Towards State-of-the-Art IDS Technology and Data Security Solutions

Actionable Intelligence for Social Policy,
Expert Panel Report

Prepared by

David Patterson, Niall Brennan, Andreas Haeberlen,
Aaron Schroeder, Adam Smith, and Ken Steif

MARCH 2017

ACTIONABLE INTELLIGENCE
FOR SOCIAL POLICY

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

Table of Contents

Table of Contents

I. Introduction

The federal government spent $3.8 trillion in 2015 across an array of programs, deductions, and entitlements. Do these programs work as intended? For every dollar spent, do we improve someone's quality of life? Lift a child out of poverty? Prevent a veteran from experiencing homelessness?

Early in 2016, House Speaker Paul Ryan and Senator Patty Murray crafted a bill, later signed into law by President Barack Obama, to form the federal Commission on Evidence-Based Policy. The goal of the Commission is to adopt data-driven decision making into the budgeting and legislative process.

The creation of this new, bipartisan-supported Commission indicates that evidenced-based policy is an issue that resonates across the political spectrum. With the requisite political goodwill in hand, the remaining challenge is to create technology that can proliferate evidence-based policy by reducing the costs associated with planning and evaluation.

While the engine of evidence-based policy is statistical evaluation, the fuel is high-resolution, individual-level administrative data collected for case management and billing purposes. Here we define evaluation as the ability to identify, from the administrative data, the specific effect of a government program on a set of outcomes across health, education, economic, and other important domains.

Administrative data are often "observational" data. These are data collected by jurisdictions and agencies in order to provide specific services. Many of these data points are specific to the agency that collects them. For example, a Department of Corrections might collect basic information on demographics, incarceration, etc. While these data by themselves help address the resource allocation or program eligibility questions for the Department of Corrections, when data are linked across agencies, the resulting dataset can help to address a variety of planning and evaluation questions across a wide range of domains, including serving as an outcome measure for other programs.

Because these data are not collected, stored, and disseminated explicitly for research, planning, or evaluation, there are large barriers even for government agencies who wish to evaluate their own programs or to engage in comprehensive, cross-domain analyses. Data are often stored in "legacy" databases managed by a small number of information technologists and in a relational format that is not conducive to extraction, linking, or statistical analysis. Individual agency databases store only client data and do so in disparate and often incompatible formats/structures, which does not allow for cross-cohort comparisons. Metadata are vague and often missing in their entirety.

All of these limitations drive up the cost of research and evaluation, a result that can delay or prevent necessary changes to vital public services. Many solutions exist for linking administrative data, but integrated data systems (IDS) have emerged as the gold standard. IDS are enterprise-scale, custom software solutions that link data across government departments.

In our work identifying best practices in jurisdictional IDS from across the United States, AISP identified only two states that use IDS for evaluation and planning: Washington and South Carolina. We located just four local jurisdictions that maintain their own IDS: Allegheny County, PA (Pittsburgh), Los Angeles County, New York City, and Philadelphia. Other IDS are run by external stakeholders, like universities. The failure of most states and localities to adopt IDS may be explained by the belief among legislators that data linking is an inappropriate or unethical function of government, or that the cost burden, technological know-how, and governance requirements are too onerous.

Our vision for AISP Innovation is to facilitate a network of jurisdictions and evaluators from across the United States in working together to further a culture of evidence-based decision making in government. This document proposes a series of technology solutions that help achieve this vision.

The purpose of this document is to inform any site who might want to develop such a system. We recommend that jurisdictions seeking to develop this solution either join the AISP effort or work as a coordinating entity for a shared infrastructure. We pay particular attention to system workflow and technology components, but ultimately, our aim is to provide enough information for a developer to offer their own technology solutions to our proposed model.

The AISP Innovation Technology and Data Security Expert Panel is responsible for this report. The Panel includes senior administrators of integrated data systems nationally, computer security experts, and experienced research evaluators.

Section II presents the results of an IDS survey distributed to help guide the technical recommendations provided here. Section III introduces the solution in brief. Section IV describes the use cases. Section V presents the system workflow, and Section VI presents the major technological components we envision being part of AISP Innovation.

II. International IDS Survey Results

Beginning in 2014, the AISP team became particularly interested in addressing the barriers to IDS development, and began to think about how IDS standards could address technological and data governance requirements that maintain the sensitive nature of the data while reducing the costs associated with custom software development. Not surprisingly, we knew we needed more information to understand the current design and function of IDS, particularly for the most robust integrated data systems. Many of these advanced IDS are from countries outside of the United States that manage data related to national healthcare programs, vital statistics, and social programs. As these systems are on the leading edge, we surveyed these organizations to guide the technology solution proposed here.[1] The most actionable data governance and technology-related insights from the survey appear below.

