INTRODUCTION TO

Data Sharing & Integration

AISP

ACTIONABLE INTELLIGENCE FOR SOCIAL POLICY MAY 2020

Contents

Suggested citation: Hawn Nelson, A., Jenkins, D., Zanti, S., Katz, M., Burnett, T., Culhane, D., Barghaus, K., et al. (2020). *Introduction to Data Sharing and Integration*. Actionable Intelligence for Social Policy. University of Pennsylvania.

Acknowledgments

Introduction to Data Sharing and Integration was created by Actionable Intelligence for Social Policy based upon more than a decade of working with and learning alongside our Network and Learning Community Sites, with generous support from The Annie E. Casey Foundation and the John D. and Catherine T. MacArthur Foundation. We are particularly indebted to John Petrila and Richard Gold, who patiently guided our understanding of the legal frameworks necessary for successful cross-sector data integration that we present here.

Introduction

This Introduction to Data Sharing & Integration was created by Actionable Intelligence for Social Policy (AISP) as a primer on the basics of using, sharing, and integrating administrative data. Administrative data are data collected during the routine process of administering programs, but can also be repurposed to support evaluation, analysis, and research.

This resource is designed to help partnerships, collaboratives, agencies, and community initiatives enter into that process of repurposing administrative data thoughtfully and of building the capacity to do so routinely with strong governance in place. We generally refer to these efforts as integrated data systems (IDS), but they have other names, including data hubs, data collaboratives, and data intermediaries.

Whatever they are called, all efforts that seek to leverage shared data to improve individual and community outcomes will likely face common ethical, relational, legal, and technical considerations. This introductory document outlines some of those considerations and provides recommended resources and references for those interested in diving deeper.

Data Sharing vs. Data Integration: What's the Difference?

DATA SHARING is the practice of providing partners with access to information (in this case, administrative data) that they cannot access in their own data systems. Data sharing allows stakeholders to learn from each other and collaborate on shared priorities.

- **Example:** A group of early childhood providers agree to share de-identified, aggregate information about their enrollment and subsidy waiting lists with a local advocacy group in order to better understand the community's early childhood capacity and unmet need.

DATA INTEGRATION is a more complex type of data sharing that involves record linkage, which refers to the joining or merging of data based on common data fields. These data fields can include personal identifiers, such as name, birth date, social security number, or a common encrypted "unique ID" that is used to link or join records at the individual level.

- **Example:** Federally funded early childhood providers share identified information about the children they serve, including name, date of birth, and address, with their state early childhood agency. The state agency links these data with birth records and records from early learning programs funded by state and local sources to estimate how many children are attending publicly funded early learning programs, how many have had no formal early learning experience prior to kindergarten, and how child outcomes vary across these groups.

Benefits, Limitations, and Risks to Data Sharing and Integration

Administrative data sharing and integration has significant benefits, and also limitations and risks.

Benefits:

- **Whole-Person View:** Integrating data across multiple sources provides a more holistic view of the experiences and outcomes of children, households, and families, supporting asset- (rather than deficit-) based approaches. Such views allow analysts to identify bright spots across communities, families, and individuals, and, ultimately, encourage investment in policies and programs that work.
- **Scale:** Analysis using administrative data can include a whole population, rather than a sample, with longitudinal views and comparison groups readily available.
- **Time and Cost:** Reusing administrative data originally collected in the course of service delivery to answer important implementation and evaluation questions can be less time- and resource-intensive than collecting new data using surveys or other means.

Limitations:

- **Availability:** Important information may not be captured in administrative records.
- **Quality:** Since administrative data are not collected for research purposes, data quality issues are common, including missing data, lack of data documentation, and questions about reliability and validity.
- **Access:** Many agencies do not have clear processes and procedures for sharing administrative data, which can make the process of gaining access difficult and time-consuming.

Risks:

- **Privacy Disclosure**: The transfer of data includes the risk of data being accessed improperly, either by accident or through a security breach. Such instances are rare with appropriate safeguards in place, but important to consider.
- **Misinterpretations of Data:** Since these data are originally collected for administrative rather than analytic purposes, data can be misinterpreted without careful understanding and consideration of fields. Such misinterpretations could include an inappropriate analytic plan (e.g., use of predictive tools with data of poor quality), misuse of a variable (e.g., incorrect assumption that PRGENT refers to date of program entry, rather than program entrance exam), or the inclusion of incorrect assumptions when explaining outcomes (e.g., explaining

a reduction in out-of-school suspensions as being a positive indicator of climate, without knowing of a code of conduct revision that changed reporting of suspensions).

- **Replicating Structural Racism:** Since administrative data are collected during the administration of programs and services for individuals in need of social services, the data represented includes people who are disproportionately living in poverty, and, as a result of the historical legacy of race in America, disproportionately Black, Indigenous, and people of color (BIPOC). Seeing these data as race-neutral is inaccurate, and such views could lead to system-level data use that unintentionally replicates structural racism.
- **Harming Individuals:** Certain individual-level uses of administrative data carry particularly high risks of causing personal harm. These include uses that provide case workers, service providers, teachers, law enforcement, etc., with personal information that could lead to biased treatments or punitive action and/or lengthen system involvement.

Because of the complexity of these benefits, limitations, and risks, it is essential that each potential use of integrated data be carefully considered by all relevant stakeholders. First and foremost, the benefit to the individual/community/society at large must outweigh the risks when sharing or integrating data. (See A Toolkit for Centering Racial Equity Throughout Data Integration for a more nuanced discussion of balancing risk versus benefit.)

Why Are You Interested in Data Sharing?

Each data sharing and integration effort is driven by a unique set of stakeholders and their shared risk/benefit assessment and rationale for collaboration. The following table describes some common stakeholder personas and rationales.

