Advancing Integrated Data Systems by States and Local Governments: Expert Panel Recommendations Regarding Governance, Legal Agreements, Data Standards, and Technology and Data Security

Dennis Culhane, John Fantuzzo, Matthew Hill and TC Burnett

Abstract

Common challenges confront states and local governments seeking to develop systems for on-going data sharing among agencies responsible for public health, education, human services, housing, justice and workforce programs, sometimes referred to as "integrated data systems" or IDS. Four panels of experts in the development and implementation of data sharing solutions were convened over the last year as part of an effort to promote scalable IDS solutions. Experts in IDS governance, legal issues, data standards, and technology and data security have developed reports intended to establish best practice standards and guidelines for the field. Those reports are summarized here. The paper concludes with a discussion of how federal, state and foundation funders can expand capacity for evidence-based policymaking through the adoption of these IDS standards, shared technology solutions based on those standards, and a shared network governance and communications infrastructure.

I. Introduction

State and local governments spend over $2.5 trillion annually on education, health care and social safety net programs.[1] Yet in spite of these expenditures, most governments lack the data required to evaluate whether or not these programs actually work. In response to this situation, there is a growing push at the federal and state level to create a more responsive, innovative and transparent government. At the Federal level, House Speaker Paul Ryan and Senator Patty Murray crafted a bill in 2016, later signed into law by the President Obama, to form the federal "Commission on Evidenced-Based Policy." The goal of this Commission is to integrate data-driven decision making into the budgeting and legislative process. A key component of this new evidence agenda entails making better use of "administrative data." Administrative data consist of data that many federal and state programs gather for program administration, and regulatory or law enforcement purposes but infrequently use for research, evaluation, or performance measurement.[2] Various "open data" initiatives by state and local governments have already provided access to many large datasets to analysts and social innovators. However, given the strong legal protections afforded confidential data regarding tax payers, students, patients and social service clients, such data are not likely to be made available via these mechanisms. Integrated data systems (IDS) have emerged as an alternative approach by states and local governments to create more streamlined access to such data, but with strict data protections not usually associated with other open data products. By harnessing the power of these data, the hope is that government can encourage innovation in the provision of

[1] See http://www.urban.org/policy-centers/cross-center-initiatives/state-local-finance-initiative/projects/state-and-local-backgrounders/state-and-local-expenditures.

[2] See https://www.whitehouse.gov/sites/default/files/omb/budget/fy2016/assets/ap_7_evidence.pdf.

education, health and human services, reducing costs while improving the quality of government-funded programs. The linkage of administrative data in an IDS makes these data all the more powerful, providing a multi-agency perspective on how individuals utilize services, both sequentially and in combination, and to examine the impact of these programs on a broad array of domains.

IDS Governance Models

Through the Actionable Intelligence for Social Policy Network of mature IDS around the country, we have studied the 13 jurisdictions that have successfully deployed IDS for policy analysis and program evaluation purposes in the U.S. Taken together, these jurisdictions comprise 26% of the U.S. population. Based on this sites, we have observed three types of IDS governance models: state/county executive led IDS; state/county agency led IDS; and university-led IDS. Executive-based models tend to be located within a non-partisan government department. They enjoy strong government support, and have often weathered several political and economic transitions. Agency-based models generally reside in a state or county's Department of Health and Human Services. Most were originally developed to complement a department's case management practices, and as their utility for policy and program decision-making was discovered, executive staff within that department initiated processes to facilitate data integration for research purposes, and have expanded over time to include data sharing partnerships with other agencies. University-based models differ from the two governmental models in that they generally do not have a formal governance process. Because of this, a university-model's research agenda is generally formed based on a combination of partner-agency need, researcher-interest, and available funding.

These three different IDS governance models demonstrate important differences. For instance, the government models—both at the executive and agency level—tend to be better funded than university-based models. The primary audience for these IDS also varies, depending on model type. Both types of government-based models mainly focus on meeting the needs of government policy makers. This in turn influences the size and scope of the IDS. So, for instance, Washington State's IDS has 72 full-time employees and South Carolina's IDS has 34 full-time employees. This is in stark contrast to the university-based models, which are smaller (with an average of 5-10 full-time employees), tend to be exclusively grant- and contract-funded, and focus on producing research, though many also have very strong partnerships with agencies and devote much of their time and energy to meet agency demands for research and evaluation studies.

