EVIDENCE TOOLKIT

Data Access and Integration

Robert Doar, Morgridge Fellow in Poverty Studies
American Enterprise Institute

January 2018

Evidence-Based
Policymaking
Collaborative

EVIDENCECOLLABORATIVE.ORG

What you need to know about data access and integration

- Government data consist of information collected through the administration of government programs and survey responses. They are often inaccessible across different agencies and generally underused.
- Expanding access to government data and facilitating integration of data from different sources can help improve service delivery and ensure government efficiency.
- Efforts to expand access and data integration can and must be pursued in a way that ensures the confidentiality of the data.
- The bipartisan Commission on Evidence-Based Policymaking has proposed the creation of a National Secure Data Service to facilitate data access and integration while protecting privacy.

Why is improving access to government data and encouraging data integration important?

When making decisions, policymakers and program managers consult a wide range of sources of information. At the federal, state, and local levels, public organizations collect data about many aspects of life in the communities they serve. Social services agencies collect data about family size and income, school districts track test scores and graduation rates, and law enforcement agencies gather information about crime and incarceration.

These administrative and survey data are used to inform decisions about individuals as well as major policy issues. However, one dilemma confronting policymakers is that even though social issues are multidimensional, these data sources (e.g., test scores from a school district or the employment status of parents) are often held closely by one agency and not used to inform decisions by other agencies or programs.

As data technology continues to improve, governments will have the chance to provide greater access to data and find new ways to integrate different data sources. As a result, governments will be able to find more comprehensive answers and new insights to pressing policy questions, offering opportunities to better monitor programs and improve service delivery, prevent waste and abuse, test innovations, and ensure continuous program improvement.

Key terms

GOVERNMENT AGENCY DATA

The information collected by government agencies originates from several sources:

- **Administrative data** are collected to meet the needs of a specific program or regulation and have several uses, including determining eligibility, ensuring compliance with program rules and requirements, evaluation, and regulatory or law enforcement purposes. Virtually every government program collects and verifies some amount of administrative data, ranging from basic facts about program participation to detailed demographic, economic, and program-use data meant to aid management and evaluation. For example, a public housing authority may track the name, age, address, and employment status of people receiving vouchers in its jurisdiction.
- **Survey data** are self-reported information usually collected from sampled respondents. The most well-known government survey in the United States is the decennial Census. The Census is complemented by the more frequent American Community Survey, which relies on smaller sample sizes but gathers new data annually. Governments at every level, as well as many government agencies, administer surveys to gather information about their constituents, economic conditions, and other topics. Recent studies have shown that survey respondents do not always provide accurate responses regarding their receipt of certain public assistance benefits. This is why survey data and administrative data need to be used together to analyze the condition of Americans.
- **Microdata** are information at the level of individual respondents. Whereas survey and census results are often published in aggregate form, microdata contain each data point for each respondent. This means microdata sets can potentially contain millions of records, allowing researchers more freedom to perform statistical analyses.

MECHANISMS FOR LEVERAGING EXISTING DATASETS TO IMPROVE PROGRAM DESIGN AND IMPLEMENTATION

Policymakers, researchers, funders, and other stakeholders can all benefit from making existing data more useful for answering questions relevant to policy. This can occur in several ways:

- **Data access** refers to enhancing the availability of datasets. Expanding access to data can include making them more widely available internally within the government and sharing information across public agencies. It can also mean making government data available to third parties for research and evaluation purposes.
- **Data integration** involves merging and layering these datasets through processes like data matching. Administrative datasets can be merged together or linked with survey data or private vendor data. Such integration can provide more reliable findings than individual data sources alone, and the newly created sets can potentially generate fresh insights about population groups and government programs and policies. For example, because of misreporting and other limitations inherent in survey data, administrative data can provide more accurate information about public

programs and policies and the people they aim to serve. Integration also takes advantage of existing data systems and provides an opportunity to get more value from what they already collect.

What are the benefits of increased data access and integration?

Public agency data integration can happen at all levels of government. Agencies can use, share, and combine data for several purposes, including policy and program evaluation and research, program eligibility determination, and service delivery improvement.

