Unleashing the Power of Administrative Data

A Guide for Federal, State, and Local Policymakers

Photo credit: City of Mesa.

October 2017

RESULTS FOR AMERICA

AMERICAN ENTERPRISE INSTITUTE

By Robert Doar and Linda Gibbs

Executive Summary

Our recommendations are grounded in our firsthand experience with government data systems as former leaders at the city and state levels, where we oversaw programs including cash welfare, food assistance, public health insurance, child welfare, homelessness, probation and corrections, and child support enforcement. In each of these policy areas, we witnessed immense potential for all levels of government—working with nonprofit organizations and academic partners—to harness the power of data to maximize the impact of taxpayer dollars and improve services for the public.

We believe what policymakers need most—especially at the state and local levels—is practical guidance for overcoming the myriad bureaucratic, legal, and cultural hurdles that prevent government leaders from unlocking the full potential of administrative data.

Our key findings culminate in five recommendations for policymakers:

First, tackle data security and privacy concerns by developing a clear and shared understanding of privacy laws, both within government and with the stakeholder community. Implement appropriate technology to ensure personally identifiable information will remain confidential.

Second, create standard definitions for reporting administrative data and require implementation as a condition of local, state, and federal funding. Technological solutions can offer users a "continuum of access" that is aligned with their legal right to see and use the data. Most importantly, data must be used by researchers and policymakers to improve their quality.

Third, take steps to instill a sharing and learning organizational mind-set. Implement a governance framework that is guided by shared values and transparency to facilitate appropriate sharing of administrative data.

Fourth, create ease and comfort with using and sharing data by implementing data sharing in a tiered approach and open greater access over time.

Fifth, at the federal level, standardize the collection of data and aggressively pursue data-sharing agreements with state and local governments. By linking administrative and survey data, US statistical agencies and independent researchers can more accurately report on Americans' real conditions.

In addition, this paper offers exemplars of where this work has been done well at the state and local level to help address social challenges, from preventing child abuse to addressing homelessness. At the federal level, it provides a road map for the US Census—in partnership with other federal agencies—to take a lead role in driving systemic change in how we share and use data.

Unleashing the Power of Administrative Data

A GUIDE FOR FEDERAL, STATE, AND LOCAL POLICYMAKERS

Robert Doar and Linda Gibbs

Every day we see how data are used to make our lives more convenient. Enterprises such as Google, Walmart, and Amazon are using data to ensure that the products we most want are at our fingertips. We cannot buy a new jacket in a city 300 miles from our home without being alerted almost instantly that our credit card is being used in a place we have never been. We do not choose a restaurant without first checking the online reviews from other customers, and we do not get there without following the quickest route—even taking into account the traffic caused by the fender bender that took place less than an hour before our departure.

At the heart of all these applications are millions of pieces of information about who we are, where we live and work, and how we earn and spend our money. Using data is the prized skill that allows the private sector to respond to our every need or desire. But in the places where effectively using data can mean saving a child from abuse, preventing high school delinquency, or helping a single mother secure a job *and* child care, the use of data to guide decisions is woefully inadequate.

To be sure, the issue is not that government agencies lack data (indeed they often have more data, more accurately collected than anyone else) but that the repositories of these data are highly protected and bureaucratically controlled. Much of the country's administrative data—collected by government entities for program administration, regulatory, or law enforcement purposes—is underappreciated, underdeveloped, and underused.

At the same time, much of the information that *is* being used for government administration purposes suffers from a variety of inherent shortcomings—mainly related to the way in which the data are collected. Survey data, in which respondents self-report information, are often more biased and less complete than less-used administrative data, which a government entity typically collects. For example, the federal government's Census Bureau relies on self-reported responses to collect and build repositories of information about who Americans are and how they live, but it struggles to overcome problems of nonresponse, lack of accuracy in self-reporting, and other issues that come with the territory of citizens telling the government about their own lives. Administrative data, which do not rely on self-reporting, can shore up the gaps in information left by relying on Census data.

Among the government's administrative data collections, many are essentially internally maintained commodities. They have developed from the point of service delivery, at the program level. Sometimes they are held in the public agency, sometimes by a nonprofit provider under contract with the government.

Only a subset of state or federal funding authorities standardize data collection and reporting. Even then, the millions of pieces of data are not fully applied to policy evaluation or important statistical reporting. More data remain in agencies' internal files and not even centralized within an agency. When Linda Gibbs was appointed by New York City Mayor Michael Bloomberg as commissioner of the Department of Homeless Services in New York City in 2002, she counted 32 separate data-tracking systems in just that agency—with little integration or sharing even within the agency's four walls.

To truly unleash the data's value, government needs a set of value propositions, tools, and structures through which it can satisfy its governmental obligation to make the resource available to the public in a meaningful way. At the core of our argument is a deep belief that public data sources—de-identified and privacy protected—should be used to benefit the public and advance social progress. Government has a unique ability, and obligation, to ensure equal access to public goods, including the vast knowledge contained in its data systems.

The ultimate goal of this paper is to explain how administrative data could be better used and more widely shared across all levels of government and made available to researchers, nonprofit service partners, policy experts, and decision makers who could leverage these data to improve outcomes. While sharing data has many barriers, none are insurmountable. The ability to link administrative data sets with each other and survey data offers significant potential to answer important questions that neither type of data can do by themselves. In this paper, we will highlight the most significant barriers and offer a set of recommendations and activities to help policymakers knock these barriers down.

Seizing the Bipartisan Momentum for Data Sharing

Achieving systemic change will require action at every level of government, and we believe the time is ripe for action. Rapid technological advances are enabling us to mine reams of administrative data to improve program management and inform policymaking while enhancing our ability to protect privacy and data security. New partnerships among social service providers, local and state governments, federal agencies, and academic institutions offer innovative models for how to tap the deep well of knowledge contained in various data sets.

These trends all point to more widespread evidence-based policymaking. An evidence-based government is one where, for all crucial policy decisions, actionable information is available when and where needed.[1] That government officials should base their decision-making process on facts is eminently sensible and is an area that attracts bipartisan support.

From City Hall to the White House, leaders of both parties have increasingly been pursuing efforts to open government data to the public. Last year, San Diego Republican Mayor Kevin Faulconer released dozens of municipal data sets as part of the city's open-data policy after asking the public to vote on which data sets to open first.[2]

The Obama administration made improved access to high-quality data a pillar of its emphasis on evidence-based policymaking. In an executive order concerning the Freedom of Information Act, President Barack Obama could not have been more direct:

> The Government should not keep information confidential merely because public officials might be embarrassed by disclosure, because errors and failures might be revealed, or because of speculative or abstract fears. Nondisclosure should never be based on an effort to protect the personal interests of Government officials at the expense of those they are supposed to serve.[3]

One of the most important developments in evidence-based policymaking came in 2016, when Congress passed bipartisan legislation creating the Commission on Evidence-Based Policymaking (CEP).[4] Established by the urging of US House Speaker Paul Ryan (R-WI) and Sen. Patty Murray

(D-WA), the CEP recently endorsed a plan to expand the availability of data while strengthening the federal government infrastructure for secure access to data. As part of this broad recommendation, the commission recently made the following important recommendations to improve the availability and use of government data for policy evaluation and research:

1. A National Secure Data Service should be established, which would generally expand and maintain federal government infrastructure aimed at linking government records and increasing access to data.

2. Such a service should have the ability to leverage public-private partnerships into new research and technologies.

3. The Office of Management and Budget should do more to make information available and searchable in existing federal inventories, data sets, and data documentation.[5]

The very creation of such a commission marked an important step in this effort, underscoring the importance of increased use of and access to data in modern governance. However, the commission's recommendations (which were limited to the policy areas specifically outlined in the federal legislation that created it) stopped short of discussing factors at other levels beyond the federal government that would mark further progress in accessing usable information. Among these are encouraging a shift in the culture of data accessibility in state and local governments, improving data sharing with services providers, and empowering the Census Bureau to create more expansive data-sharing arrangements with state agencies to allow information to travel back down to the localities that produced it.

Many of our suggestions build on the commission's recommendations and echo some points made in the commission's final report. We hope to add to these recommendations and provide an insider's perspective of the agencies the commission hopes to influence.

Administrative Data: What Is It, and Why Is It Important?

Administrative data are information collected by government entities charged with administering a public program. Typically, beneficiaries provide information as a condition to receiving a particular service. For example, state workforce development programs collect information on job seekers, while nutrition assistance programs collect data on families receiving aid. Public administrative data vary across programs and government entities, but their fundamental purpose is to aid in the administration of a government program.

From city to state to federal governments, each agency has numerous unique data sets containing information on the populations they serve. These agencies may also contract for services with nonprofit organizations that collect a great deal of population-based information. These data sets sometimes track the capacity and use of a service, such as receipt of nutrition assistance or housing vouchers or Medicaid or Medicare use. They can also go so far as to maintain detailed information on client use and experience in the service and financial data associated with program operation.

While they are not designed for explicit evaluative or statistical purposes, administrative data are usually collected routinely on the full universe of individuals affected by a particular program. In addition to having large numbers and offering longitudinal study opportunities, administrative data suffer from fewer problems associated with attrition, nonresponse, and underreporting in survey information. In particular, when data elements are deciding factors for program eligibility, agencies go to great lengths to verify their accuracy.[6] With this detailed information, administrative data sets can help answer questions about educational success, labor market outcomes, health risks, justice involvement, housing stability, and other areas related to program participation and results.