Through our study of IDS sites in the AISP network, one factor became immediately clear—regardless of organizational model, all sites agreed that data holders must be at the center of the IDS' purpose in order for it to operate in a sustainable, ongoing manner. This focus is paramount as it provides incentive for the data holder to participate in and support the IDS. Without data holder participation, the IDS has no ability to function, much less exist. Therefore, the purpose of the IDS must align with the data holder's needs. This is in strict contrast to many European-based IDS, which are directly set up to handle a broad array of population studies. In the U.S., data holders generally do not support research for research's sake as a primary value; rather, the research aims must seek to address key areas of policy and programmatic need.

While other significant efforts to link administrative data are ongoing, for instance, at many university research centers, they often do not encompass a broad policy focus similar to the aforementioned AISP sites. In particular, such centers are driven by academic research priorities, complementary to the needs of specific public agencies, but are often intramural in their focus, rather than interfacing directly with a broad array of government and evaluation partners.

Additionally, states and local governments are also continuously integrating administrative data for operational purposes, and a number of IT vendors operate business divisions specifically to serve these aims. For instance, every state shares administrative data across agencies for eligibility and income verification purposes. This routinely takes place for determination of eligibility for things like school lunch programs, nutrition assistance, and housing assistance. States therefore have ample experience integrating and using integrated data for program, regulatory, and legal purposes. The next step in the evidence-based policy agenda is helping states to leverage these data for reflective practice, and the creation of a more responsive government.

The AISP Network has engaged in self-study for nearly ten years, identifying the best practices and common challenges faced by state/county and university-led IDS systems. Based on that experience, in this paper we reflect on that learning to discuss the future potential for the IDS field. The insights expressed here will also draw on four expert panels convened in 2016, funded by the Linda and John Arnold Foundation, to address the most common challenges and barriers faced by sites seeking to develop an IDS. These include: 1) legal challenges related to the integration of protected, person-level data; 2) governance challenges arising from the right mix of stakeholders, and policies and

procedures required to operate an ethical and effective IDS; 3) data challenges, including data quality and the data elements most likely to be available with sufficient reliability, and with potential scalability across the country; and 4) technology challenges associated with the cost and effort of building an IDS, and ensuring the security of its data.

II. IDS Challenges and Best Practices

A. Legal Issues

The biggest legal barrier that states encounter in seeking to integrate their administrative data is the *perception* on the part of data owners, agency leaders, and elected officials that "this is not legal." This idea is buttressed by the notion that there are federal statues (e.g. FERPA, HIPPA, 42CFR Pt.2, Privacy Act) that prohibit the sharing of administrative data, and that state statutes further restrict such activities. This interpretation is predicated on a range of factors. These include: misunderstandings of actual privacy laws and policies; inconsistencies between federal, state and local privacy laws; and concerns about potential liability and lawsuits, coupled with a bias within government agencies against sharing data. When taken together, these misperceptions combine to produce an artificial – and surmountable -- barrier to data sharing and integration.

Our expert panel has concluded that current federal law is not a barrier to such integration. Exemptions and exceptions exist in each of the prevailing federal laws that permit the sharing of data for evaluation, planning, and audit purposes, all of which are routine business activities within these agencies. For instance, the federal Privacy Act permits the disclosure of personally identifiable information (PII) (e.g. education, medical, employment) for "routine use" (Petrila 2015:43). Federal agencies have interpreted this

exemption to include the inter-agency exchange of data to assist them in verifying eligible recipients of federal program, the evaluation of programs, and the detection of fraud and waste (cf. Coles 1990). Similarly, the Health Insurance Portability and Accountability Act (HIPPA), while protecting personal health information (PHI) in the control of "covered entities" (e.g. health plans, health care providers, health care clearinghouses) permits the sharing of such data for research purposes in "deidentified form" or alternatively in "limited data sets" with IRB approval and a data use agreement (Peters 2011). Finally, while the Family Educational Rights and Privacy Act (FERPA) protects PII from education records from unauthorized disclosure, it makes exceptions for school officials, "studies" aimed at improving instruction, and "audits and evaluations" of federal or state supported education programs. More recent guidance further permits the designation of an IDS and its staff as an "educational authority" if the proposed data linkages are for a legitimate educational purpose—i.e. the data are used on behalf of a school district to improve instruction, or to evaluate a federal or state supported education program (Hawes 2015).