Stakeholder Personas and Rationales for Data Integration

Stakeholder/User Persona	Rationale — Integrated Data Needed to:
Internal agency analyst	Better understand overlap in the populations served by multiple programs and assess how best to maximize resources and impact
Community collaborative or organizer	Better understand the lived experiences of residents, for example around housing instability, including homelessness, housing subsidies, evictions, and community support services
Project evaluator at a research firm	Evaluate how a program did or did not impact other areas of program recipients' lives
Director of a large regional nonprofit	Gain a better understanding of the backgrounds and experiences of the people served by the nonprofit, as well as how they fare after leaving the programs
University-based researcher	Better understand whether certain characteristics or experiences are predictive of later outcomes
Deputy director of a state agency	Look at how predictive analytics could be used to reform a particular system in a county/state
Staffer at the governor's policy office	Draft a strategy to target new resources to support vulnerable populations, for example, new investments in early learning for under-resourced neighborhoods

Before beginning any cross-sector data flow, we suggest that you start with a firm understanding of why data sharing is necessary for your community. Reflecting on these user personas can help you develop and articulate a clear purpose for sharing or integrating data. This is essential, as **data flows at the speed of trust**, and trust can be difficult to establish across agencies and organizations that lack a common language and are each focused on their own distinct duties and priorities. It is therefore important that agencies engaged in new data sharing efforts spend time up front discussing motivations, concerns, and expectations in order to build trust and document the rules of the road and the needs of the community before beginning a data project.

Who Are Your Stakeholders?

In this section, we have adapted content from AISP's foundational resource IDS Governance: Setting Up for Ethical and Effective Use (2017) to help you consider a broad range of stakeholder perspectives in your data sharing process from the beginning.

Who are your potential stakeholders?

- **Core Stakeholders:** Who is needed for the effort to achieve success?

  Examples:

  - Data owners and stewards
  - Funding sources
  - Public agency leadership/elected officials

- **Direct Stakeholders:** Who else can help facilitate (or impede) IDS success?

  Examples:

  - Data users (researchers, analysts)
  - Data custodians and technical experts
  - Privacy advocates
  - Community organizers
  - Funders
  - Administrators

- **Other Stakeholders:** Who can broaden the interest of the integrated data and deepen its constituencies? What community groups are "in the data"? Who can this effort help? Who can it hurt?

  Examples:

  - Business groups
  - Good government groups
  - Other citizen and public interest groups

Resource:
For additional strategies and framing, see A Toolkit for Centering Racial Equity Throughout Data Integration, specifically **Toolkit Activity 1, Who Should Be at the Table**

How Will Your Purpose Drive Design?

Ultimately, the shared purpose of your stakeholders should drive your approach to data sharing and integration. The table summarizes three core purposes—Indicators and Reporting; Analytics, Research, and Evaluation; and Operations and Service Delivery—and related implications for approaches to data sharing and integration.

Core Purposes and Approaches for Data Sharing and Integration

Purpose for data sharing and integration	Indicators and Reporting	Analytics, Research, and Evaluation	Operations and Service Delivery
Approach	Data can be summarized and reported at the aggregate	Data must be curated, shared, linked, and then de-identified for statistical purposes	Data must be identifiable and may include case notes to support client-level services
Legal Framework	Data may be publicly available already or may require a simple Data Use Agreement to receive in de-identified format	Data access will generally require multiple agreements, including a Memorandum of Understanding and Data Use License/Agreement to clearly outline permissible access and use	Data access may require client consent and non-disclosure agreements. Data agreements must outline parameters for role-based, credentialed access
Data Frequency	Data may be updated based on reporting cycles, quarterly or annually	Archive of select data may be updated periodically depending on availability and analytic requirements	Daily or real-time updates of entire client records may be required
Privacy and Security	A lack of identifiers or small cell sizes means minimal risk of redisclosure, although demographic information, dates of service, diagnoses etc., mean that data are potentially reidentifiable and must be covered by a Data Use Agreement (except for statistically approved aggregate measures)	Minimal access to identifiable data and small group of approved users means that security requirements are essential but basic	Many users and identifiable data mean that complex permissions and audit trail will be necessary
Cost	Minimal	Moderate	Significant
Difficulty getting started?	Can be difficult, depending on the familiarity of partners	Difficult; labor- and time-intensive	Significant investment of time, labor, and financial resources
Governance	Minimal	Clear parameters around access and use are required, shared processes involving all agencies	

While it is possible to design infrastructure that combines several of the approaches described above, it is always important to differentiate them during design and planning. In most cases, we recommend starting with data sharing and integration work that can be reported in the aggregate and build on early successes toward the much more challenging work of coordinated operations and service delivery at the individual level.

What Data Will You Need?

When considering your purpose and approach to data sharing, it's also important to begin with a realistic assessment of what data would be required to achieve your aims and how easy (or hard) access may be. It is therefore helpful to first broadly classify the accessibility of administrative data into one of three categories:

OPEN DATA	RESTRICTED DATA	UNAVAILABLE DATA
Data that can be shared openly, either at the aggregate or individual level, based on state and federal law. These data often exist in open data portals.	Data that can be shared, but only under specific circumstances with appropriate safeguards in place.	Data that cannot or should not be shared, either because of state or federal law, lack of digital format (paper copies only), data quality or other concerns.

***Open Data:** Data that can be freely used, re-used, and redistributed by anyone. For more information, see What Is Open Data?

Another way of visualizing these differentiations is using a matrix where the data are categorized based upon the technical and legal ability to share. Most data can be shared, with safeguards in place. Those data that cannot be shared could be classified this way because of legal and technical considerations. Classifying high-value data assets of agencies involved in a data sharing effort is an important first step in determining what data sharing and integration is feasible in your context.