A prominent example is Pennsylvania's Allegheny County Department of Human Services, which for many years has been a leader in integrated social service data use. Their experience demonstrates how greater use of existing administrative records improved decision-making and program outcomes.

In 1999, the county's Department of Human Services (DHS) created a centralized repository for all records related to human services and client information, referred to as the Data Warehouse (DW). Initially linking internal data from health, welfare, and homeless services, the department slowly added sources from other agencies such as the Department of Public Welfare, local public school authorities, and the criminal justice system. The DW now includes nearly one million client records with demographic information (e.g., name and date of birth), receipt of past or current services and associated costs, and provider information (e.g., name, location, type of providers, and services delivered).

The Allegheny DW is used for a variety of analysis and research purposes, resulting in greater efficiency in the delivery of services while complying with federal, state, and local laws protecting privacy and data ethics.[7] For example, the greater ease of data analytics allows the DHS to implement the Allegheny Family Screening Tool, a predictive-risk modeling tool that identifies children at risk of abuse and aligns financial resources and staff to these at-risk children.[8] Administrative data can also be used to conduct research or improve delivery of a particular service.

Data Collected by Government and Used by Research and Service Delivery Partners. While serving in government in New York State and New York City, we clearly understood the usefulness of the data our own agencies collected. Using the data systems common in every state, we could see who was receiving food stamps or other welfare benefits, in what neighborhoods, and in what types of families. Regarding some of the nation's largest safety-net programs, we knew how much assistance New Yorkers were receiving from various government programs operated under our roof.

Nevertheless, our staff and other state and local officials did not share this information with each other in any organized, comprehensive, or effective way. This limited our ability to truly appreciate the complexities, challenges, and opportunities in our clients' lives. With the goal of creating that holistic view, we began an arduous process of weaving together the data from nine different agencies serving a total of more than 2.5 million New Yorkers.[9]

Our own approach to data integration was driven by our desire to share case records at the front line with workers simultaneously involved in a single family's life. Worker Connect emerged from this effort.[10] It is a case-management solution that merges six foundational data files through a process of "entity resolution," connecting unique clients across multiple services and presenting a single, comprehensive record of the entire family and their current service use and involvement across agencies. It is fully privacy protected, and authorized users are allowed access only to the information they are legally authorized to see.[11]

The greater ease of data analytics allows the DHS to implement the Allegheny Family Screening Tool, a predictive-risk modeling tool that identifies children at risk of abuse and aligns financial resources and staff to these at-risk children.

Worker Connect has allowed workers to understand the bigger picture when a new case is added to their caseload, permitting child-welfare workers to appreciate the full family composition on their way to a child protective investigation, helping emergency room assistants authorize services for families that appear in the middle of the night without documentation, and allowing probation officers to understand the household circumstances surrounding a young person whose life they are trying to keep on track. The system gets 220,000 hits a month from more than 2,000 authorized users and connects clients to digital images of important documents needed for eligibility verification, which often were otherwise lost or misplaced in the multitude of moves poor households often make. Some agency caseworkers from nonprofits with contracts from the city have recently been afforded access to the Worker Connect system, extending its benefits to those who work daily with clients across the city.

Administrative data sharing has also been shown to yield valuable information when employed by social scientists and researchers, informing decisions beyond eligibility for program receipt and answering complex questions. For example, using naturally occurring school choice lotteries and student-level administrative records from public schools, Howard Bloom and Rebecca Unterman in 2014 showed that New York City's large-scale high school reform efforts had increased graduation rates and reduced education expenditures per graduate.[12] In a separate study by the Education and Incarceration Project, academics used state administrative records on the Supplemental Nutrition Assistance Program (SNAP) and other welfare programs to study the effects of earning a GED while in prison.[13]

In a parallel effort to Worker Connect, the Center for Innovation through Data Intelligence (CIDI) has brought an analytic approach to data integration. CIDI arranges standing data-sharing agreements among agencies interested in exploring multiagency analytic questions. These agreements are triggered by an original research question.[14] Once established, the agreements remain in place as additional partners are added to the networked data approach.

Whereas Worker Connect provides snapshot case-management data, CIDI conducts longitudinal research and evaluation. When CIDI researched the conditions that led youth who exited youth shelters, family shelters, or foster care to become homeless, they found that housing vouchers reduced recidivism and that youth in youth shelters and foster care did better than youth in family shelters (which had fewer youth-centered services). They are now working on a coordinated entry system so youth receive positive youth-development services no matter which door they come through.

When the hard work of sharing and integrating these incredibly rich data sets is complete, remarkable power is unleashed at the individual case level and the policy level.

Implementing Change at the State and Local Level: Barriers and Solutions

While there are many barriers to unleashing the full potential of administrative data, fortunately there are also many practical solutions. The overarching philosophy that grounds these solutions is that administrative data should be seen and used as a public good. Local governments should embrace broader administrative data sharing because doing so fully aligns with their own values of transparency, accountability, and collaborative problem-solving. It can reveal the benefits (or unintended consequences) of programs that tackle shared concerns. And these data sources can provide the best evidence for evaluating if the investments return benefits that are worth the costs.

For predictable reasons, data are jealously guarded, and access is stingily given. But the obstacles facing agencies—real and perceived—can be overcome. We want to see more places using data like Allegheny County and New York City do and greater opportunity for researchers, advocates, and communities to offer their expertise and advice on complex policy questions.

What we have now, however, is a situation in which each locality is attempting this on its own, repeating efforts and developing unique solutions. This is wasteful and possible only for those jurisdictions able to apply the time and money to the undertaking.

More can be done to move to a common set of approaches, definitions, platforms, and uses that can allow localities to build off each other. The end goal is to have state and local agencies and their partners have greater ease and ability in sharing data sets and researching the effectiveness of policy and programs.

To help achieve this goal, we have outlined common barriers to using administrative data and paired them with recommendations to address those challenges and tools to navigate the effort (Table 1).

Barrier 1: Privacy and Data Security. One of the biggest barriers is the need to secure privacy and assure service recipients that their personally identifiable information will remain confidential. Since administrative records are collected with a particular decision-making purpose, a responder's unique identifiers are generally recorded, making privacy a justifiable concern. Adding to the complexity, privacy laws have been at best misunderstood and overbroadly applied and at worst used as a pretense to deny sharing requests that could be satisfied.

For example, the Family Educational Rights and Privacy Act is interpreted differently among states, localities, and researchers about whether and how it is permissible to use information from student education records for evidence and policy evaluation purposes.[15] In many cases, when the federal and state laws around the authority to share data are imprecise or silent, states default to "no."[16]

Related to privacy is the argument about data security. Under existing protocols, individual de-identified data are shared after personal identifying information is removed or encrypted such that the individual cannot be identified. Two objections are raised by security skeptics. One is that the security systems themselves are inadequate and reidentification is possible. The other is that some data sets are so small that it is possible to figure out who the individual is anyway. The combination of confidentiality challenges and security concerns often makes agencies reluctant to share data with their government or research partners.

Solution 1: Privacy Guidance and Technology Approaches. Federal privacy laws set national standards for protecting information that might identify subjects, and state laws may go further to offer additional protections. A core value of the philosophy that data are a public good is that carefully guarding these protections must be core to the mission of unleashing

Table 1. Barriers and Solutions Summary

Category	Barrier	Solution
Privacy and Data Security	Privacy laws are often misunderstood, and many agencies lack access to the proper technology to ensure personally identifiable information will remain confidential.	Provide guidance on privacy laws. Communities must reach a consensus on application and appropriate technology to ensure personally identifiable information will remain confidential.
Data Curation	Administrative data are collected with diverse and inconsistent goals, definitions, and reporting units. The data often contain errors and reflect organizational biases.	Create standard definitions and require implementation as a condition of state and federal funding. Technological solutions can be used as an alternative. Most importantly, data must be used to improve their quality.
Culture and Governance Framework	Many agencies have a culture of restricting data access for false reasons rather than instilling a sharing and learning organizational mind-set.	Instill a belief that data are a public good. Implement a governance framework that is guided by shared values and transparency to facilitate appropriate data sharing.
Capacity	There is a lack of ease and comfort with using and sharing data among many government agencies.	Implement data sharing in a tiered approach, opening up greater access over time.

Source: Authors.

population-level information. Stewards of public data must ensure data are effectively de-identified before they are publicly shared.

In this regard, the CEP's report is instructive. To the commission's credit, it took great pains to envision a system of data sharing that removes and deletes direct identifiers from accumulated administrative data before they are cleared for public release. Local and state agencies should follow their lead.

Additionally, in some cases, private data may be shared. That is typically true at the casework level, where a worker's responsibility for a client's case authorizes him or her to access information that is not otherwise shareable. Integrated data systems can greatly facilitate effective case management by providing a holistic view of the client and the household.

Hence, protection of privacy requires clear guidance of what privacy laws protect and permit and the standard tools that can be used to easily and uniformly adopt these locally. And it requires that data gathering is responsibly constructed, that integration and dissemination technology is used to ensure privacy is protected, and that access to levels of secure data is carefully structured to align with users' legal right to see and use the data.