The forthcoming AISP expert panel report will address these common misconceptions, and demonstrate how the various data sharing provisions can be accommodated through two key legal agreements: an interagency data sharing Memorandum of Understanding (MOU) and the end user Data License Agreement (DLA). The first of these agreements, the MOU, is a foundational agreement between a lead IDS agency (i.e. the entity administering the IDS), the various data contributors (i.e. government agencies or sub-agencies), and data licensees (i.e. researchers or program evaluators). It defines the core features of the IDS structure, and the respective legal rights and responsibilities of the participating parties. IDS lead agencies can either draft separate MOUs with each participating data provider, or

a single, 'enterprise' MOU (e-MOU) with all of the data contributors.[3] The DLA, by contrast, is a legal agreement that defines the terms and conditions under which researchers or evaluators may gain temporary access to a limited set of data in an IDS for research, evaluation, or audit purposes. It contains many of the same requirements regarding the legal use and security protections for the data as the MOU, as well as additional provisions covering the specific data elements involved, the handling of the data, the right of agencies to review results and work products, and the duration of the license.

B. Data Governance

While the MOU and DLA provide the necessary legal framework required to develop an IDS, they can also serve as the foundational process for the establishment of an IDS governance system overall. They can do so by bringing diverse stakeholders (e.g. data owners, funding sources, public agency leadership, data users, technical experts) together, enabling them to define roles and responsibilities, create joint goals, and establish trust by working together to address shared concerns. The agreements also help stimulate broader conversations about the reasons for creating the IDS. These include: its ideal organizational structure, the data it should include, the policies and procedures governing use of the data, and the research priorities that it should address. In this manner, the legal agreements create the legal authority which gives rise to, and is ultimately incorporated in, the governance process, which goes further by describing the principles and operating rules that govern the IDS.

In a forthcoming report, the AISP Policies & Procedures expert panel has concluded that given the value proposition of integrated administrative data to improve services to the

[3] The Commonwealth of Virigina has developed such an e-MOU facilitating the streamlined sharing of data among participating government agencies. See https://ehhr.virginia.gov/media/5884/emou_one_pager.pdf.

public while providing sufficient protections to the security and confidentiality of the data, there should be an assumption that these data represent a public good more than a security risk, and that they should therefore be shared for research and program evaluation purposes whenever possible and unless specifically restricted by law. Sharing such data can provide government with fundamental knowledge essential to program planning, targeting, implementation, outcomes measurement and cost effectiveness analysis, thereby broadly serving the public interest. The default position should therefore be that administrative data should be made available and even incentivized for these important uses, unless there is a valid reason for their exclusion. Data sharing should be understood as ultimately about supporting population health, civic participation, educational achievement, and economic opportunity, and data governance seen as the means of creating partnerships in support of those goals. The ethic that guides the work of data transparency then is one of promoting improved services for the public good.

A corollary of this ethic of data transparency is that IDS "shall do no harm" by building in safeguards to minimize the potential risk of re-identification of personally identifiable information. For an IDS, such potential harm includes damage to the individuals whose data is being used by the IDS, as well as to the public agencies that serve those individuals (Fantuzzo et al. 2015:28). While we discuss the risks of re-identification in the Technology and Data Security section of this report, we note here that IDS must embed confidentiality and privacy protections in the data governance process to ensure that there is no risk to the general public.