Data Classification Matrix

Data can be shared with agreement and approval through governance	**Open data; can be shared without an agreement**
Not shareable	**Technology and/or data structure limits ability to share data**

This is a list of restricted data that AISP network sites are currently integrating;

- Vital records
- Health
- Child welfare
- Juvenile justice
- Adult justice
- Adult protection
- Early childhood
- Education
- Economic security
- Homelessness
- Housing
- Nonprofit social service provider

When classifying data, it is important to include data owners, data stewards, and data custodians in the discussion, as all three will have different perspectives on benefits, limitations, and risks.

The Role of Data Owners, Data Stewards, and Data Custodians

	Role in data sharing process	Role within agency
Data Owner	Accountable for the quality and security of the data and holds decision-making authority regarding access and use	Typically agency leadership that has signatory authority
Data Steward	Responsible for the governance of data, including metadata. Support established processes and policies for access and use	Typically the subject matter experts and data analysts that work with data
Data Custodian	Responsible for the technology used to store and transport data	Typically an IT person or team

Before beginning a data sharing and integration initiative, it is important to conduct a landscape scan of what efforts are in development and underway. Our experience is that every agency and site is sharing and integrating data in some form, typically through ad hoc projects. Engaging with data owners, data stewards, and data custodians can be an effective way to better understand current data access and use practices, and to build upon pieces that are working. Conducting this scan can also prevent duplicative efforts.

Consider the following questions to determine the data you will need and how complicated it will be to access that data:

1. What data are needed to answer the question?

2. Can these data be shared from one agency, de-identified?
    a. **Yes** ≫ Great. Ask the agency about their process for data requests.
    b. **No** ≫ Next question.
    c. **Not sure** ≫ Ask the agency about their ability to share data.

3. Is there a collaborative/institution/agency/data intermediary that can provide access to these data?
    a. **Yes** ≫ Great. Figure out their request process and pay the fee.
    b. **No** ≫ Next question.
    c. **Not sure** ≫ Search for integrated data efforts (e.g., AISP, NNIP, survey relevant agencies and institutions).

4. Will these data need to be shared and integrated one time (also referred to as an ad hoc data request)?
    a. **Yes** ≫ Ask the agency/ies about their process for data integration, access, and use.
    b. **No** ≫ This will be an ongoing project. Next question.

5. Is this a cross-sector effort that requires regular integration of data?
    a. **Yes** ≫ You need to spend some time thinking about governance. And good governance starts with getting the right people at the table, working together.

How to Begin Data Sharing

Once you have the right stakeholders engaged together around shared priorities for data sharing, it's time to consider three simple questions:

Core question	Key considerations	Who decides?	Resources on this topic
Is this data sharing legal?	Are there federal or state statutes that prevent or constrain this data access or use?	This is typically determined by agency-involved legal counsel.	Legal Issues for IDS Use: Finding a Way Forward (2017)
Is this data sharing ethical?	Do the benefits outweigh the risks, particularly for vulnerable populations?	This is typically determined during the review process for data requests that should include data owners.	Data Ethics Workbook (2018)
Is this data sharing a good idea?	What action can be taken as a result of this analysis? What can reasonably be changed or improved based upon the findings?	This is typically determined by a data governance group, including data owners who will respond to insights that emerge from the analysis.	

These questions are typically not linear, and all must be considered prior to cross-sector data sharing and/or integration. The decision-making process should take place within a clear data governance framework, and decisions should be made by a variety of stakeholders. The question that typically receives the most attention, "Is this legal?," is discussed below. However, all three questions are essential. We strongly encourage you to grapple with each to help you decide, together with your stakeholders, whether and how to move forward with data sharing and integration.

Data Governance

> **Data Governance:** The policies and procedures that determine how data are managed, used, and protected.

All agencies and organizations make decisions about their data assets. Without clear data governance, these decisions are often made by individuals, usually data custodians, who are responsible for the technology used to store and transport data, and these decisions may not be consistent or transparent. While data custodians are essential to the work of data sharing and integration, a variety of stakeholders—most importantly, data stewards and data owners—should also be involved in decision making for cross-sector data efforts and that the rationale for these decisions be clearly articulated to the public.

Data governance for a cross-sector data sharing effort can draw upon existing data governance practices within one agency, involve a separate set of policies and procedures, or be a hybrid of the two. Regardless of what approach is taken, we strongly encourage that policies and procedures be explicit and collaboratively agreed upon, rather than implicit and driven by any one individual.

> Data governance for ongoing data sharing and integration should include clearly defined policies and processes to support decision-making, routine meeting structures, and well-documented proceedings—**all fostering a culture of trust, collaboration, and openness.**

The particulars of the policies and procedures will vary widely based on the vision, mission, and guiding principles for data sharing established by the data partners involved. A narrow goal of creating an academic research database to support indicators and reporting will suggest one governance approach, which will differ significantly from the approach needed to support an ambitious agenda to create access to real-time integrated data for credentialed users to support operations and service delivery. We recommend that a site devote time up front both internally and with partner organizations to build consensus around what data sharing and integration is intended to achieve. Taking the time to do this at the outset allows each site to establish tailored rules of engagement that best meet its needs and goals.

Strong and inclusive data governance for cross-sector data sharing and integration should be:

1. Purpose-, value-, and principle-driven: We encourage sites to first identify the purpose for sharing, and then develop vision, mission, and guiding principles. These should include clear value statements around mutual benefit for data partners and the broader community.

2. Strategically located: Before determining the optimal organizational roles and legal framework for data sharing and integration in your context, it is helpful to consider two major functions of data governance:

a. Stakeholder engagement and procedural oversight: Relationship management, convenings, developing policies and procedures, communications, agenda setting, etc.

b. Data management and integration: Secure data transfer, storage, linking, and access for analysis, etc.