Privacy Guidance. In the wake of growing concern about data privacy and cybersecurity breaches, government entities and private firms have focused on establishing policies to inform consumers of a data breach and guide them through the aftermath of such an event. Forty-eight states and the District of Columbia have established policies requiring government or private entities to notify individuals of security breaches of personally identifiable information.[17]

At the federal level, the Department of Justice in 2015 released guidance to organizations on preparing plans for addressing cybersecurity incidents. Cities and counties have addressed the problem by adopting formal plans and frameworks for addressing and preventing breaches. Private entities have also been instrumental in establishing plans, as evidenced by a data breach response policy developed by the SANS Institute.[18] The bottom line is that governments should establish frameworks for both preventing security breaches and limiting the fallout if they occur.

Achieving this goal requires improved guidance on privacy protections and what they do, and do not, allow. Because local laws vary, it also requires establishing the steps a locality must take to determine what special considerations should apply. This includes understanding the legal status of various data sources (internal versus external sharing), any special state and local privacy laws that add on to federal protections, and who may and may not share access to confidential data.

Jurisdictions would also be aided by a common set of standard data-sharing agreements, client confidentiality waivers, and guidance to support this work.

Appendix A discusses these support tools in more detail.

Definitions of Levels of Allowable Access Open Data

Open Data: Data that are not unique to an individual (tabulated summaries) or are unique to an individual but are stripped of identifying information, which do not require user authorization and are immediately accessible to the public

Public Data Shared with Controlled Access: Data that may be tabulated summaries or unique to an individual but stripped of direct identifiers, which for security or policy reasons are not publicly shared but are available to authorized users through secure access (e.g., de-identified information available to administrators without public release)

Secure Data with Provisioned Access: Data that specific users with authorization are given access to based on a unique user identity (e.g., a caseworker's access to the integrated data file of a client being served)

The City of Seattle has implemented this kind of approach in an effort to use data and evidence to help the homeless find permanent housing. Executive Order 2016-05 directed the City of Seattle departments to expand their use of data and analytics in everyday management and strategic decision-making to ensure performance and accountability measures are integrated across city government.[19] These actions are intended to apply results-driven methodologies to city programs to better analyze and measure good governance, transparency, and effectiveness.

To further this, All Home King County and United Way of King County signed a Memorandum of Understanding (MOU), committing to a shared set of performance measures for the agencies they fund.[20] The MOU became a way to align the community priorities across the entire network and tie funding to outcomes that improve the system's effectiveness. The data sharing needed to facilitate this work was enabled by the Homelessness Partner Agency Privacy and Data Sharing Agreement. This effort was further supported by the Washington Homeless Client Management Information System Law.[21]

To ensure protection of privacy, localities must have a strong understanding of how the technology has advanced and what requirements are in place to ensure best practices.

Enhanced Technology. As program administrators, data owners and policy leaders are not typically experts in IT engineering. Their job is to hire the right people and make sure they do the job with the strongest technology for the task and with great integrity. To ensure protection of privacy, localities must have a strong understanding of how the technology has advanced and what requirements are in place to ensure best practices.

One of the great facilitators of increased data sharing is that the technology has advanced to provide better and better security while costs are plummeting. Building systems that more securely encrypt privacy-protected information allows for greater access to the data.

Nonetheless, for integrated data systems that depend on the ability to integrate facts across multiple data sets, connecting individual data is key. This process is referred to as data linkage through entity resolution, in which both data sets must have overlapping identifying information on an individual level. The necessary IT security systems must be in place to ensure no breach of privacy results from the technical solutions. Coding solutions are available that conduct the matching behind a privacy wall and expose only the matched data sets at a population level through a "one-way hash"—a coding algorithm—without the ability to re-identify any one individual.

Improved data-integration techniques ease the smart use of data. Refinements to approaches allow jurisdictions to understand clients' relationships among service systems and build early warning alerts on the client level when life goes awry. To achieve all this, jurisdictions should be supported with clarity on role definition, straightforward explanations of basic coding techniques, and management practices to ensure continued integrity.

All these advances have all but eliminated the opportunities for system breaches. Nonetheless, authorized users can abuse their privilege. Continued integrity of a system must include clear standards, strictly enforced, if any breach does in fact occur.

Guidance on these privacy and technology elements is included in Appendix A.

Barrier 2: Data Curation. Data curation is the management of data throughout their life cycle, from creation and initial storage to when they are archived for posterity or become obsolete and are deleted. The main purpose of data curation is to ensure that data are standardized when compared across jurisdictions and are reliably retrievable for future research purposes or reuse. There are several barriers to effectively managing the data curation process, including errors resulting from incorrect merging of data, the cost to produce some data, bias that can be reflected in the data, and lack of analytical capacity.

The challenge begins from the fact that administrative records are not intended for research purposes and can be quite heterogeneous and unstructured. This can also be seen as a consequence of the organic development of data collection—each data system has its own definitions, and the reporting units are defined by the originating source. Even within cities, multiple agencies can define and measure similar activities very differently. Conversely, the same label can be given to two distinct activities.

Another frequent refrain is that the data are "dirty"—too filled with error to be valid, not worth the cloud they are stored in, and worse, if used could produce such distorted conclusions and reports that their release would be reckless. Additionally, because data mirror the behavior of people and institutions and because bias, both implicit and explicit, persists, data sets mirror that bias and run the risk of perpetuating that bias if not interpreted with an eye for where and how the information is true.

These obstacles require a number of steps to "clean" or curate data so that they are suitable for public use or integration with other data sets. Crucially, staff and technical infrastructure must be available to structure and convert the data into usable formats. The capacity to do this varies across localities and may be limited or, in many cases, nonexistent. In addition, agencies increasingly have internal IT offices (although these have been subject of late to consolidation in central citywide serving entities), where internal competing demands can put them in the position as arbiter of priorities for data production.

This is in contrast to the principal statistical agencies (e.g., the Census Bureau and the Bureau of Labor Statistics), whose missions dictate investing in dissemination tools and distributing the data sets that external sources could use. Local agencies, whose mission is administering a particular program, generally do not make comparable investments.

Solution 2: Standardization of Data Definitions. The solution to flawed data curation includes standardizing through common definitions and measurement, standardizing through state technology and statistical products, and cleaning data through ongoing use by program managers, policymakers, service providers, and the public.

Where a state or federal funding source has standardized the field, localities have adopted common definitions and measurements. State and federal governments can create data standardization through reporting in three forms: (1) required as a condition of funding, (2) voluntary but incentivized with funding, or (3) simply recommended.

An example of a mandatory reporting format was in the field of education, where recipients of federal education funding were required to report educational outcomes in prescribed ways, helping to move the field as a whole to a more standardized format.

An example of an incentivized system is the child welfare field, where the federal government made substantial investments to incentivize states to build automated data systems, at first through statewide automated child welfare information systems, updated through the Comprehensive Child Welfare Information System in 2016. While not mandatory, accepting federal funds to automate the child welfare case-management records subjected states to review and approval of federal parameters. The process allowed the federal government to move state systems to a more standardized set of data definitions and measures in child welfare.

The most significant completely voluntary effort to offer standard definitions has been the National Information Exchange Model (NIEM).[22] NIEM has attempted to create a standard set of data definitions to facilitate information exchange across public and private organizations.[23] Unfortunately, adoption of this standard has been limited. This may be because the categories were too refined to be useful for the intended user.

NIEM would have been useful, for instance, when we in New York City automated our entire human services procurement system, converting the requests for proposals, submissions, awards, and contracts onto an electronic platform. To do this, we needed to standardize categories across agencies for the first time, putting similar services into similar buckets, allowing all qualified vendors to learn about and bid on contracts. We explored NIEM as a standard but found it incomplete and inappropriate to our needs. We built our own categories, which are now in use for contracting more than $1 billion in services in New York City annually.[24]

Alternatively, the nature of available data technology and statistical products is evolving so rapidly that the requirement of data standardization and curation is fading away as a legitimate reason for why data cannot be shared. Another approach would be to accept that uniformity may not be achievable, given the unique local nature of a vast array of service programs and delivery mechanisms, and have the technology structure create the solution. In a process referred to as extract, transform, and load (ETL), unique data sets are guided through translation processes to transform underlying data into shared formats to create a standard presentation across jurisdictions. This is valuable for creating apples-to-apples comparisons of outcomes and advancing national knowledge from local data.

Once the data have been standardized, the challenge of data error remains. Errors in the data will be a reality, as surely as humans transpose information incorrectly or adopt shortcuts that repurpose underused fields for more urgent data-collection needs that engineers did not anticipate. We have found that the best way to clean up dirty data is to use them.

For example, when New York City was moving to performance-based contracting for child welfare providers, a huge barrier came from providers who would put in front of us extensive errors in the underlying data—some of their own origin and some the public agency's. Rather than succumb to the conclusion that we could in no way use the data, we adopted a transition approach that put the data out initially as informational and did not attach rating consequences until a period of use expired. We then added an audit challenge as part of the process, allowing legitimate ongoing issues to be corrected before final performance rankings were adopted. Putting dirty data into the light of day is the best and fastest cure and one that all levels of government can accomplish.

Montgomery County, Maryland, implemented an open-data bill in 2012 and has seen innovation and improved effectiveness in government services, better transparency in budgeting, and increased evidence-based decision-making as a result.