In order to promote beneficence with regard to this work, there has to be a transparent oversight and review process that includes all of the key stakeholders involved in the

development and ongoing operation of the IDS. At the forefront of these efforts is the IDS' *Executive Board*, which is charged with the supervision and direction of the IDS, and ensuring the ethical use of its data. It defines the agenda of the IDS by regularly convening the relevant stakeholders (researchers, practitioners, the general public) to provide input on critical issues to the community, selecting research priorities that align with the IDS' principles and the public good, and ensuring the feasibility of projects based on available data and resources. The *research community* must similarly ensure that research proposals support the priorities of the agencies that comprise the IDS, and undergo an IRB review process to ensure that they adequately protect the rights of human subjects. Moreover, the *IDS staff* must play a role by leading the approval and review process for all data and research requests. They must also convene researchers and practitioners to review and translate key findings, discuss policy and practice implications, and raise new questions for further research. Finally, the *data sub-committee* must review the final research products prior to public dissemination to ensure compliance with the initial data request and broader agency priorities, while soliciting feedback from the data providers whose data were used in the research. In sum, this entire stakeholder review process depends on the establishment of an IDS with agreed upon processes and protocols for bringing data together, and assuring that the use of these data aligns with agency priorities.

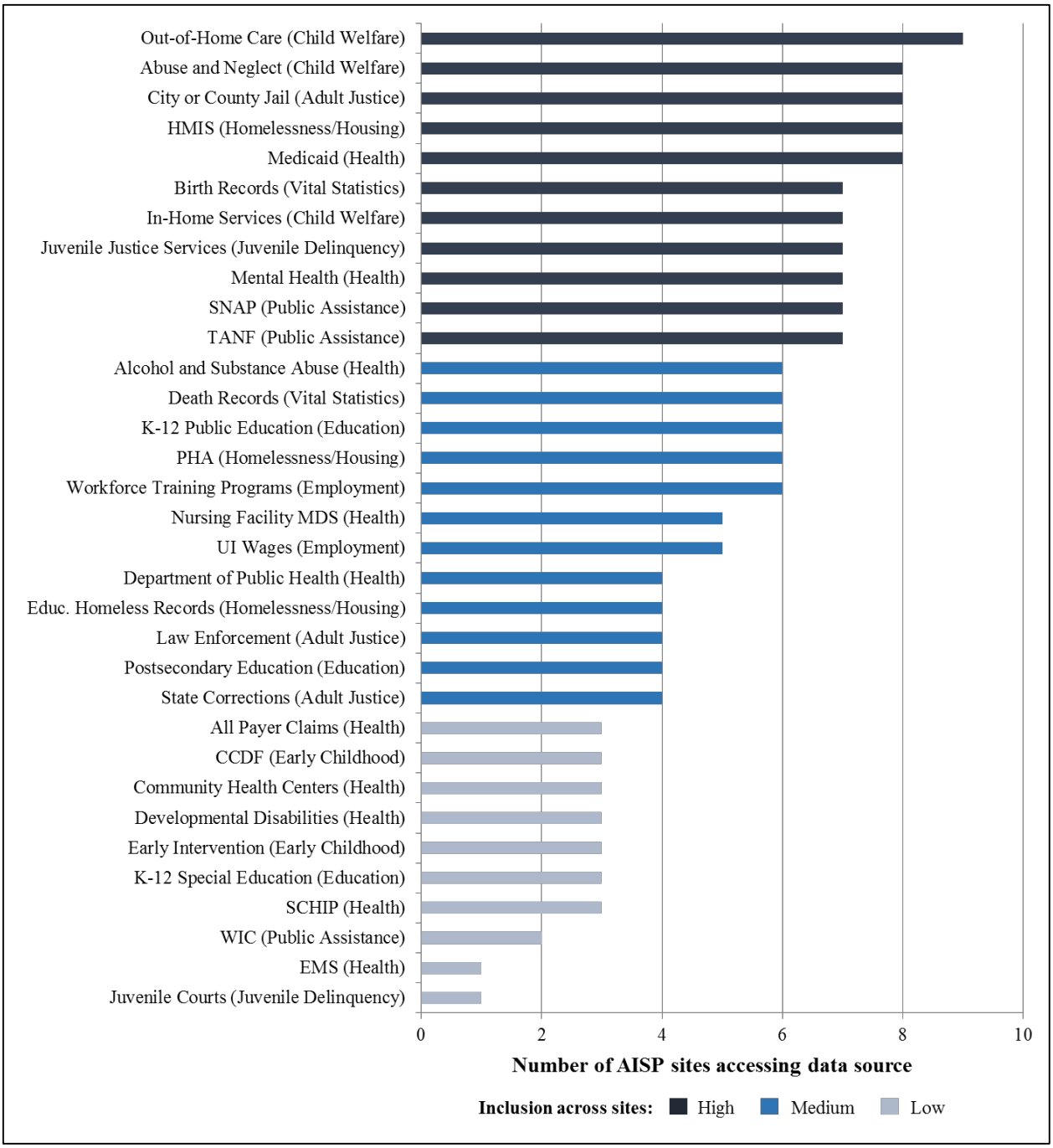
C. Minimum Data Sets and Data Quality Standards