- **Service delivery improvement:** By integrating government datasets, agencies can improve program design and provide better customer service to citizens. For example, metrics developed for and gathered through one program can be used to inform the design of other programs or to identify needs in other program or policy areas. Moreover, citizens interact with government agencies in many ways, yet different programs often ask them to fill out similar forms or share similar information, slowing down service delivery. With integrated data, health care providers can quickly determine a patient's health insurance status by accessing state Medicaid files and significantly speed up the intake process.
- **Policy and program evaluation and research:** Data sharing can help government agencies undertake continuous improvement through policy evaluation and research. With more data and possibly new metrics at their disposal, agencies can more easily measure their own successes and shortcomings. These efforts can support further data-driven decisionmaking, inform better program design and implementation, and allow testing of new innovations.
- **Program eligibility determination:** Data sharing and integration can help government agencies determine who is eligible to participate in certain government programs. By doing so, agencies can better target services to their intended recipients. This can improve efficiency in government operations by streamlining the determination process and creating fewer barriers for beneficiaries. Programs, in turn, could experience either cost increases or reductions. For example, New York City was one of the first cities to share Medicaid data with its department of education to more efficiently determine eligibility for its school meals program. Today, federal policy encourages all states and local education agencies to match their public benefit records against student enrollment lists and directly certify eligible children for free lunch and breakfast.[1] Because improved data access and integration may help program administrators better identify those eligible for services, some may see rising costs because of increased enrollment. At the same time, data integration in other contexts may reveal program inefficiencies or fraud, which could result in potential cost savings once resolved. Data integration can also be a useful tool for policymakers hoping to encourage

[1] US Department of Agriculture, Food and Nutrition Service, "Direct Certification in the National School Lunch Program: State Implementation Progress, School Year 2013-2014: Report to Congress – Summary" (Washington, DC: US Department of Agriculture, 2015).

personal responsibility. For instance, child support enforcement agencies rely on many data sources to ensure the proper and timely payment of child support from noncustodial parents.

What are some success stories involving expanded data access and integration?

Integration of government agency data has been an important method for developing improved public policy and programs, based on evidence of the realities of American life.

EQUALITY OF OPPORTUNITY PROJECT

A team of researchers led by former Harvard economists Raj Chetty and Nathaniel Hendren released new research in 2015 on how neighborhoods affect upward economic mobility. Linking administrative and survey data from multiple government entities, including the Census, the IRS, and the Integrated Postsecondary Education Data System, the researchers conducted a national quasi-experimental study of 5 million families and re-analyzed the Moving to Opportunity Experiment. The research looks at how economic mobility and opportunity are shaped by the neighborhood you grow up in. The findings show that every year of exposure to a lower-poverty neighborhood improves a child's chances of success, especially among younger movers. The findings also brought attention to evidence that growing up in neighborhoods with a high percentage of single-parent families has a negative impact on upward mobility. This work helps illustrate how place and family structure matter for economic mobility and supports the importance of residential mobility programs and place-based strategies, as well as programs that promote marriage before having children. More broadly, it is an example of how data integration can be leveraged for policy evaluation and research that benefits the field and informs future policy decisions.

EFFORTS TO IMPROVE POVERTY MEASUREMENT

Although poverty experts have long acknowledged limitations in our official statistics on poverty, several scholars have used new information available through data integration to shine a light on one particular inadequacy: the underreporting of government benefits. As Bruce Meyer, a University of Chicago professor and AEI visiting scholar, documented in his 2015 study, official government surveys used to calculate the poverty rate "sharply understate the income of poor households" because survey respondents deliberately or accidentally do not correctly report how much they earn or receive in government transfers. However, this discrepancy can only be seen when survey responses are matched against program administrative data. Furthering data integration efforts will allow researchers to develop a more accurate understanding of the material well-being of low-income households. Senator Mike Lee (R-UT) has introduced the Poverty Measurement Improvement Act, which would authorize the Census Bureau to create a new survey that could be linked with agency data to provide a more accurate picture of poverty in the United States.