Another strategy for increasing the use of data and thereby the quality of data is to create open-data policies. Almost 70 cities, states, and counties across the country have implemented open-data policies. Montgomery County, Maryland, implemented an open-data bill in 2012 and has seen innovation and improved effectiveness in government services, better transparency in budgeting, and increased evidence-based decision-making as a result.[25]

Finally, bias in data mirrors bias in society. Basing policy and program decisions on the "facts" as presented in the data guarantees those biases will persist. In fact, data scientists have determined that machine learning actually exacerbates these data biases. Data scientists are quickly developing strategies that detect and adjust when biases exist. These are far from perfect and will evolve quickly. Users of data need to be aware of the potential for bias, use techniques to detect and minimize it, and interpret results with an eye for questioning the reported data when pursuing policy and program decisions.[26]

Appendix A provides additional information on tools for data curation.

Barrier 3: Culture and Governance Framework. In addition to technical capacity barriers, jurisdictions may face barriers from cultural issues at agencies, including beliefs about the ownership of data, views about data's role in public policy, and awareness (or unawareness) of the data's potential to inform policy and practice. Commonly, agency leaders do not perceive themselves as stewards with a responsibility to share; rather, they see themselves as owners who treat this valuable public asset as a private good. This position may be formed by fear of what would be revealed—no administrator wants someone else to know something about their agency or service before they do, good or bad—or simply a sense of defensiveness over a particular populace and its governance. Either way, administrators would most often prefer to keep requested data hidden than expose the agency to the potential for embarrassment, demand for improved outcomes, or even program elimination. Other cultural issues include resistance to being reduced to "a bunch of numbers," sensing a lack of respect for the skilled professionalism and expert judgment of staff and policymakers with years of irreplaceable experience.

Many times these cultural issues are masked by an agency's claims that the public, without the right credentials, is simply not in a position to responsibly use the data. This claim serves the administrator, not the client or the public, and looks past the ability of information gatherers and analysts to affect public goods in the long term.

Solution 3: Introduction of Governance Structures. Offering a governance framework can ensure that participants come to the table guided by shared values and transparency. Importantly, it also creates an infrastructure that will live beyond the tenure of the committed leader, and it will allow successors to carry the work forward. The starting place for this is an Integrated Data System charter.

One effort underway that follows this approach is the Obama administration's My Brother's Keeper Equity Intelligence Platform (EIP). In an effort to shed light on the gaps in outcomes for boys and men of color, the EIP is being developed as a model for gathering and sharing local data. It is a national effort, with a national advisory board governed by a national charter.[27] Additionally, the platform is being prototyped in Oakland, and a local advisory board has its own project charter.[28] These documents have been useful in clarifying roles among partners and communicating the purpose to the broader community.

Stakeholder Engagement. A data-sharing effort's success will be defined by trust. As costs drop and security improves, the time and effort barriers will dissipate. What will remain is the need for citizens, clients, agencies, and leaders to develop an appreciation for the value of the data to help them achieve the vision they each have for their neighborhoods and city. To do that, there must be trust that the data are not to be used against them but for them.

Engaging each stakeholder in the process of achieving this vision is the most surefire way to get there. From designing the platform to agreeing on the data to collect to establishing agreements on access and prioritizing analytic resources, a collaborative engagement process provides confidence that the data are being generated in the public's interest.

In Oakland's work on the EIP, the local Youth Ventures Joint Powers Authority has conducted extensive engagement with public agencies and leadership, while PolicyLink and Urban Strategies have led grassroots-level communications with youth, community members, and nonprofits. The design firm, ISL, engaged representatives of all these groups in providing user input to the platform's functionality to ensure it is meeting local users' needs and interests.

Policies and Procedures. While the charter can lay out the guiding framework, issues will be raised and resolved and choices will be made on the day-to-day functionality of the system. It will be wise to capture these in documentation that preserves the decision and offers transparency to all involved. This would include procedural bylaws, standards for access and special user provisioning, decision-making on data prioritization and research agenda setting, rules and procedures for Internal Review Board reviews and authorizations when needed, and guidance on analysis and interpretation. In many cases this will also be facilitated by standard use templates.

Appendix A includes more information on structuring effective governance agreements.

Barrier 4: Lack of Incentives. Even when all the hard work is done and the data are ready to be shared, willing partners may balk at the most ambitious data-sharing plans. This will likely be true particularly in the early stages of sharing. What may be technically possible and legally allowable to share may not be comfortable to share. Enthusiastic encouragement to "do the right thing" will only get you so far. Trust will have to be developed and comfort gained in being the steward of data as a public good.

Compounding this problem is the lack of incentives for public officials to share data. These officials are often placed in a position whereby they may fear punishments for perceived misuse of data—including the confidentiality and security concerns mentioned above—while seeing limited potential reward to the improved outcomes associated with data sharing. This asymmetric incentive structure often leads to public officials limiting or preventing data sharing in their own best interest.

Solution 4: Progressive, Tiered Release of Data. Some approaches to consider to build confidence and fluency over time include providing tiered access to data. Essentially, this would involve having large data sets with the most expansive access among a limited set of super users, such as mayors, agency heads, and program staff. Winnowing down from there, a broader set of users with preapproved status could be provisioned to see much but not all of those data. This might include research partners whose skills are trusted to responsibly extract from and interpret the data. It might also include frontline staff and non-profit partners, as access to unique data sets might facilitate their practice and understanding.

The final tier would be to expose a defined highest level of data to open, public access. Users at this level would include, essentially, everyone else. They might be students, community groups, advocacy organizations, and citizens. This would require structuring technology to allow access in a process referred to as provisioning.

Another concept that would ease access to the data would be to adopt standards on aging data before use. A partner unwilling to share data in real time—letting the whole world see it at the same time the mayor or director sees it—might be more comfortable if a delay in release postpones public access for a defined period of days or weeks.

Enhancing the Strategic Use of Administrative Data at the Federal Level: The Role of the Census Bureau

Overcoming the challenges for state and local agencies will not be easy, but once accomplished, the potential benefit for policymakers is immense. Armed with data that have been unified in measurement and language and that provide insights into trends from all corners of American life, leaders at all levels of government can begin to draw a clearer picture of what the data tell them about various social, economic, and policy phenomena.

Data accrued from state and local agencies will directly inform policy in tangible, practical ways. Precedent for using data to measure trends with significant policy implications can be found in the work of economist Raj Chetty. Together with several coresearchers at the Equality of Opportunity Project, he used administrative data from the IRS to measure the effects of several factors on economic mobility.

With the power of administrative data in tow, Chetty was able to show how families' geographic moves affect children's chances of being upwardly mobile,[29] which counties present the best opportunities for mobility,[30] and even which colleges have seen the best results in graduates' mobility.[31] These findings not only challenged years of conventional knowledge but also reinvigorated the policy discussion of housing vouchers, college admissions, and antipoverty initiatives. Chetty's work shows that data use is not strictly academic; there are significant policy implications to the knowledge gained from data.

Efforts like Chetty's also showcase data's power to effect two-pronged change in governance. First, the quality of the programs provided by the federal government can be improved, by identifying and targeting specific communities that stand to benefit from injections of incentive-laden programs that might increase mobility. And secondly, the operations themselves can be streamlined. More efficient operations and targeted and better-informed policies are both possible and within reach; the key is opening a pipeline of data between local- and state-level agencies that hold the keys to such data and federal agencies that could harness and direct the data's power.

Beyond state agencies sharing data with each other and their partners, data-sharing arrangements with federal statistical agencies, the Census Bureau in particular, should be pursued with vigor. Building these arrangements not only produces valuable statistical products for all parties involved but also promotes evidence-based policymaking.

To meet this goal, federal statistical agencies must continue to address gaps in their knowledge and adapt their methodology to meet the demands of an increasingly data-driven nation.[32] This will involve the Census Bureau creating more robust data-sharing arrangements with state agencies, particularly those administering safety-net programs. Most promising is the linkage of state-level, administrative data with household survey data conducted by the Census. Such linkages have been shown to allow statisticians to exploit the best aspects of these data sources while minimizing their weaknesses.

This partnership faces formidable barriers. The perceived and real legal challenges, as well as the aforementioned technical and financial constraints, generally make state administrations reluctant to share their data with statistical agencies unless required to do so for regulatory reasons. The necessary technical infrastructure and methodology for linking administrative and survey data, however, has already been successfully employed by the Census Bureau. For decades they have routinely done so in a cost-effective and secure manner.[33]

The quality of the programs provided by the federal government can be improved, by identifying and targeting specific communities that stand to benefit from injections of incentive-laden programs that might increase mobility.

Building this partnership will require clarifying existing laws around data sharing and expanding the Census Bureau's role as a hub for data integration. Such arrangements are crucial for improving some of the most important policy-guiding statistics.

The Decline of Household Surveys. Linking administrative and survey data is growing increasingly important in the wake of the declining quality of household survey data. Survey data, in contrast to administrative data, are collected by a statistical agency to understand greater economic or social trends or the impacts of various social programs. Examples of important and highly used surveys include the Survey of Income and Program Participation (SIPP),[34] the American Community Survey (ACS),[35] and the Current Population Survey (CPS);[36] data from these surveys are used by policymakers to shape and evaluate programs at all levels of government, are a primary resource for economic studies, and are the source of important economic trends such as the official rates of poverty, unemployment, and inequality. Unlike administrative data, survey respondents typically have limited incentives to accurately or completely answer survey questions.