One major potential benefit from multiple IDS installations across the country is to work toward a nationally scalable and replicable approach, with a common data model to the extent that it is feasible. AISP convened an expert panel to examine what data sources would exist in nearly every state/country in the U.S., and what the 'most reliable' data

elements would be within these data sources. Towards this end, the panel focused specifically on data elements that would be subject an auditing process, thereby helping to ensure their accuracy. In preparing its report, the expert panel recommended that the data elements selected for inclusion in the IDS should be organized from a life course perspective. In this view, public institutions play an important role in helping individuals navigate key life transitions, and IDS research and policy questions should focus on the efficacy of these institutions in facilitating this process. Key transitions that organize the data sources that an IDS should incorporate include: birth and infancy; early childhood; school-age children; transitions to adulthood; adults, workers and parenthood; old age and death. In addition to data sources organized around these life transitions, the data in the IDS should also include contextual information, specifically temporal and place-based identifiers, providing background on social organizations and their interaction with individuals in a specific time frame and physical location.

The panel identified ten domains for initial consideration for inclusion in a nationally scalable IDS. These include: vital statistics, healthcare utilization, child welfare, early childhood, education, juvenile justice, adult justice, employment, public assistance, and homelessness/housing assistance. Within each of these data sources, it also identified specific data elements for inclusion in the IDS, such as birth and death records for vital statistics, or with regard to abuse and neglect investigations, out-of-home care, and in-home services for child welfare. Moreover, it recommended that each of these data elements, e.g. birth records, should be further organized by four key contextual factors. These include: 1) 'person' descriptors (e.g. an expectant mother's age, marital status, education level, race/ethnicity, sex, height, weight); 2) types of public system encounters

(e.g. for prenatal care, the total number of visits, previous live births, cigarette smoking history, risk factors during pregnancy, and infections treated during pregnancy); 3) the place of service provision (e.g. mother's residential address, facility name); 4) and the timeframe (e.g. date of first and last prenatal care visit, data of birth of child). While the panel identified ten data sources, and as many as thirty data elements within these sources, they prioritized their importance based on their frequency of actual inclusion in existing IDS within the AISP network. As Figure 1 shows, the most common data domains included 'out-of-home care' and 'abuse and neglect' data in the child welfare area, while the least common data elements included EMS records and juvenile courts within the health and juvenile delinquency domains.


Out-of-Home Care (Child Welfare)
Abuse and Neglect (Child Welfare)
City or County Jail (Adult Justice)
HMIS (Homelessness/Housing)
Medicaid (Health)
Birth Records (Vital Statistics)
In-Home Services (Child Welfare)
Juvenile Justice Services (Juvenile Delinquency)
Mental Health (Health)
SNAP (Public Assistance)
TANF (Public Assistance)
Alcohol and Substance Abuse (Health)
Death Records (Vital Statistics)
K-12 Public Education (Education)
PHA (Homelessness/Housing)
Workforce Training Programs (Employment)
Nursing Facility MDS (Health)
UI Wages (Employment)
Department of Public Health (Health)
Educ. Homeless Records (Homelessness/Housing)
Law Enforcement (Adult Justice)
Postsecondary Education (Education)
State Corrections (Adult Justice)
All Payer Claims (Health)
CCDF (Early Childhood)
Community Health Centers (Health)
Developmental Disabilities (Health)
Early Intervention (Early Childhood)
K-12 Special Education (Education)
SCHIP (Health)
WIC (Public Assistance)
EMS (Health)
Juvenile Courts (Juvenile Delinquency)
0
2
4
6
8
10
Number of AISP sites accessing data source
Inclusion across sites:
High
Medium
Low