ACCESS NYC

Data integration would not only benefit the government agencies looking to make their programs more efficient. It would also help people receiving government benefits receive all they are eligible for by giving them a comprehensive view of available social services. The City of New York provides such a view by integrating datasets from local, state, and federal agencies to create ACCESS NYC, an online public screening tool New Yorkers can use to determine their eligibility for and apply to a range of local, state, and federal health and human services programs. The platform includes over 30 programs focused on topics such as early child care and education, employment and training, housing programs, health care services, food and nutrition, and financial assistance. ACCESS NYC is part of HHS-Connect, New York's comprehensive initiative launched to break down barriers between various public agencies and provide better information about client needs. Sharing data across agencies also helps city workers conduct faster child welfare investigations and determine Medicaid eligibility more efficiently.

What issues must be considered when expanding access to and integrating government agency data?

Efforts to combine and share agency data can be challenging because of real and perceived legal issues, cultural or institutional barriers, and technical and scientific limitations. However, policymakers and practitioners have found strategies to manage potential obstacles and pursue integrated data systems and data sharing agreements.

LEGAL

Efforts to improve data access and sharing must be balanced with individuals' rights to be protected against unwarranted disclosure of their personal information (see "How do you ensure privacy and confidentiality?" below). Authorization to access and share data varies based on their source and the purpose for which data will be used. The Privacy Act of 1974 regulates the federal government's collection, maintenance, use, and dissemination of personal information. Individually identifiable information cannot be publicly released. Additional federal legislative guidelines protect individual health records (via the Health Insurance Portability and Accountability Act) and education records (via the Family Educational Rights and Privacy Act). For tax data, the Internal Revenue Code stipulates that tax returns must be confidential, though exceptions have been made. State and local governments also have their own regulations for enforcing privacy protection within their data systems.

The complex legal framework created by these rules can make data integration without clear federal guidance seem risky to state policymakers. But it is by no means an intractable barrier to doing this work. Good legal counsel can help identify statutory flexibility and waivers, and data licensure, de-identification, and aggregation are all mechanisms that help make data integration possible within the legal landscape. Further, as the number of data integration efforts grows, policymakers will benefit from more case studies to guide their own work. Concerns regarding privacy should be taken seriously.

However, with the available methods of de-identification, these concerns should not inhibit data integration efforts that respect privacy laws.

INSTITUTIONAL

Impediments to data integration can also be cultural or institutional. Public agencies are often risk averse and protective of their data and may be hesitant to engage in efforts to increase accessibility. This is partly the result of reputational concerns about how researchers or other agencies might use their data and how sharing data exposes a program to potential criticism about management and effectiveness.[2] But government officials afraid of personal criticism or change to the status quo should not be barriers to innovation and improvement in social services. Developing strong, trusting relationships within and across agencies is critical to overcoming these challenges and building an institutional culture that encourages and supports cross-agency collaboration and data sharing.

TECHNICAL

The utility of integrated datasets depends on the quality of the data. Government data acquisition, auditing, and linkage procedures must ensure data integrity and completeness. Sometimes, an agency will collect information about a service recipient that conflicts with information collected by another agency; when that happens, the agencies must have a process in place to resolve the differences. These problems are often caused by uncommon language in data, when agencies collect similar but not identical information using similar but not identical data categories. Further, if agencies are collecting data as a matter of course and without specific purposes in mind, such as eligibility determination and policy research, the data may not be obtained or stored in a manner conducive to those activities. More importantly, the data may not track the outcomes of real interest.

Data integration requires multiple datasets to be merged using common fields. Agency databases are often legacy systems that were not developed in coordination with one another, so record linkage can be complicated if IT systems are incompatible. Creating standards and parameters for data collection and storage can help mitigate some of these challenges, and there have been efforts to facilitate infrastructure compatibility. Following the passage of the Affordable Care Act, the Centers for Medicare and Medicaid Services enhanced its funding matching rate for eligibility and enrollment modernization, increasing the level of federal support to 90 percent for new systems builds and 75 percent for maintenance and operations. The Office of Management and Budget also offers a cost allocation waiver to support the integration of eligibility systems for health and human services programs such as the Supplemental Nutrition Assistance Program and Temporary Assistance for Needy Families.