A growing body of evidence suggests that survey quality has been in decline, not only impairing our ability to implement and evaluate government programs but also distorting our view of the true economic condition facing Americans.[37] Households have simply become less inclined to respond to surveys, and when they do, they are less likely to answer certain questions and provide accurate information, particularly when they are being asked about receiving public assistance.

Bruce Meyer, a leading scholar on the quality of household surveys and current commissioner on the CEP, illuminates the consequences of this decline. Through linking New York State administrative data on four transfer programs—SNAP, Temporary Assistance for Needy Families (TANF), General Assistance, and housing subsidies—with New York CPS data, Meyer and his coauthor, Nikolas Mittag, revealed receipt of program benefits over a four-year period (2008–11) was missed in survey data for more than one-third of housing assistance recipients, 40 percent of food stamp recipients, and 60 percent of TANF and General Assistance recipients.[38]

These findings are largely consistent with a more comprehensive study on several other important household surveys conducted by Meyer, Mok, and Sullivan in 2015.[39] Through comparing survey results with administrative data from federal and state welfare agencies, Meyer and his coauthors found that survey measures of public assistance receipt and the value of that receipt were both sharply biased downward. In one of our country's largest cash welfare programs, TANF, four of five surveys failed to capture more than half the dollars given out. Even in SIPP, which is designed to capture nonlabor income, more than one-third of TANF dollars were missed.

Meyer explains that the major contributors to this underreporting are beneficiaries' lack of response to surveys, assumptions made about unanswered questions, and most importantly, measurement error, which have all increased markedly over the past three decades. There are several proposed reasons for this phenomena: People are concerned about privacy, have less leisure time to spend on answering surveys, or feel a stigma around questions pertaining to dependence on "welfare." Even if a survey recipient fullheartedly wants to participate to the best of his or her ability, sometimes the requested information is simply difficult to recall.

The underreporting of transfer receipts and incomes leads to an overstatement of poverty and inequality. In a time when citizens and government officials alike are demanding greater accountability over public resources, it is crucial that we more properly evaluate some of our largest public programs. Experts have long been pointing toward administrative data as an increasingly valuable source for correcting underreporting in surveys.

The Census Bureau's Experience and Other Examples. As far back as 1977, the *Report to the President from the Privacy Protection Study Commission* recognized the benefits of using administrative data for statistical purposes.[40] Accordingly, the Census Bureau has been pulling administrative data from the IRS, the Department of Housing and Urban Development, and the Centers for Medicare and Medicaid Service, among other sources.[41]

The Census also uses certain administrative data to frame and design many of their surveys. For example, the US Postal Service and local government data are primary sources to derive the master address list for the Decennial Census.[42]

Much of the methodology required to link administrative and survey data has already been established, and the Census is leveraging this expertise for several projects:

1. The Longitudinal Employer-Household Dynamics program combines federal and state administrative data and survey data with Unemployment Insurance earnings data and the Quarterly Census of Employment and Wages, providing states and localities detailed information on geographic and industry job flows.[43]

2. The Census Longitudinal Infrastructure Project integrates administrative records held at the Census Bureau with core linkable files from the ACS, the CPS, and the 1940, 2000, and 2010 Census, enhancing information on the American population across several decades and offering more opportunities to evaluate the quality of surveys.[44]

3. The American Opportunity Study is a collaborative project with the National Academies of Science, Engineering, and Medicine and researchers at Stanford, the University of Michigan, and the University of Wisconsin, looking to expand the data linkages backward in time to enable new research on social and economic mobility.[45]

4. The Mortality Disparities in American Communities project combines ACS data with death certificate information to provide differentials in demographic and socioeconomic characteristics on mortality.[46]

5. The Next-Generation Data Platform is a Census partnership with the Food and Nutrition Service and Economic Research Service that links participating state-level SNAP administrative data to the ACS survey data to measure program performance in US Department of Agriculture food assistance programs.[47]

How Survey and Administrative Data Are Linked. Linking administrative and survey data is more easily done than many perceive. The previously listed projects demonstrate that linking administrative data sets with each other and survey data could answer important questions that neither type of data can do by itself.

For the purposes of the Census household surveys, linking administrative data can enhance the accuracy of some variables, correct underreporting of transfer incomes, and reduce survey burden by eliminating questions that can be satisfied from the administrative record. To accomplish this, holders of administrative data would need to be required to share what they have with the Census Bureau, and the Census would need to expand its role as a hub for data acquisition and integration. The following steps could then facilitate this linkage:

1. State agencies and Census officials establish an ongoing process to collaboratively identify data sets and data elements with potential for statistical use, with the Census providing technical documentation and other assistance to assess the quality of particular data sets.

2. Safeguards are established to ensure data remain confidential and are compliant with the Privacy Act of 1974.[48] The Census must ensure that data are accessed only by those who have a statistical need for the data, as shown in the Confidential Information Protection and Statistical Efficiency Act (CIPSEA) Implementation Guide.[49]

3. Program and statistical agencies use interagency agreements or other similar tools to document terms and conditions governing data access and use when program agencies provide data that are not publicly available to the Census.

4. State agencies extract the required data elements and use appropriate encryption techniques to send data. Identifiable information should be provided only if the need cannot be met by relying on non-identifiable information.

5. After the Census receives and decrypts the data, overlapping unique identifiers from both data

sets are linked using probabilistic matching software. To append unique and consistent linkage identifiers, referred to as Protected Identification Keys, personal identifiers are compared to data in a reference file constructed by the Social Security Administration numerical identification file and other federal agency administrative data, using different combinations of social security number (SSN), full name, full date of birth, and address.

6. Identifiable data elements are eliminated when no longer required.

Even without an overarching data-sharing mandate between states and the Census, some state agencies already share their data with external partners voluntarily or with federal regulators when it is required to do so. In any new arrangement with the Census, however, attention must be paid to data privacy laws, data set quality, and data security and usage. A mandate from Congress and a clarification of current data-sharing laws and statutes will naturally be immensely helpful in this area. States can also be incentivized by the statistical product of the sharing arrangement and curating assistance from the Census, but they need the legal cover to do so. Standardization on data collection across state agencies must also be encouraged.

Conclusion

To advance the cause of data as public good, we must make progress on two fronts. To make the operations of government more effective, we need to make it easier for state and local agencies to share their information in real time—so that more places can do what Allegheny County does to address the problems of its most troubled citizens. To improve our ability to see how we are doing as a nation, we also need to make much faster progress on allowing the statistical agencies of the US to provide more accurate reports on the real condition of Americans.

To do that, Congress and other leaders must follow the CEP's recommendations and establish a service for data aggregation with a strong focus on information security, which can parlay public-private partnerships into new research and technology, and urge the Office of Management and Budget to make information available and searchable. Furthermore, they should begin the process of encouraging states and localities to share data, a crucial step toward making data sources the public resource they should be.

To achieve the vision of data as a public good, some localities may simply need help. Legal, administrative, data management and curation, IT, and security infrastructures are required to carry out data-sharing activities effectively and securely, and the legal and financial constraints facing agencies limit their ability to carry out these tasks. Statistical agencies may become invaluable sources to the heads of departments and other agencies in these efforts, calling on their own experiences and helping identify the benefits to department leadership.

Implementing these recommendations regarding privacy protection, security of IT systems, standardization, governance, and cooperation with the Census Bureau can help unlock the trove of information state and local agencies hold and in the process greatly improve outcomes for all Americans.

Appendix A

Privacy Guidance and Supports

Documents crucial to preventing security breaches would include the following.

Federal Privacy Protections. There would need to be definitive, user-friendly guidance on primary federal privacy protections, including the Family Educational Rights and Privacy Act of 1974 and the Health Insurance Portability and Accountability Act of 1996.

State and Local Laws on Privacy. Similarly, there would need to be user-friendly guidance on typical state and local laws on privacy, as well as word searches and research strategies to guide users to find these laws.

Spheres of Privacy. Steps should be provided to understand the different spheres of privacy: open public, provisioned public, and provisioned secure.

Some personal data may be shared openly within a single legal entity. That may be as large as a governmental jurisdiction (city, county, or state) or as small as a program if the program has an independent legal status. In general, this is widely misunderstood, and conservative misunderstandings cause shareable information to be hoarded.[50] Therefore, the term "legal entity" needs to be defined.

While personal information may be shared within a legal entity, portions of those data are privacy protected and may not be shared without a legally permissible justification. Clarity on this distinction would open up shareable information more widely.

Support on How to Effectively Share Data. Protections and limitations must be in place for sharing data outside a legal entity or sharing private information.

Guidance on steps to create data-sharing agreements between legal entities is required.[51] Additionally, guidance on circumstances in which privacy-protected information may be shared and when the information may be shared in only a de-identified fashion (i.e., public data) is necessary.

Considerations should be provided for when public data should be open versus provisioned versus not shared. Our bias is in favor of maximizing access to data that may be public. Some thoughts on how to navigate to this new reality over time would be useful, and there may be circumstances under which the government has a public interest in not releasing public data.[52]

Interagency Agreements. Once there is clarity about what can and cannot be shared and in what form, localities need a set of tools to formalize those agreements and understandings. These are embodied in standard data-sharing MOUs. Developing a standard template (or more likely identifying an existing one) would be helpful.