In addition to these recommendations on standardized data sources and elements for inclusion in an IDS, the data standards panel also considered the fact that there is a science to IDS that is currently underdeveloped. This science must examine the quality of the data for inclusion in an IDS, and develop ways of characterizing variations in data quality. This includes attending to the *completeness* of the data for a specific purpose (e.g. characterization of missing data and population coverage); the *consistency* of the data within a particular dataset (e.g. a local zip code that aligns with a state code); and the *uniqueness* of the data, an index of the variety and richness of the data within a dataset, and its corresponding utility for research. This science must also address how to measure the quality and nature of specific record linkages, which in the social services typically takes place around a person or family, but can also take place through geographic area, caseworker, or service provided. While identifying opportunities for record linkage is relatively easy, accomplishing the linking is far more challenging. Often, multiple steps are required to achieve successful linkages, including multiple matching and validation steps such as deterministic geocode matching, probabilistic name matching, and human review (Kumar, 2015). A common language should be developed to communicate to reporting audiences what choices are made and why, and how data linkage decisions can impact interpretation of results.

Our hope is that the federal government and the EBPC will provide guidance that validates the positive ethic represented by viewing these administrative data as an asset that not only can be used, but should be used to create a more responsive government.[4] This is particularly true with highly valued data sources that some state governments may

[4] See Culhane and Hill (2016). Response to 'Request for Comments', Commission on Evidence-Based Policy Making, Department of Commerce. Actionable Intelligence for Social Policy. November 12.

be hesitant to share such as the Department of Labor (DOL) wage records and All-Payer Health Claims data, two of the most important datasets frequently requested by policy analysts and evaluators, but available in a handful of states and counties with an IDS. The federal government should incentivize state and local governments to provide access to these and other data through competitive grant programs that include IDS development such as the Performance Partnership Pilots (P3) and through programs that provide bonus administrative funds for IT development such as the Centers for Medicaid and Medicare Services' recent 90/10 ruling.[5] It could also require multi-system evaluations where necessary, and provide guidance on how federal funds can be used for this purpose. The federal government should also encourage a pathway between state and local government datasets and federal datasets that are going to reside at the Bureau of Census. States seek access to federal data like national earnings records, and college attendance and completion records, so that they can track the outcomes of citizens who leave their jurisdictions. There thus needs to be a transactions platform and standard model for sharing data such that a common encryption system allows state and local data to be linkable to federal data. This would likely take place on a project-by-project basis, as states and local governments at the current time are unlikely to routinely share and store their data with Census.

D. Technology and Data Security

A wide variety of IDS technological solutions have been developed within our AISP Network. These range from highly sophisticated systems that are refreshed on a 24-hour basis through more modest 'sneaker net' solutions that rely on the manual transport of files

[5] See http://www.healthcareitnews.com/news/cms-9010-ruling-increasing-funds-medicaid-it.

via disk that are subsequently uploaded. Technology can present a significant barrier to IDS development to the extent that there is currently little appetite for high-cost high-technology projects within states or local governments. Fully automated IDS systems are typically expensive to develop, given their proprietary, 'one-off' nature, and the reliance on high-cost consultants to maintain them. In fact, only two states in the AISP network have managed to develop highly automated IDS with daily data refreshes —i.e. Washington State and South Carolina—and only a handful of large counties have developed similar systems. This is in part due to the perception among agency executives and legislators that the cost burden, technological expertise, and governance requirements of such systems are too onerous. However, the development of a shared technology solution would allow states and local governments the potential to more easily, inexpensively, and securely store and integrate their administrative data, as well as share tools that would enable them to administer access to, analysis of, and management of that data. Moreover, a shared solution would address one of the biggest fears governments face in developing IDS—the risk of disclosing personally identifiable information, either through a data breach, or by releasing potentially reidentifiable data to external researchers.