Which partner or partners are best positioned to conduct these two functions will depend on a range of factors, including legal authority to use the data as intended by the identified purpose, perceived neutrality among data partners, staff capacity, and technical capacity for data management. In our experience, it is worth the time and effort to consider these practical and strategic questions early on, which can help avoid major stumbling blocks later in executing agreements and allowing data to flow.

3. Collaborative: Governance policies and procedures should be developed cooperatively, and focus on building trust and strong relationships among the partnering organizations. In practice, this may require multiple layers of engagement. Many successful sites have at least three groups that support governance functions:

a. **Deciders**: Executive leader group that supports strategic decision-making

b. **Approvers**: Data subcommittee that supports review and oversight

c. **Doers**: Staff who are charged with daily operations

4. Iterative: Data governance is an iterative process that should guide the whole project life cycle and be revisited and honed regularly as your data sharing effort evolves.

Is This Legal?

Common Legal Terms

Anonymized Data: Data that have been de-identified and then anonymized[1] (including, but not limited to, the removal of all personally identifiable information and aggregated at sufficient geography and cell size or perturbed)

Confidential Data: Data that are not anonymized and not meant for public dissemination.

Data Provider/Data Owner: The owner of confidential data that has agreed to grant access for approved use.

Data User: Individual receiving data for approved use (could be internal or external).

Privacy: Privacy applies to the individual. Privacy measures are concerned with the settings and methods of information gathering. Privacy is also concerned with the type of information being collected.

Security: The process of protecting data from unauthorized access and use throughout the data life cycle.

While the question "Is this legal?" is typically the first asked when beginning a data sharing and integration project, we strongly encourage you to grapple with broader considerations—Is this ethical? Is this a good idea?—to help you decide, together with your stakeholders, whether and how to move forward with data sharing and integration.

There is no simple answer to whether data sharing is legal. It all depends on:

- **WHY you want to share information**
- **WHAT type of information will be shared**
- **WHO you want to share it with**
- **HOW you will share the information**

1 Finch, K. (2016, April 25).

You will need to answer these questions in order to better understand the legal parameters around your data sharing efforts. Use the following prompts as a guide:

- **WHY do you want to share information?**
  - Identify a target population?
  - Identify geographic areas of greatest impact?
  - Evaluate program outcomes?
  - Improve services at the point of intervention?
  - Conduct data analytics?

- **WHAT type of information do you want to share?**
  - Information that does not identify individuals?
  - Information that does identify individuals?
  - Information that might identify a person?
  - Health information?
  - Housing status?
  - Demographics?

- **WHO do you want to share it with?**
  - Law enforcement on the street?
  - The jail?
  - Probation officers?
  - A community treatment provider?
  - A hospital emergency department?
  - A researcher?

- **HOW will you share the information?**
  - How will data be accessed?
  - How will data access and use be approved by the data owner?
    - What legal framework will determine access and use? Ongoing data sharing and integration usually requires multiple agreements, with different purposes (discussed in more detail below):
      - Memorandum of Understanding
      - Data Sharing Agreement
      - Business Associate Agreement (for covered entities)
      - Data Use License/Data Use Agreement
    - Who/what will manage the data governance?
  - How will data be secured during transfer, integration, and use? Who will be responsible for data security?

Legal Considerations

When determining the appropriate legal framework to guide data sharing and integration, begin by identifying relevant legal considerations and authority of data access and use. Four federal statutes and regulations are most relevant to data sharing and integration: the Privacy Act of 1974, the Health Insurance Portability and Accountability Act (HIPAA), 42 CFR Part 2, and the Federal Education Rights and Privacy Act (FERPA). In addition, states have statutes, regulations, ordinances, orders, and rules that may exceed federal protections for administrative data sharing. For this reason, it is important to work with legal counsel to ensure that all relevant legal considerations, specifically authority, are considered prior to developing a legal framework. For example, many state agencies regularly share and integrate administrative data based upon legislative mandate and/or Executive Order.[2]

Do we need consent?

Whether consent is needed largely depends upon the type of data, who is accessing the data, and how the data will be used. There are many considerations, and often no clear answer. Data Across Sectors of Health have created some helpful guidance on Informed Consent, particularly in thinking about specific elements that are required based upon HIPAA, 42 CFR Part 2, and FERPA. We strongly recommend that any decisions around consent be carefully considered with a variety of stakeholders through data governance processes. In general, however, consent is not usually required for research, evaluation, and planning efforts where individual identifiers will not be seen or used by analysts.

Privacy Act of 1974

- Stringent confidentiality provisions, but permits disclosure without consent for "routine use."
- Routine use: the use of a record for a purpose compatible with the purpose for which it was collected.
- This has been used to permit researchers and evaluators access, even to identifiable data, so long as the project meets an administrative purpose such as program planning or improvement.

Health Insurance Portability and Accountability Act (HIPAA), 1996

- Protected Health Information (PHI): Any information that can be linked to an individual about health status, provision of health care, or payment for health care that is created or collected by a "covered entity" or "business associate" of a covered entity; includes 18 identifiers that must be treated carefully.

2 See AISP Network Site Case Studies for more information regarding the legal authority that AISP Network Sites operate under.

- The Privacy Rule permits a covered entity to use and disclose PHI in a limited data set for research; can be disclosed for purposes of research, public health, or health care operations.
- Data Use Agreement (DUA): An agreement into which the covered entity enters with the intended recipient of a limited data set that establishes the ways it may be used and how it will be protected.
- Limited Data Set: Refers to PHI that excludes 16 categories of direct identifiers and may be used or disclosed, for research, public health, or health care operations, without obtaining either an individual's authorization or a waiver or an alteration of authorization for its use and disclosure, with a DUA.