Data-sharing agreements differ in form, depending on whether they are for:

- Ongoing client-specific data sharing,
- Ongoing provisioned public data sharing,
- Open data sharing, or
- Time-limited or research-question-specific data sharing.

Client Waiver of Confidentiality. All this is obviated when clients waive confidentiality. This is a convenient approach that is widely used. The risks are that clients blindly sign waivers without understanding what they are signing or the implications. Therefore, guidance on the ethical use of client waivers would be valuable.

This would include a standard waiver form. Ideally the waiver would be a blanket waiver, for current and future services. Alternatively each point of service could require an independent waiver. It would also include guidance on how to advise clients about the waiver's meaning and content, before securing their signature.

Data Security Approaches

Secure data sharing requires using state-of-the-art technology solutions that have vastly improved data protection at a fraction of the cost. Localities need standardized solutions and guidance on this, enabling them to become better purchasers of the solutions, and more low- or no-cost shared environments, to avoid starting from scratch.

Some common approaches that would benefit from plain-language guidance include:

- **Encryption.** Integrated data systems can encrypt data that are publicly facing, which puts queries that access personal identifying information behind security protections, allowing tailored questions to be queried without revealing personal information in the response. This uses an approach referred to as a one-way hash.

- **Expert Determination Method.** This method uses statistical and scientific principles to render information not individually identifiable.

- **Entity Resolution.** Entity resolution is a process that connects files on a particular client across programs, creating a common client indicator. This is the basis for case-management systems that have carefully guarded, provisioned access but that are also used to create de-identified, population-level, multisystem reports.

- **De-Identification.** All information that could reasonably be used to identify an individual (e.g., name, address, and SSN) has been removed or replaced before sharing.[53]

- **Provisioning.** Technology solutions can create tiers of access to a common data set. Without authorization through provisioning, users can get only public-access open data. A user log-in may be nonetheless desired to track system use. Access to deeper levels of detail within the data is created through special user accounts, referred to as provisioning users.

- **Storage.** Solutions appropriate for sensitive government data storage are being developed and should be understood.

Data Curation

The variability in local definitions and collection methods means data sharing within jurisdictions and beyond local boundaries requires identifying and resolving discrepancies in the data themselves. Work must be done to advance appreciation of this challenge and offer tools to resolve it. Ultimately, the federal government or strong national intermediaries should fill this void and move local data systems to standard approaches in each field.

Several resources will facilitate standard data definitions:

1. The Food and Drug Administration has endorsed the SAS Institute as a resource.[54]
2. The NIEM effort should be reviewed for lessons learned and potential components for use.
3. The New York City Accelerator should be reviewed as a standard data model.
4. The US Department of Education Investing in Innovation grants to encourage standardization could be a useful model.

Until there is standardization in the field, and in recognition that perfection is unlikely, tools that identify and smooth out discrepancies will be needed and must be understood. The standard ETL approach to architecture would be helpful with this.

Table A1. Examples of Integrated Local Data Systems

Location	System	Organization
Allegheny County, Pennsylvania	Allegheny County Data Warehouse	Allegheny County Department of Human Services
Mecklenburg County, North Carolina	Institute for Social Capital Community Database	Institute for Social Capital Inc., University of North Carolina at Charlotte
Florida	Policy and Services Research and Data Center	Department of Mental Health and Policy, University of South Florida
Illinois	Integrated Database on Children and Family Programs	Chapin Hall, University of Chicago
Cuyahoga County, Ohio	Childhood Integrated Longitudinal Data System	Center on Urban Poverty and Community Development, Case Western Reserve University
Los Angeles County, California	Enterprise Linkages Project	Los Angeles County (Executive Office and Department Public Social Services)
Los Angeles and State of California	Children's Data Network School of Social Work	University of Southern California
New Jersey	Integrated Population Health Data Project	Center for State Health Policy, Rutgers University
New York City	Center for Innovation through Data Intelligence	Office of the Deputy Mayor for Health and Human Services
Rhode Island	DataSpark	The Providence Plan
San Mateo, Santa Cruz, and Santa Clara Counties	Silicon Valley Regional Data Trust	University of California, Santa Cruz
South Carolina	South Carolina Integrated Data Warehouse	South Carolina Revenue and Fiscal Affairs Office
Washington State	Department of Social and Health Services Integrated Client Database	Washington State Department of Social and Health Services, Research and Analysis Division

Source: Authors.

Governance

The basis for making advances in open data sharing is trust and transparency. A clear governance structure is crucial to both of these. A charter and a data breach enforcement policy are both needed to establish governance structure.

Charter. A standard charter provides the foundation's shared vision for the work being done and sets forth the basic framework for action. It is a grounding document and perhaps the most fundamental action leaders of local or state agencies can take toward establishing and achieving concrete goals.

Vision, Mission, and Principles. A clear statement of the values that drive the work, and the end game, can provide a beacon of focus that will help participants see their shared goals and not be defined by their differences.

Membership, Committees, Tasks, and Roles. Laying out roles and responsibilities offers clarity in what can be a confusing tangle of partners. Typical organizational units may include an executive board, a data sourcing committee, a data utilization committee, and a research and analytics advisory group.

Data Breach Enforcement Policy. Crucial to effective data security is a data breach policy that includes guidance on notification if a breach occurs and clear standards of enforcement when the liable party is identified—swift and certain is the recommendation. For an example of best practices on this topic, see the 2015 Department of Justice guidance on best practices for victim response and reporting of cybersecurity incidents.[55]

Federal Laws Relevant to Privacy Protection and Data Stewardship

The Privacy Act of 1974 established several requirements pertaining to records that are maintained in a "system of records," as defined in the statute.[56] The act generally prohibits agencies from disclosing individuals' records without their prior written consent, but it provides exceptions for matching data records if there are "matches performed to produce aggregate statistical data without any personal identifiers" and "matches performed to support any research or statistical project, the specific data of which may not be used to make decisions concerning the rights, benefits, or privileges of specific individuals."

The E-Government Act of 2002 requires agencies to conduct "an analysis of how [personally identifiable] information is handled: (i) to ensure handling conforms to applicable legal, regulatory, and policy requirements regarding privacy, (ii) to determine the risks and effects of collecting, maintaining and disseminating information in identifiable form in an electronic information system, and (iii) to examine and evaluate protections and alternative processes for handling information to mitigate potential privacy risks."[57]

The Federal Information Security Modernization Act of 2014 amended Title III of the E-Government Act to require that "each Federal agency shall develop, document, and implement an agency-wide information security program to provide information security for the information and information systems that support the operations and assets of the agency, including those provided or managed by another agency, contractor, or other source."[58]

Title V of the E-Government Act (CIPSEA) established uniform confidentiality protections for information acquired by agencies, including principal statistical agencies, so that "the description, estimation, or analysis of the characteristics of groups [is done without] identifying the individuals or organizations that comprise such groups."[59]

Title 42 of the US Code protects the information states provide to the HHS's Federal Parent Locator System, which includes the National Directory of New Hires, permitting the disclosure of this information to specific agencies for limited purposes.

FERPA protects the privacy of student education records, limiting both access and use.

The 1996 Health Insurance Portability and Accountability Act and the Health Information Technology for

Economic and Clinical Health Act address the use and disclosure of protected health information by specified "covered entities."[60]

Title 13 of the United States Code provides the Census with the authority to acquire and use administrative records collected by other federal agencies; state, tribal, or local governments; and private organizations.[61] The Census Bureau is obligated to protect the confidentiality of these records just as it protects the information it gathers directly from individuals and businesses. These data can be used only for statistical purposes, no individual or business may be identified in a published report, and individual records may be accessed only by sworn officers or employees of the Census Bureau.

Examples of Administrative Data Used for Research and Policy Evaluation

Administrative data sharing has already been shown to yield valuable information when employed by social scientists and select government agencies:

- Combining data from one state's Departments of Mental Health, Social Services, Public Safety, Corrections, and the Division of Court Supported Services, researchers examined how justice involvement affected behavioral health treatment costs.

- Researchers used administrative records to map the resources available to support young children through public and nonprofit providers in a major metropolitan area. They estimated that the city's efforts to coordinate family services saved $3 in future health expenditures for every $1 invested.

- The Center for Medicare and Medicaid Innovation uses administrative health data to conduct rapid cycle evaluation to identify effective approaches to reducing expenditures without degrading quality of care.

- In New York City, researchers designed an experiment to test behavioral "nudges" in SNAP applications to encourage complete and accurate reporting by randomly assigning applicants to one of four redesigned online applications. Results using administrative data were used to inform improvements to the application process.

- Researchers compared program impacts on math and reading achievement using both aggregate school-level and individual student-level data. For some research questions, aggregate and individual data produced similar results, meaning that researchers should carefully consider whether aggregate data could be sufficient, as these data are often more readily available than individual-level data.[62]

Relevant Data Tools

There are three primary systems that agencies use to control access with researchers.

Online data query systems are analysis tools that allow the public to examine restricted-use data dynamically, creating tables, rates, and models.

On-site access allows eligible researchers the opportunity to gain access to restricted-use data for select research projects at the agency. Generally, interested researchers submit a project proposal that, if approved, allows them to conduct work with on-site microdata at little to no cost to them or their institution or organization.