We recently conducted an international survey and discovered that many countries/states in Europe, Canada and Australia have found common solutions to these threats. First, they designate a single, trusted entity for storing their data in a secure data warehouse, with a small number of trained individuals within these entities designated to handle personal identifiers. These data warehouses protect personally identifiable information by segregating it from other data, and assigning a unique encrypted identifier to the associated record, ensuring that privacy is maintained, but that the data remains

linkable. Once such a warehouse is established, these IDS have a process in place for reviewing and approving research requests, typically by either an oversight board, or an internal committee of researchers. As a first line of protection, typically only analysts from authorized institutions are allowed access to the data, such as universities, ensuring institutional support, supervision and liability over the end users (researchers). They also have tools for creating the discrete research datasets that are tailored to the specifications of authorized researchers. Once those datasets are prepared, typically by the staff of the IDS, they are set aside, and researchers either come on site to analyze them, or are given remote access through a secure portal (i.e. VPN). Finally, the statistical output are inspected manually, to ensure that the cell sizes are adequate and data visualizations sufficiently aggregate or nonspecific so as to prevent reidentification of individuals within the data. At no point in this process are researchers capable of printing or downloading the prepared datasets, or is a jurisdiction's data transferred away from the jurisdiction. Rather, researchers either work at a secure workstation on site, or send queries via the data portal, and receive statistical output (tables, charts, etc.).

Based on the findings from this survey and from their assessment of IDS implementation challenges in the US, our expert panel has proposed the development of a shared technology solution designed to enable the secure storage and analysis of linked and encrypted individual-level IDS data. States and local governments would maintain their own installations of the system, with site control over the data, and a local data governance structure. While states and counties have many technological solutions in place for integrating data for operational purposes (e.g. enterprise systems used for case management or program eligibility purposes) the solution proposed by the expert panel is

an archival system. Data would be used for planning, evaluation and analysis purposes only (not for patient care or case management), and updated on an annual or semi-annual basis. It would also only contain a thin stream of audited data sources and elements (such as proposed by the expert panel on minimum data sets) rather than a comprehensive set of data variables that governments typically gather in managing agency programs. The goal of this solution would be to create a common, open source infrastructure that could be scaled up on a national level, and accessed by a broader policy analysis, planning and evaluation community.

The expert panel also recognizes major human resource capacity limitations on the part of states and local governments, complicating their ability to project manage an IDS enterprise. The panel therefore recommended the creation of a clearinghouse that would act as both a project management and transaction gateway on behalf of the state and local installations. The clearinghouse would provide state and local governments with the ability to share site-specific metadata to potential users, templated data request submission and approval, DLAs for end users, an authentication process for researchers approved for access to research datasets, a portal for accessing research datasets, project tracking, and invoicing and payment for data use. At the state and local level, the responsibilities for maintaining an IDS would be limited to four main functions: data governance, including review of requests for data and resulting work products, standardized ETL processes, creating research datasets (with tools provided for record linkage and cohort construction), and the manual disclosure review of results. Most of the other project specific transactions involved with the IDS would be handled the clearinghouse. The clearinghouse could also operate as a place where agency leaders, subject matter experts

and other stakeholders with common substantive interests could convene, and have a common space to collaborate or discuss research and evaluation projects. This IDS solution set could be overseen by a governing board of representatives from participating jurisdictions, social scientists, and technology experts, to assure its continuing development and responsiveness to state and local government needs.

III. Conclusion

The U.S. is currently on the verge of a new frontier in social science research and technology fostered by the promise of integrated administrative data systems. Administrative datasets at the federal, state and local level represent a rich source of untapped information about government policies and programs, for determining what works and what doesn't, and for informing how government can better serve the needs of its citizens with limited tax dollars. The good news is that this frontier in the social sciences isn't uncharted or dangerous territory. Governments in Europe, Australia, Canada and in the U.S. have created secure IDS system architectures, and have decades of experience on how to organize a secure and effective IDS. Given advances in technology, and a new climate in favor of evidence-based policy, the US has the opportunity to grow the IDS sphere so that more states and local governments can do this work, and thereby collectively inform the common good.