[2] Focusing research questions on strategies for program improvement—for example, examining why some officers in a given program are more effective than others—can reduce these concerns, as opposed to research that makes broader assessments about program effectiveness.

How do you ensure privacy and confidentiality?

Every data sharing effort must have strong protections in place to ensure the privacy and confidentiality of personal information. Proactive data governance principles, the policies and procedures an organization uses to manage data from acquisition to disposal, are critical to this.

ESTABLISHING PROTOCOLS FOR DATA SHARING

There are several preliminary tasks human services leaders should complete when establishing a data sharing initiative. Leaders should first develop memorandums of understanding or data use agreements "outlining the information [to] be shared, how [it] will be shared, and the protections of the information once shared." A team should be designated to determine what the "minimally necessary information" to be shared will be and who needs the information. Another team, made up of privacy officials and IT staff, should be responsible for determining how the information will be shared and protected. Extensive training on the policies and procedures of the information sharing project should be required for all members.

Brady and colleagues have outlined several similar guiding principles for providing data access while maintaining confidentiality. They suggest that departments or agencies handling data designate a "data steward" and structure their staff to adequately manage data access requests. These staff members should keep a catalog of written confidentiality agreements or memorandums of understanding signed by the researchers or outside staff interested in the data. Agencies should also have mechanisms in place for identifying and addressing data security breaches.

METHODS OF PRIVACY PROTECTION

There are several ways to protect the privacy and confidentiality of data while providing researchers, other organizations, and other federal agency staff with data access. One way is to limit and credential who has access to data and for what purposes. Researchers would submit a project proposal that, if approved, would allow them to work with typically on-site microdata. In a report from the chief data officer for the City and County of San Francisco for the Commission on Evidence-Based Policymaking, it was suggested that a "risk management perspective" be used when considering limitations on data access and use, where limits should be "commensurate to the privacy or security risk posed by the data." Instead of focusing on "whether others can *have* data," a risk management approach allows agencies "to designate *who* should have access and *for what purposes* based on a data classification scheme." Under this approach, review committees could be established to assess the risks and benefits of the proposed research, data sharing infrastructure could support role-based access, and data use agreements or terms of use agreements for data sharing could be created. For example, the UK's Administrative Data Research Service has developed methods for researcher training and credentialing. In addition, the European Union's Data Without Boundaries project designed a researcher "passport" to provide credentialed access across all European statistical agencies.

Another method of maintaining data confidentiality is by creating secure data enclaves. Data enclaves can be physical locations operated by on-site federal employees, or they can be virtual. Virtual enclaves typically use a client-server environment or virtual private network or allow credentialed researchers to host data on their own computers (with strict conditions). Margaret Levenstein of the Inter-University Consortium for Political and Social Research believes this method is particularly important because “many tasks associated with the work of turning administrative datasets into useful analytical datasets, including data cleaning, the production of metadata, and dataset linkages, can only be accomplished with access to identifiable data.”[3]

There are also techniques to help maintain privacy when integrating or matching administrative data from various sources. For instance, privacy-preserving record linkage matches records across databases in such a way that no information about the source data can be learned by parties involved in the linkage.

Collaboration with academic organizations can also help protect the privacy and confidentiality of data because excessive centralization (i.e., data housed under one federal agency) might compromise data security. For example, the Institute for Research on Innovation and Science, a collaboration of dozens of universities based at the University of Michigan’s Institute for Social Research, uses secure multiparty computing to create a networked data infrastructure that allows for the aggregation of administrative datasets from member universities without them residing in one location. Other academic collaborations do the same for administrative data from state and local governments.