42 CFR Part 2, Federal Regulations Governing the Confidentiality of Alcohol and Substance Abuse Records

- While HIPAA protects PHI in possession of covered entities, 42 CFR Part 2 protects information around alcohol and substance abuse treatment records regardless of who has possession.
- Allows for use of covered information for research without consent if certain conditions are met.
- There is crossover with HIPAA.
- State laws may exceed federal protections in 42 CFR.

Federal Education Rights and Privacy Act (FERPA)

- FERPA applies to all schools that receive federal funds.
- There is no "research exemption," but a total of 15 exceptions in FERPA.
- IDS can use FERPA-protected data through a process using a variety of exceptions, depending on the project.
- The exceptions most pertinent to IDS are the:
  - School Officials Exception
  - Studies Exception
  - Audit/Evaluation Exception
- See Department of Education Privacy Technical Assistance Center Guidance, 2017.

The table summarizes the legal approaches and signatories needed for three common uses of administrative data: Indicators and Reporting; Analytics, Research, and Evaluation; and Operations and Service Delivery.

Legal Approaches and Signatories for Data Sharing and Integration Uses

Purpose	Indicators and Reporting	Analytics, Research, and Evaluation	Operations and Service Delivery
Legal Approach	Typically, data are already publicly available or can be accessed in an aggregated format with a simple Data Use Agreement.	Data access will generally require multiple agreements, including a Memorandum of Understanding and Data Use License/Agreement to clearly outline permissible access and use.	Data access may require client consent and non-disclosure agreements. Data agreements will also need to outline clear parameters for role-based, credentialed access.
Signatories	Data owner/s and data users	Data owner/s, data integration staff (if separate from data owner/s), data users/ licensee	Individual receiving service/s (through informed consent), data owner/s, data integration staff (if separate from data owner/s), data users/licensee if applicable

Foundational Agreements for Data Sharing

Data sharing and data integration efforts are as much relational as they are technical. Several documents formalize the relationship between agencies to ensure data sharing complies with all federal and state statutes and that everyone involved is clear on the rules of the road. The following sections will help you familiarize yourself with the types of documents that commonly govern data sharing practices.

Memorandums of Understanding and Data Use Licenses

At least two legal agreements are needed for governing data access and use for integrated data: a Memorandum of Understanding (MOU) and a Data Use License (DUL). State agencies may use different terms to refer to these documents, including data security agreement, information sharing plan, memorandum of agreement, data sharing agreement, and data use agreement. It is helpful to learn the terminology used by the agencies you hope to partner with and to use this terminology consistently.

Memorandum of Understanding (MOU): A foundational agreement between the host organization and the data contributors. The MOU sets forth the core features of the data sharing/integration structure as well as the legal rights and responsibilities of each party involved. A good MOU will codify both the legal requirements and operational structure. An MOU should be written in plain language so that anyone can understand its terms.[3] It may also memorialize the mission, values, and ethical framework of the data sharing effort.

3 To learn more about the legal framework of an IDS and for specific guidance on MOUs, see Legal Issues for IDS Use: Finding a Way Forward; Guidelines for Developing Data Sharing Agreements to Use State Administrative Data for Early Care and Education Research; and The State Data Sharing Initiative.

Data Use License (DUL): This document outlines the duties of any approved recipient of data or data user, which likely includes the protection of confidential data, use limited only to what is outlined in the license, date for termination, and immediate notification if data privacy is breached. The DUL might also include requirements for citation, peer review, or advance notification prior to publication of any research findings.

Data Security

Data security is often seen as a technical consideration, but it is a multi-dimensional process that also includes legal, procedural, and physical components. These components vary widely by site, but could include:

Legal:

- Organizational structure (e.g., entity with authority to conduct data integration, entity with liability/board/cyber insurance)
- Data sharing agreements, including MOUs, DULs, Cooperation Agreements, and Non-Disclosure Agreements (NDAs)
- Data license process
- Data security plans

Technical:

- Regular security audits
- Passwords (dual authentication)
- Encryption (data at rest, data in transfer)
- Secure servers (e.g., public cloud, private cloud, on-premise)
- Data integrity measures (e.g., backups)
- Controlled, limited access
- Private network
- De-identification/anonymization standards and procedures

Procedural:

- Regular communication among staff, both vertical and horizontal
- Business procedures/process that explain "how we work"
- Regular staff training
- Oversight of board that includes data stewards/data owners
- Incident response protocols
- Logs (audit trail)
- Data quality review
- Collaborative checklist for data license requests
- Separation of duties for staff

Physical:

- Hardened work stations
- Locked offices

The goal for these multiple layers of security is to prevent a data breach or security incident. Any data effort needs a clear incident protocol to prepare for potential data breach and security incidents. While breaches and incidents are uncommon, how a site prepares for and responds to such threats is essential for building and maintaining trust among data contributors and the broader community.

> A **data breach** is the intentional or unintentional release and use of protected data (generally understood as data that can lead to identification of a person), e.g., a malicious intruder with intent to use stolen data.
>
> A **security incident** is an event that leads to a violation of established security policies and puts protected data at risk of exposure. e.g., a malware infection, unauthorized access, insider breach, or loss of equipment.

Technical Approaches for Data Sharing and Integration

You may wonder why it's taken this long for us to get to this section. When agencies begin to share and integrate data, the work is commonly approached as a technical project. While we understand this tendency, we encourage sites to view the technical components of this work as a process to support analytics and insights, and ultimately, improvements in policies, practice, and outcomes. **The technical approach should not be the focus or end goal, but rather a tool or means to the goal.** Most importantly, the technical approach will change as data sharing and integration develops and expands, and as technological advances shift best practice. For this reason, we encourage starting small, and initially investing more in relationships and human capacity than in large data IT infrastructure.