Data enclaves are secure environments, on site or virtual, in which qualified researchers may access restricted-use microdata for statistical purposes.

About the Authors

Robert Doar is the Morgridge Fellow in Poverty Studies at the American Enterprise Institute (AEI), where his research focuses on how improved federal and state antipoverty policies and safety-net programs can reduce poverty, connect individuals to work, strengthen families, and increase opportunities for low-income Americans and their children. He is also a senior fellow at Results for America. While at AEI, Mr. Doar has served as a co-chair of the National Commission on Hunger and as a lead member of the AEI-Brookings Working Group on Poverty and Opportunity, which published the report titled *Opportunity, Responsibility, and Security: A Consensus Plan for Reducing Poverty and Restoring the American Dream*. He is also the editor of *A Safety Net That Works: Improving Federal Programs for Low-Income Americans* (AEI Press, 2017), in which experts discuss major federal public assistance programs and offer proposals for reform. Before joining AEI, he worked for Mayor Michael Bloomberg as commissioner of New York City's Human Resources Administration. While administering 12 public assistance programs, including cash welfare, food assistance, public health insurance, child support enforcement services, and others, he oversaw a 25 percent reduction in the city's welfare caseload. Before joining the Bloomberg administration, he was commissioner of social services for the state of New York, where he helped to make the state a model for the implementation of welfare reform. Mr. Doar has testified numerous times before Congress, and his writing has appeared in the *Wall Street Journal*, *USA Today*, the *Hill*, and *National Review*, among other publications. Mr. Doar has a bachelor's degree in history from Princeton University.

Linda Gibbs is a principal with Bloomberg Associates and a senior fellow at Results for America. She served as New York City deputy mayor of health and human services from 2005 to 2013. Supervising the city's human service, public health, and social justice agencies, she spearheaded major initiatives on poverty alleviation, juvenile justice reform, and obesity reduction. She helped shape "Age Friendly NYC," a blueprint for enhancing livability for older New Yorkers, and "Young Men's Initiative," addressing race-based disparities facing black and Latino young men in the areas of health, education, employment training, and the justice system. Ms. Gibbs also improved the use of data and technology in human service management and spearheaded efforts to improve contract effectiveness and create evidence-based program development. Before her appointment as deputy mayor, Ms. Gibbs was commissioner of the New York City Department of Homeless Services and held senior positions with the Administration for Children's Services and the Office of Management and Budget. She is a graduate of SUNY Buffalo School of Law.

Notes

1. Terence Lutes, *Data-Driven Government: Challenges and a Path Forward*, IBM Analytics, April 2015, http://www-01.ibm.com/common/ssi/cgi-bin/ssialias?htmlfid=GQW03008USEN.

2. City of San Diego City Council, "Open Data Policy," January 1, 2014, https://www.sandiego.gov/sites/default/files/legacy/opengov/pdf/agendas/draftopendatapolicy.pdf; and City of San Diego, "Mayor Faulconer Makes Dozens of Public Data Sets Available Online to Increase Transparency: San Diego's New Open Data Portal Automatically Provides Up-to-Date Data for the Public, Software Developers to Utilize," press release, July 6, 2016, https://www.sandiego.gov/mayor/news/releases/mayor-faulconer-makes-dozens-public-data-sets-available-online-increase-transparency.

3. The White House, Office of the Press Secretary, "Freedom of Information Act," January 21, 2009, http://nsarchive2.gwu.edu//news/20090121/2009_FOIA_memo.pdf.

4. Evidence-Based Policymaking Commission Act of 2016, Pub. L. No. 114-140, https://www.cep.gov/content/dam/cep/about/public-law-114-140.pdf.

5. Commission on Evidence-Based Policymaking, *The Promise of Evidence-Based Policymaking*, September 2017, https://www.cep.gov/content/dam/cep/report/cep-final-report.pdf.

6. Bruce D. Meyer, Wallace K. C. Mok, and James X. Sullivan, "Household Surveys in Crisis," *Journal of Economic Perspectives* 29, no. 4 (Fall 2015): 199–226, http://www.nber.org/papers/w21399.

7. Erika M. Kitzmiller, *Allegheny County's Data Warehouse: Leveraging Data to Enhance Human Service Programs and Policies,* Actionable Intelligence for Social Policy, May 2014, https://www.aisp.upenn.edu/wp-content/uploads/2015/08/AlleghenyCounty-_Case-Study.pdf.

8. Allegheny County, "Predictive Risk Modeling in Child Welfare in Allegheny County: The Allegheny Family Screening Tool," http://www.alleghenycounty.us/Human-Services/News-Events/Accomplishments/Allegheny-Family-Screening-Tool.aspx.

9. Exec. Order No. 114, New York City (March 18, 2008), https://drive.google.com/file/d/0B1dHL-KJpuw5bW91QlJoeDJlZVk/view.

10. Mayor's Office for Economic Opportunity, NYC Opportunity, "Worker Connect," http://www1.nyc.gov/site/opportunity/portfolio/worker-connect.page.

11. City of New York, "Inter-Agency Data Exchange Agreement," November 2010, https://drive.google.com/file/d/0B1dHL-KJpuw5aU5ISzIxUUVyMm8/view; NYC Connect, "Terms of Use for Confidential Data," https://drive.google.com/file/d/0B1dHL-KJpuw5OW5EcklUMERJcGc/view; and NYC Connect, "Worker Connect Use Case Template," https://drive.google.com/file/d/0B1dHL-KJpuw5X1JyektsTmJYVTA/view.

12. Howard S. Bloom and Rebecca Unterman, "Can Small High Schools of Choice Improve Educational Prospects for Disadvantaged Students?," *Journal of Policy Analysis and Management* 33, no. 2 (March 2, 2014): 290–319, http://onlinelibrary.wiley.com/doi/10.1002/pam.21748/abstract.

13. Mark Prell et al., "Profiles in Success of Statistical Uses of Administrative Data," April 2009, https://www.bls.gov/osmr/fcsm.pdf.

14. Center for Innovation Through Data Intelligence, "CIDI Data Hive Protocol," July 3, 2012, https://drive.google.com/file/d/0B1dHL-KJpuw5UVVLSVhGOUJXRGs/view.

15. US Department of Education, "Family Education Rights and Privacy Act (FERPA)," June 26, 2015, https://www2.ed.gov/policy/gen/guid/fpco/ferpa/index.html.

16. Office of Management and Budget, "Barriers to Using Administrative Data for Evidence-Building," July 15, 2016, https://obamawhitehouse.archives.gov/sites/default/files/omb/mgmt-gpra/barriers_to_using_administrative_data_for_evidence_building.pdf.

17. National Conference of State Legislatures, "Security Breach Notification Laws," April 12, 2017, http://www.ncsl.org/research/

telecommunications-and-information-technology/security-breach-notification-laws.aspx.

18. SANS Institute, "Data Breach Response Policy," https://www.sans.org/security-resources/policies/general/pdf/data-breach-response.

19. Exec. Order 2016-05, City of Seattle (June 9, 2016), https://drive.google.com/file/d/0B0kI_w5VZQDFZ214a1B3UElvem8/view.

20. City of Seattle Human Services Department, City of Seattle Office of Housing, King County Department of Community and Health Services, and United Way of King County, "Memorandum of Understanding: Implementation of Revised System Wide Performance Targets and Minimum Standards," September 6, 2016, https://drive.google.com/file/d/0B0kI_w5VZQDFc2VpZUVyU0kxck0/view.

21. Washington State Legislature, "Washington Homeless Client Management Information System," RCW 43.185C.180, http://apps.leg.wa.gov/Rcw/default.aspx?cite=43.185C.180.

22. National Information Exchange Model, https://www.niem.gov/.

23. From the website: "If I say 'vessel' and you say 'boat', and he says 'ship' and she says 'conveyance', we may mean the same thing, but we have no way to tell our computer systems to treat the words as having the same meaning. Until we do, we'll all have separate facts about the same world—pieces of the big puzzle—but no common understanding or way to connect them. . . . This is the idea behind NIEM. It lets your system and my system speak—even if they've never spoken before—by ensuring that information carries the same consistent meaning across various communities." National Information Exchange Model, "About NIEM," https://www.niem.gov/about-niem.

24. NYC HHS Accelerator, "Client and Community Services (CCS) Catalog," http://www1.nyc.gov/assets/hhsaccelerator/downloads/pdf/services_catalog.pdf.

25. Montgomery County Open Government Act, No. 23-12, http://montgomerycountymd.gov/open/Resources/Files/SignedOpenDataBill_23-12.pdf.

26. Niels Bantilan, "Themis-ml: A Fairness-Aware Machine Learning Interface for End-to-End Discrimination Discovery and Mitigation" (paper, Bloomberg Data for Good Exchange Conference, New York, NY, September 24, 2017).

27. Equity Intelligence Platform, "Equity Intelligence Platform: Project Charter," 2017, https://drive.google.com/file/d/0B1dHL-KJpuw5VVRiMUxtMU1aN2s/view.

28. Equity Intelligence Platform, "Equity Intelligence Platform: Oakland Project Charter," 2017, https://drive.google.com/file/d/0B1dHL-KJpuw5Z1A3TGRuRUJuUUE/view?usp=sharing.