A. IDS Governance

- Four of the six responding IDS have governance structures that allow one "director" to set policies and procedures. The remaining sites have a governing board that plays this role. However, in all cases, either the governing board or a secondary oversight board reviews and approves projects.
- All IDS respondents say that program evaluation priorities are set by both government stakeholders and external stakeholders.
- All but one IDS support their operation by charging evaluators for externally proposed projects. In every case, the IDS charges a fee depending on the scope of the project.

B. Technology

- Sixty percent of IDS respondents say that interdepartmental data exchange happens by way of hand-to-hand contact (e.g., hard drive). The remaining 40% use a secure FTP.
- Eighty-three percent of respondents say that data are stored in one centralized database.

1 The AISP international IDS survey was sent to IDS directors in Australia, Canada, Denmark, Finland, France, Germany, the Netherlands, New Zealand, Norway, Sweden, the UK, Brazil, and Mexico. Fifteen surveys were sent out, and six individuals responded.

- Four of six respondents do not store personally identifiable information, such as social security numbers.
- All but one IDS have a person responsible for linking data across data sets.
- All but one IDS have an automated approach for ensuring that data cannot be re-identified.
- All but one IDS have a system that allows researchers to interface with data remotely.

The common element among most, if not all, of these systems is the presence of a governance structure, including a designated group of experts who can approve projects, authorize users, and collect fees. All but one IDS employ personnel whose job is to link data across data sets on demand, and use an automated approach for preventing the re-identification of data. Finally, all but one IDS respondent use remote technology that allows researchers to interface with data without ever actually possessing them. This helps the IDS maintain the privacy of the data while lowering the costs of having evaluators working on site.

III. The AISP Innovation Solution

The AISP team is focused on creating a set of standards that replicate or improve upon the features of these advanced IDS. As many jurisdictions have neither the financial wherewithal nor the personnel to manage enterprise-scale data systems, our solution attempts to automate many of the human-scale IDS processes without trading off privacy and security. AISP Innovation does so by way of two interconnected systems that (a) help to standardize the governance process and (b) allow evaluators to remotely interact with data without ever having the actual data in hand.

With respect to governance, AISP Innovation features a web-based management information system called the Clearinghouse. Accessible by a browser, the Clearinghouse enables evaluators and jurisdictions to propose and negotiate planning and evaluation projects, sign data use agreements and financial contracts, and help communicate results to relevant stakeholders. AISP will provide support to help jurisdictions set up a governance committee to help facilitate these processes, if one does not already exist. The Clearinghouse then becomes the tool through which such a committee can communicate its planning and evaluation agenda to evaluators internally and across the nation.

While the Clearinghouse is a cloud-based solution managed by AISP, the data accessed by evaluators remain owned by the jurisdictions that participate in AISP Innovation.

The DataHub is a standalone hardware solution—an appliance, which is located at a jurisdiction and houses that jurisdiction's administrative data. To keep staffing costs down, data would be updated annually or semi-annually.[2] Each jurisdiction's DataHub would be completely independent of existing legacy database systems. The AISP Data Standards Expert Panel has identified a core set of variables and databases that would sit on a jurisdiction's DataHub (see Wulczyn et al., 2017). While we envision an extract, transform, load (ETL) process for the DataHub, we wish to leave the design specifics up to the eventual developer.

To manage access to their data, governance officials at a given jurisdiction can access a series of graphical administrative functions that are part of the cloud-based Clearinghouse software. This might look something like Figure 1.[3]

2 We envision that future versions might allow for data to be updated with more frequency.
3 This figure is only meant to be suggestive, and does not represent the final form of the administrator panel.

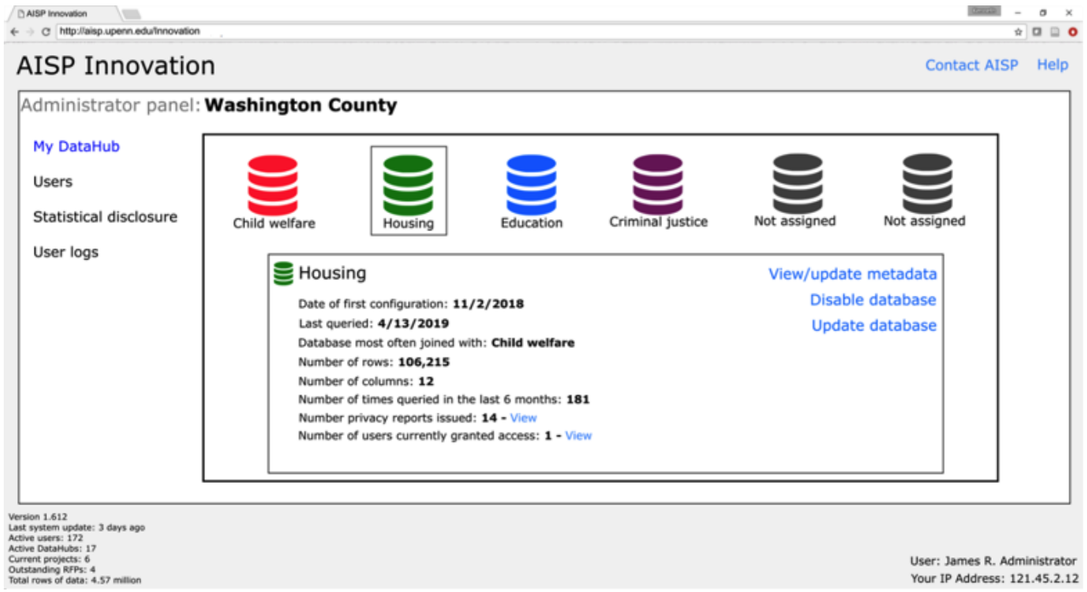


Figure 1: A Mockup of the Clearinghouse Administrator Panel.