There are many technical approaches to sharing and integrating data, and it's important that **purpose drives design**. Prior to building or procuring anything, we encourage you to think through the following considerations.

Legal Framework:

- What legal entity is charged with data security?
- How will this impact where data "lives"?

Staffing[4]

- Who will manage stakeholder coordination and the project management components of data sharing and integration, including communication among data owners, data stewards, and data custodians?
- Who will conduct the technical components of the work?
- Is there a need for additional internal capacity? External capacity?

Data Collection

- How are data currently collected?
- How are metadata stored, updated, and communicated?
- How will data be extracted from current data management structure?
- How will data be transferred?

Data Management

- What existing data management structures can be used for sharing and integration?
- What will need to be built/procured/partitioned/etc.?
- Where will data be stored? Cleaned? Integrated? Analyzed?

Security & Privacy

- Who/what agency will be in charge of ensuring data security?
- What governance structures will provide oversight to ensure data privacy?

Data Linking

- How will data be linked? What identifiers will be used?
- Will data be linked ad hoc (as needed) or will data be linked and integrated upon receipt?

Data Access & Dissemination:

- How will data be de-identified and anonymized?
- How will data be accessed by internal analysts? External analysts?
- How will data be disseminated?

Determining the technical approach for any data sharing and integration efforts is a bit like building a bridge while walking on it. While planning is important, even more important is the ability to be flexible and quickly adapt to challenges and opportunities as they arise. The technical solutions in this space are disruptive and fast-changing, so we encourage the reuse of current systems when possible and starting with incremental investments prior to committing to enterprise-level shifts in data management. Most importantly, we encourage sites to look around at what has worked in other places and learn from prior efforts. We all benefit from building strong technical approaches for data access and use that are legal, ethical, and actionable.

4 See Appendix D of the IDS Governance: Setting Up for Ethical and Effective Use (2017).

Best Practice: An emerging best practice standard as of 2020 is to share data with external partners, such as researchers, through a secure portal, where statistical queries are submitted remotely, and there is no transfer or view of personally identifiable information by the data user. See California Health and Human Services Record Reconciliation and Research Data Hub in the next section.

Resource: For additional framing and guidance, see Technology for Civic Data Integration, a report by MetroLab Network, AISP, and NNIP (2018).

Developing a Data Model

A data model is the abstract framework used to organize data elements and standardize how they relate to one another in the context of complex social and administrative systems. Data models describe structure and integrity aspects of the data stored in different data management tools and databases used for integration.

Example: If data sharing and integration involve PreK-12 education, then any data model should include information that allows the analysis to control for attendance. A data model would include absences (excused and unexcused), suspensions (short- and long-term), and days in membership (how many days for the attendance period). If possible, teacher attendance would also be included, so a data model would also include teacher of record and how many days the teacher was absent.

When you are selecting which data will be shared and integrated as part of your data model for a given analysis, we encourage you to consider the following broad principles:[5]

Organize around the life course: Administrative data can be used to describe service involvement over a lifetime. While connecting a lifetime of records is unnecessary, intentionally focusing on key developmental periods and transitions can be important for high-impact analytics.

Include contextual factors: Individuals live within households and families, and in neighborhoods, and they attend schools in cities, counties, states, and regions. Using individual-level records without broader context (including place-based information) limits insights and opportunities for action.

5 Adapted from Establishing a Standard Data Model for Large-scale IDS Use (2017)

Ensure reliability and validity: For data sharing and integration, administrative data are collected during the routine process of administering programs and reused in a way that was not originally intended (e.g., for analytics and insights). Data quality can impact reliability and validity, and is a primary consideration when developing a data model.

- Reliability refers to data that produce similar results under consistent circumstances, e.g., a record consistently links an event date to an event (such as a program start).
- Validity refers to the extent to which conclusions drawn from analysis are accurate, e.g., an evaluation has a sufficient data model to measure outcomes related to a specific programmatic intervention, rather than other changes (such as improved funding or change in enrollment).

For further guidance on human service–oriented data models, see AISP's Establishing a Standard Data Model for Large-scale IDS Use.

For guidance on early childhood–specific data elements, see Early Childhood Data Definitions: A Guide for Researchers Using Administrative Data. For an example of a state process of developing a comprehensive data model, look to Shared Measures of Success to Put North Carolina's Children on a Pathway to Grade-Level Reading.

For guidance on elements relevant to courts and child welfare, see Data Sharing for Courts and Child Welfare Agencies.

For guidance on elements relevant to homelessness, see Market Predictors of Homelessness: How Housing and Community Factors Shape Homelessness Within Continuums of Care.

Use Cases of Integrated Data

AISP recommends a developmental approach to data integration, beginning with basic data sharing for aggregate, descriptive analysis, and building with complexity as use cases are successful.

Community-based indicator projects can be a high-impact and relatively simple way to build trust and gain momentum for data sharing.[6]

Common types of indicators include:

- input indicators (i.e., measuring the funding, staff, and other key inputs necessary for program implementation)
- process indicators (i.e., measuring the program's key activities and outputs, such as the number of families served)
- outcome indicators (i.e., measuring the short- and long-term effects or changes)
  - See the California Strong Start Index and Charlotte/Mecklenburg Quality of Life Explorer for examples of data sharing and integration for indicators.
  - Learn more about the specifics and development of indicators for program evaluation from the CDC.

Below, we describe six additional, more complex use cases for integrated data that successfully built on existing momentum, established relationships, and strong governance to advance cross-agency policy and practice.