CASE STUDY OF METHODS OF PROTECTION

The US Census Bureau combines legal frameworks and its own policies and procedures on good data stewardship to protect confidentiality and privacy. The agency provides a model for securing administrative data that involves a series of protections throughout the data’s lifecycle: acquisition, preparing files and provisioning data, and monitoring projects. The Census Bureau acquires data through data sharing agreements, which detail methods for file transfers and data retention and requirements for approved uses. After the agency obtains the data, designated staff work in a secure location to conduct quality control of the data and remove any identifying information. Only after the data are reviewed and cleaned will the “research versions” of the data be registered in the agency’s centralized data management system.

Project contacts submit proposals to the agency’s Policy Coordination Office outlining their work, including their methodology, objectives, anticipated output, and the benefit that the research provides the agency. If the proposal is approved, all of the project’s researchers are reviewed to ensure they have completed all necessary trainings before gaining access to the data. At the conclusion of the project, its results are reviewed by a disclosure avoidance officer to ensure that no identifying information is included. For administrative records, the Census Bureau adheres to applicable records retention schedules in regards to the disposal of data. The agency only retains original records for two years, and files that it creates are also destroyed after two years or when they are no longer necessary. This ensures that privacy and confidentiality are protected at every step.

[3] *Hearing of the Commission on Evidence-Based Policymaking* (2017) (statement of Margaret Levenstein) https://www.cep.gov/content/dam/cep/events/2017-01-05/2017-1-5-Levenstein%20Update.pdf.

What are the opportunities for expanding access to and integrating government agency data?

The huge range of opportunities for sharing and integrating data are evident in projects currently under way at government agencies, universities and other research institutions, and private vendors. A key next step is to better coordinate and institutionalize data access procedures and data linking efforts to avoid high transaction costs and duplication of effort, especially at the federal level.

In September 2017, the Commission on Evidence-Based Policymaking released 22 recommendations as part of its final report. Central to these recommendations is the establishment of a National Secure Data Service that would temporarily link datasets for researchers inside and outside of government agencies on an individual project basis. It would also provide researchers with a more streamlined process for accessing data. The report also makes several other recommendations related to security, which are particularly instructive, as well as evidence building and intragovernmental data sharing.

Although much of the focus on data access issues has been on increasing program benefits for eligible citizens and ensuring that safety net programs target the right people, attention should also be given to overall government efficiency and detecting and preventing fraud, abuse, and waste. More broadly, the constant collection and analysis of data reflecting the performance and shortcomings of various government agencies should help create a culture of continuous innovation and improvement.

Federal agencies need to be clearer and bolder in letting state agencies know they are permitted and encouraged to share data across programs. The appointed leaders of major safety net programs can help by setting a tone of collaboration across programs.

Where can I learn more?

- Researchers at the University of Pennsylvania have written about the promise of integrated data systems.
- The Office of Management and Budget released a set of white papers that provide background on how data can be used for evidence building.
- From the perspective of social services administrators, the American Public Human Services Association hosted a conference with presentations on how data integration tools could be used most effectively.
- An overview of how integrated data systems can improve the quality of services is available as part of a conversation with Dennis Culhane on Andy Feldman's Gov Innovator podcast.
- The Administration for Children and Families has created a Confidentiality Toolkit detailing how confidentiality requirements can be addressed in state and local information sharing and service coordination initiatives.
- The work of the Commission on Evidence-Based Policymaking is being continued at the Bipartisan Policy Center.

URBAN INSTITUTE

BROOKINGS

AEI American Enterprise Institute

THE PEW CHARITABLE TRUSTS | MacArthur Foundation

Evidence-Based Policymaking Collaborative

With support from the Laura and John Arnold Foundation, researchers from the Urban Institute, Brookings Institution, American Enterprise Institute, and The Pew-MacArthur Results First Initiative have formed the Evidence-Based Policymaking Collaborative. The Collaborative brings together researchers from organizations across the ideological spectrum to create tools to support evidence-based policymaking at the federal level. The Collaborative's work is assisted by an Advisory Group consisting of stakeholders throughout the evidence-based policymaking field. The opinions expressed in this brief do not necessarily reflect the views of all members of the Evidence-Based Policymaking Collaborative.