out approved projects that use confidential federal data.
ICSP	Interagency Council on Statistical Policy	Council made up of statistical-agency heads and heads of statistical-units within other federal agencies that works with the Chief Statistician of the U.S. (CSOTUS) on policy issues affecting the U.S. statistical system.
ICPSR	Inter-university Consortium for Political and Social Research	Consortium of 825 universities, colleges, and other research institutions that works to systematically archive and provide access to research databases, and training in research methods. Housed at the University of Michigan.
NCSES	National Center for Science and Engineering Statistics	A statistical agency housed in the National Science Foundation, which is the institutional location of the NSDS-D.
NSDS-D	National Secure Data Service – Demonstration Project	5-year demonstration project set up by the CHIPS and Science Act of 2022 to develop a shared-services model for expanding and improving secure, efficient data-sharing and linkage of confidential federal data for evidence-building purposes.
OMB	Office of Management and Budget	Charged by statute with coordinating the U.S. statistical system. Houses the Office of the Chief Statistician
SAP	Standard Application Process	One-stop portal for finding confidential federal data housed by different statistical agencies and units and applying to access them for statistical purposes.