Early Childhood Iowa Statewide Needs Assessment

Early Childhood Iowa uses its integrated data system to better understand early childhood service utilization and the early childhood workforce. Data sources included public health, education, and human services data. Initial efforts focused on determining an unduplicated count of children in care from birth to age 5 across the state, and found that 73% of children had at least one center-based experience during the year before kindergarten entry. Importantly, the analysis revealed significant gaps for vulnerable children, particularly those in rural counties. Analysis also found shortages in both the quantity and quality of the early childhood workforce, with staffing challenges being particularly acute in rural counties, which comprise 89% of Iowa counties. This project was an important precursor to receiving a Preschool Development Grade, Birth-5 grant in 2019.

Charlotte-Mecklenburg Family Homelessness Snapshot, 2014-2015

The Housing Advisory Board of Charlotte-Mecklenburg, with support from Mecklenburg County Support Services, used the county's integrated data system to better understand families experiencing housing instability and homelessness. Analysis found a disconnect between students identified by schools as experiencing homelessness (using McKinney Vento records) and children and youth identified as literally homeless in an emergency shelter, transitional housing facility, or unsheltered location (using Housing Management Information Systems [HMIS] records). Some individuals were not identified as experiencing homelessness by the school system, but had, in fact, experienced homelessness in an emergency shelter. This discrepancy was communicated

6 Learn more about the National Neighborhood Indicators Partnership.

to the county, local providers, and the school district; as a result, additional social workers were placed within the emergency shelter system to identify children for and connect them to services.

Miami-Dade IDEAS Consortium for Children

The Miami-Dade IDEAS Consortium for Children used their integrated data system to map aggregate outcomes of early childhood education that better inform decision-making at local agencies, including resiliency mapping by census tract to identify areas of persistent need and areas where children are outperforming socioeconomic expectations. The Consortium determine d that while 83% of children entering kindergarten had a preschool experience, countywide, there are significant opportunities to increase access to high-quality preschool programs, as only 31% of preschool programs are licensed. Using preschool attendance data and K-12 data, analyses also found that children who consistently attended preschool scored higher on math and reading assessments in preschool and in kindergarten, especially for children living in census tracts with higher concentrated disadvantage. These analyses have supported work to improve early childhood program attendance.

California Health and Human Services Record Reconciliation and Research Data Hub

California's Health and Human Services Agency (CHHS) partnered with the USC Children's Data Network (CDN) to conduct a record reconciliation that linked and organized administrative, individual-level client records across eight major CHHS programs (31+ million records prior to deduplication) and generated an encrypted master client identifier for interagency use. This record reconciliation supports everyday operations as well as the development of longitudinal, cross-sector evaluation and research that facilitates a holistic view of client experiences. These efforts have spurred the development of a secure, cloud-based research enclave for hosting record-level research data sets and accompanying linkage keys. Once operational, this environment will provide carefully controlled, role-based access to analysts within CHHS.

In the longer term, the goal is to develop protocols that, with necessary approvals, will give external university-based and other research partners access to curated data sets and statistical resources within this analytic environment. It is anticipated that this secure platform will advance rigorous evaluation, improve the reproducibility of research, create efficiencies in data management, and further the engagement of university-based researchers with government. Additionally, this research data hub will enhance record security and client confidentiality through data access and security protocols that can be more carefully audited.

The Use of Integrated Data to Inform Quality Pre-K Expansion in Philadelphia

The City of Philadelphia, in partnership with the Penn Child Research Center at the University of Pennsylvania, used its CARES integrated data system to inform how Philadelphia could best leverage revenue from its newly established soda tax to expand universal pre-k offerings. This is an example of how data integrated at the individual level can be aggregated and mapped geographically to drive resource allocation to neighborhoods and families with complex needs. The city was also able to build on this initial success and execute a separate legal agreement allowing trusted practitioners access to information about families who might benefit from the new pre-K slots so that those most in need could be connected directly to services.

The Massachusetts Opioid Epidemic — A data visualization of findings from the Chapter 55 report

After the passage of Chapter 55 legislation in Massachusetts, multiple state agencies collaborated to use integrated data to better understand trends in opioid-related overdose and death. Their efforts resulted in a more holistic view of those affected by the public health crisis and improved their ability to target resources and interventions, leading to promising reductions in opioid-related deaths in 2019.

Conclusion

Data sharing and integration have enormous potential to improve cross-agency collaboration and outcomes for families and communities. However, the potential benefits of data sharing and use must always be weighed alongside the limitations and risks. To ensure that this balance is carefully considered, start with a strong "why" for your data sharing or integration effort and include diverse stakeholders in planning and ongoing governance. Asking three simple questions at the outset of your effort—Is it legal? Is it ethical? Is it a good idea?—can help guide the initial work and ensure you are on the right track.

Data flows at the speed of trust, and we cannot overstate the need to focus on relationship building. We encourage you to start with a strong vision, mission, and purpose. Then, identify and regularly seek input from a core group of stakeholders that can inform approach, processes, and policies. Spend time thinking through the data that will be needed to conduct high-interest use cases that can lead to action, and then craft a legal framework, including data security approaches, that can facilitate this data access and use. Develop a technical approach and data models that can support your use cases and evolve to match changing needs and capacity. And most importantly, remember that sites have approached this work in countless different ways, and while this work is challenging, the benefits are worth the effort. Good luck!

Additional Reading

This document is intended to provide a high-level introduction to data sharing and integration. For those embarking on new data efforts, we also recommend some additional reading to expand upon topics discussed above.