29. Raj Chetty et al., "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility" (working paper, National Bureau of Economic Research, Cambridge, MA, January 2014), http://www.equality-of-opportunity.org/assets/documents/mobility_trends.pdf.

30. Raj Chetty and Nathaniel Hendren, "The Impacts of Neighborhoods on Intergenerational Mobility I: Childhood Exposure Effects," May 2017, http://www.equality-of-opportunity.org/assets/documents/movers_paper1.pdf.

31. Raj Chetty et al., "Mobility Report Cards: The Role of Colleges in Intergenerational Mobility," July 2017, http://www.equality-of-opportunity.org/papers/coll_mrc_paper.pdf.

32. Nicholas Eberstadt et al., "In Order That They Might Rest Their Arguments on Facts: The Vital Role of Government-Collected Data," Brookings Institution and American Enterprise Institute, March 2, 2017, https://www.brookings.edu/research/in-order-that-they-might-rest-their-arguments-on-facts-the-vital-role-of-government-collected-data/.

33. US Census Bureau, "Administrative Data Acquisitions," August 29, 2016, https://www.census.gov/about/adrm/linkage/technical-documentation/AR-Acqu.html; and US Census Bureau, "Data Ingest and Linkage," September 7, 2016, https://www.census.gov/about/adrm/linkage/technical-documentation/processing-de-identification.html.

34. US Census Bureau, "Survey of Income and Program Participation," https://www.census.gov/sipp/.

35. US Census Bureau, "American Community Survey (ACS)," https://www.census.gov/programs-surveys/acs/.

36. US Census Bureau, "Current Population Survey (CPS)," https://www.census.gov/programs-surveys/cps.html.

37. John L. Czajka and Amy Beyler, "Declining Response Rates in Federal Surveys: Trends and Implications," Mathematica Policy Research, June 15, 2016, https://aspe.hhs.gov/system/files/pdf/255531/Decliningresponserates.pdf.

38. Bruce D. Meyer and Nikolas Mittag, "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness and Holes in the Safety Net" (working paper, American Enterprise Institute, Washington, DC, August 26, 2015), https://www.aei.org/publication/using-linked-survey-and-administrative-data-to-better-measure-income-implications-for-poverty-program-effectiveness-and-holes-in-the-safety-net/.

39. Bruce D. Meyer, Wallace K. C. Mok, and James X. Sullivan, "Household Surveys in Crisis" (working paper, National Bureau of Economic Research, Cambridge, MA, July 2015), http://www.nber.org/papers/w21399.

40. Jim Cannon, "Memorandum for the President: S. 3435—Privacy Protection Study Commission," September 1, 1976, https://www.fordlibrarymuseum.gov/library/document/0055/1669480.pdf.

41. US Census Bureau, "Combining Data—A General Overview," April 24, 2017, https://www.census.gov/about/what/admin-data.html.

42. US Census Bureau, *Design and Methodology: American Community Survey*, April 2009, https://www.census.gov/content/dam/Census/library/publications/2010/acs/acs_design_methodology.pdf.

43. US Census Bureau, "Longitudinal Employer-Household Dynamics," https://lehd.ces.census.gov/.

44. US Census Bureau, "Census Longitudinal Infrastructure Project (CLIP)," https://census.gov/about/adrm/linkage/projects/clip.html.

45. US Census Bureau, "American Opportunity Study (AOS)," https://census.gov/about/adrm/linkage/projects/aos.html.

46. US Census Bureau, "Mortality Disparities in American Communities (MDAC)," https://census.gov/about/adrm/linkage/projects/mdac0.html.

47. Mark Prell, "Illuminating SNAP Performance Using the Power of Administrative Data," US Department of Agriculture, Economic Research Service, November 7, 2016, https://www.ers.usda.gov/amber-waves/2016/november/illuminating-snap-performance-using-the-power-of-administrative/.

48. The Privacy Act limits the collection, use, maintenance, and dissemination of information about individuals maintained by agencies but does permit intra-agency disclosure of data for statistical purposes if there is a valid "need to know." US Department of Justice, Office of Privacy and Civil Liberties, "Privacy Act of 1974," July 17, 2015, https://www.justice.gov/opcl/privacy-act-1974.

49. Office of Management and Budget, "Implementation Guidance for Title V of the E-Government Act, Confidential Information Protection and Statistical Efficiency Act of 2002 (CIPSEA); Notice," June 15, 2007, https://www.gpo.gov/fdsys/granule/FR-2007-06-15/E7-11542.

50. For instance, Allegheny County, Pennsylvania, has established that the county is the legal entity of relevance for sharing client information that is not confidentially protected at the individual level. Alleghany County Department of Human Services, "Memorandum of Understanding," https://www.alleghenycountyanalytics.us/wp-content/uploads/2016/06/Data-Sharing-Memorandum-of-Understanding-DHS-School-District.pdf.

51. The Sunlight Foundation has some useful guidance and resources. Emily Shaw, "Sharing Sensitive Data Within Government," Sunlight Foundation, February 11, 2015, https://sunlightfoundation.com/2015/02/11/sharing-sensitive-data-within-government/. For an analysis of the legal foundation for data sharing in Allegheny County, see John Petrila, "Report on Privacy and Security Issues in the Allegheny County Data Warehouse," Allegheny County Department of Human Services Data Warehouse, https://drive.google.com/file/d/0B1dHL-KJpuw5RUdBRmFJWTZ1TEE/view.

52. US Government Accountability Office, "Sustained and Coordinated Efforts Could Facilitate Data Sharing While Protecting Privacy," February 8, 2013, https://www.gao.gov/products/GAO-13-106.

53. See US Health and Human Services, "Guidance Regarding Methods for De-Identification of Protected Health Information in Accordance with the Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule," November 6, 2015, https://www.hhs.gov/hipaa/for-professionals/privacy/special-topics/de-identification/index.html.

54. US Food and Drug Administration, "Study Data Specifications," July 18, 2012, https://www.fda.gov/downloads/forindustry/datastandards/studydatastandards/ucm312964.pdf.

55. US Department of Justice, Computer Crime & Intellectual Property Section, Criminal Division, Cybersecurity Unit, "Best Practices for Victim Response and Reporting of Cyber Incidents," April 2015, https://www.justice.gov/sites/default/files/opa/speeches/

attachments/2015/04/29/criminal_division_guidance_on_best_practices_for_victim_response_and_reporting_cyber_incidents.pdf.

56. The term "record" means "any item, collection, or grouping of information about an individual that is maintained by an agency, including, but not limited to, his education, financial transactions, medical history, and criminal or employment history and that contains his name, or the identifying number, symbol, or other identifying particular assigned to the individual, such as a finger or voice print or a photograph." National Archives, "The Privacy Act of 1974," August 15, 2016, https://www.archives.gov/about/laws/privacy-act-1974.html.

57. E-Government Act of 2002, Pub. L. No. 107-347 § 208, https://www.gpo.gov/fdsys/pkg/PLAW-107publ347/pdf/PLAW-107publ347.pdf.

58. Federal Information Security Management Act of 2002, 44 USC § 3541.

59. Federal Register, "Statistical Policy Directive No. 1: Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units," December 2, 2014, 71614, https://www.federalregister.gov/documents/2014/12/02/2014-28326/statistical-policy-directive-no-1-fundamental-responsibilities-of-federal-statistical-agencies-and.

60. Office of Management and Budget, "Privacy and Confidentiality in the Use of Administrative and Survey Data," July 15, 2016, https://obamawhitehouse.archives.gov/sites/default/files/omb/mgmt-gpra/privacy_and_confidentiality_in_the_use_of_administrative_and_survey_data_0.pdf.

61. 13 USC § 6 (1954), https://www.gpo.gov/fdsys/pkg/USCODE-2007-title13/pdf/USCODE-2007-title13.pdf.

62. Mark Prell, "Profiles in Success of Statistical Uses of Administrative Data," April 2009, https://s3.amazonaws.com/sitesusa/wp-content/uploads/sites/242/2014/04/StatisticalUsesofARData.pdf.

Results for America is helping decision makers at all levels of government harness the power of evidence and data to solve great challenges. Our mission is to make investing in what works the new normal, so that when government policymakers make decisions, they start by seeking the best evidence and data available, then use what they find to get better results. We accomplish this goal by developing standards of excellence which highlight the government infrastructure necessary to be able to invest in what works, supporting policymakers committed to investing in what works, and enlisting champions committed to investing in what works.

American Enterprise Institute is a public policy think tank dedicated to defending human dignity, expanding human potential, and building a freer and safer world. The work of our scholars and staff advances ideas rooted in our belief in democracy, free enterprise, American strength and global leadership, solidarity with those at the periphery of our society, and a pluralistic, entrepreneurial culture. We are committed to making the intellectual, moral, and practical case for expanding freedom, increasing individual opportunity, and strengthening the free enterprise system in America and around the world.

Acknowledgment

This report has been produced with the generous support of the Kresge Foundation and the Nielsen Foundation. The report is an independent work product of Results for America and the American Enterprise Institute, and the views expressed are those of the authors and do not necessarily represent those of the funders.

The American Enterprise Institute (AEI) is a nonpartisan, nonprofit, 501(c)(3) educational organization and does not take institutional positions on any issues. The views expressed here are those of the author(s).