- A Toolkit for Centering Racial Equity Throughout Data Integration (2020), AISP, by Hawn Nelson et al.
- Nothing to Hide: Tools for Talking (and Listening) About Data Privacy for Integrated Data Systems (2018), Future of Privacy Forum & AISP.
- Technology for Civic Data Integration (2018), MetroLab Network, AISP, and NNIP
- Unlocking the Value of Data Sharing: A Look Across Five Sectors (2018), Data Across Sectors for Health, by Eckart
- The Integrated Data System Approach: A Vehicle to More Effective and Efficient Data-Driven Solutions in Government (2017), by Fantuzzo, Henderson, Coe, & Culhane
- Connecting the Dots: The Promise of Integrated Data Systems for Policy Analysis and Systems Reform (2010), by Culhane, Fantuzzo, Rouse, Tam, & Lukens
- AISP case studies of Integrated Data System Network Sites, 2014-2018
- Research Connections, Working with Administrative Data (ongoing), National Center for Children in Poverty

References

California Strong Start Index. (2019). First 5 Association of California, Children's Data Network.

Centers for Disease Control and Prevention. (2016, December 2). CDC Approach to Evaluation: Indicators. CDC Program Performance and Evaluation Office.

Children's Data Network. (n.d.). CHHS Annual Record Reconciliation and Research Data Hub. California Health and Human Services.

City of Charlotte/Mecklenburg Quality of Life Explorer. (2017). Mecklenburg County, City of Charlotte, UNC Charlotte.

Clark, A., Lane, J. T., & Marcus Gaines, A. (March 2017). Charlotte-Mecklenburg Family Homelessness Snapshot 2014-2015. University of North Carolina at Charlotte Urban Institute.

Data Across Sectors for Health & The Network for Public Health Law. (2018, November). Data Sharing and the Law: Deep Dive on Consent.

Department for Digital, Culture, Media & Sport. (2018, June 13). Data Ethics Workbook. Gov.UK.

Early Childhood Iowa. (2019). Early Childhood Iowa Statewide Needs Assessment.

Evans Harris, N., Di Maura-Nava, S., Hawn Nelson, A., Hendey, L., & Kingsley, C. (2018). Technology for Civic Data Integration. MetroLab Network, Actionable Intelligence for Social Policy (AISP), and National Neighborhood Indicators Partnership (NNIP).

Finch, K. (2016, April 25). A Visual Guide to Practical Data De-Identification. Future of Privacy Forum.

Future of Privacy Forum & Actionable Intelligence for Social Policy (2018). Nothing to Hide: Tools for Talking (and Listening) About Data Privacy for Integrated Data Systems.

Gibbs, L., Hawn Nelson, A., Dalton, E., Cantor, J., Shipp, S., & Jenkins, D. (2017, March). IDS Governance: Setting Up for Ethical and Effective Use. Actionable Intelligence for Social Policy, University of Pennsylvania.

Hawn Nelson, A., Jenkins, D., Zanti, S., Katz, M., Berkowitz, E., et al. (2020). A Toolkit for Centering Racial Equity Throughout Data Integration. Actionable Intelligence for Social Policy, University of Pennsylvania.

King, C. & Maxwell, K. (2017). Early Childhood Data Definitions: A Guide for Researchers Using Administrative Data. OPRE Research Brief # 2017-67. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

LeBoeuf, W., Barghaus, K., Henderson, C., Coe, K., Fantuzzo, J., & Moore, J. The Use of Integrated Data to Inform Quality Pre-K Expansion in Philadelphia (2017). Research Briefs.

Massachusetts Department of Public Health. (n.d.). The Massachusetts Opioid Epidemic: Data Visualizations from the Chapter 55 Overdose Report.

Miami-Dade IDEAS Consortium for Children. (n.d.). Integrating Data to Improve Early Learning and School Readiness for All Children. University of Miami.

Nisar, H., Vachon, M., Horseman, C., & Murdoch, J. (2019). Market Predictors of Homelessness: How Housing and Community Factors Shape Homelessness Rates Within Continuums of Care. U.S. Department of Housing and Urban Development, Office of Policy Development and Research.

North Carolina Early Childhood Foundation. (2017). Shared Measures of Success to Put North Carolina's Children on a Pathway to Grade-Level Reading. North Carolina Early Learning Foundation.

Open Data Handbook. (n.d.) What Is Open Data? Open Knowledge Foundation.

Petrila, J., Cohn, B., Pritchett, W., Stiles, P., Stodden, V., Vagle, J., Humowiecki, M., & Rozario, N. (2017). Legal Issues for IDS Use: Finding a Way Forward. Actionable Intelligence for Social Policy, University of Pennsylvania.

Privacy Technical Assistance Center (PTAC). (2017). Integrated Data Systems and Student Privacy. U.S. Department of Education.

Shaw, S. H., Lin, V., & Maxwell, K. L. (2018). Guidelines for Developing Data Sharing Agreements to Use State Administrative Data for Early Care and Education Research. OPRE Research Brief #2018-67. Office of Planning, Research and Evaluation, Administration for Children and Families, U.S. Department of Health and Human Services.

State Data Sharing Initiative. (n.d.). Data Sharing for Policy Analysis & Program Evaluation. The Center for Regional Economic Competitiveness.

U.S. Department of Health and Human Services, Administration for Children and Families. (2018). Data Sharing for Courts and Child Welfare Agencies.

Wulczyn, F., Clinch, R., Coulton, C., Keller, S., Moore, J., Muschkin, C., Nicklin, A., LeBoeuf, W., Barghaus, K. (2017). Establishing a Standard Model for Large-Scale IDS Use. Actionable Intelligence for Social Policy, University of Pennsylvania.

Actionable Intelligence for Social Policy
University of Pennsylvania
3701 Locust Walk, Philadelphia, PA 19104
215.573.5827 www.aisp.upenn.edu

AISP