The Clearinghouse also features a user interface that allows those researchers who have been granted access to interact with a given DataHub remotely. Unlike our survey respondents who employ statisticians and other experts to link data and ensure they cannot be re-identified, AISP Innovation, by way of a jurisdiction's DataHub, will automate this process by building cell size routines and other privacy checks directly into the system.

By way of a front-end Clearinghouse interface and a backend application program interface (API), the evaluator issues queries; the system links records and performs a privacy check to ensure that these data cannot be used to re-identify individuals.[4] Assuming the privacy check is accepted, a data "view" is created, and the researcher is permitted to analyze the data.

At no time will the jurisdiction's identifiable and unencrypted data ever be transferred away from the jurisdiction. The analyst issues a command on the Clearinghouse that is relayed to the DataHub, which performs the calculation and returns the result to the analyst via the Clearinghouse. The only feedback the evaluator receives is in the form of statistical output (tables, charts, etc.), reports that describe the data linkage process, or error messages related to queries that do not fit a given privacy protocol.

Within the automation, AISP Innovation will make privacy and data security paramount. The administrator controls will enable jurisdictions to manage how their data are used and how the system as a whole employs industry-standard encryption and privacy protocols.

4 We leave the full functionality of this feature, like that of many others described in this report, up to respondents to the forthcoming request for proposals.

Although a jurisdiction might already run a large-scale IDS for case management, the separate, standalone, AISP Innovation solution will enable faster and more cost-effective evaluation. Further, smaller jurisdictions who cannot currently afford a fully customizable deployment of commercial enterprise software may elect to adopt AISP Innovation as their lone IDS solution.

The development of a business plan is currently under way, and our work thus far has identified a willingness among external evaluators to pay for access to the kind of data that will be housed in AISP Innovation. Whatever this fee might be, the majority of revenue will remain with participating contracted jurisdictions, while AISP collects a percentage to maintain the system.

We now turn to use cases—describing who will use the system and to what end.

IV. Use Cases

We begin our discussion of the AISP Innovation system by describing the six use cases we envision.

1. The first use case is evaluators who want to analyze high-quality observational administrative data at a given jurisdiction (via the jurisdiction's DataHub).

2. The second use case is evaluators who already have data on a specific cohort of individuals and want to link those data to additional observational data on a jurisdiction's DataHub.

3. The third use case enables jurisdictions (states, counties, cities) to evaluate their own programs and expenditures.

4. The fourth use case is jurisdictions that would like to partner with an evaluator to evaluate their own programs and expenditures.

5. Future versions of AISP Innovation may enable secure linking of records across existing jurisdictions (i.e., across multiple DataHubs), opening numerous analytic possibilities, including the ability to follow individuals who have migrated outside of a jurisdiction and to utilize vast amounts of data to produce studies of rare events with much more statistical power.

6. Future iterations of the system may also allow data from DataHubs, with jurisdictional permission, to be securely linked to federal data systems like those housed at the U.S. Census Bureau. Similar technology is already used elsewhere. See the below section entitled *Linking across a distributed network* for more information.

V. System Workflow

To describe the system workflow, we take the example of "Washington County," a fictitious jurisdiction that is looking to fulfill Use Case 4, the ability for a government to partner with an evaluator to assess one of its own programs. Washington County has a custom IDS that it uses to provide case management services for at-risk populations. This in-house IDS is operated by the County's Health and Human Services Department. The IDS pulls data nightly from three County agencies and provides real-time data to counselors who work with disadvantaged youth across the County. In an effort to evaluate several of its education and public safety initiatives, however, the Director of Washington County's Health and Human Services has requested access to AISP Innovation.

The workflow is discussed from two points of view. The first is from Washington County's perspective as they install and configure AISP Innovation. The second point of view is that of an external evaluator, "Dr. Jane," who is using the system to evaluate a taxpayer-funded intervention.

A. Washington County's Workflow

Washington County has a custom IDS solution supported by three full-time employees and a Director, who every year must defend the IDS budget to representatives of the state legislature. In an effort to increase revenues while showcasing the public policy value of their data, the Director works with AISP to join the AISP Innovation network.

AISP, which receives grant funding to provide technical assistance to incoming jurisdictions, works with the County to adopt the appropriate data governance regime. This human component is vital to the success of the system. Assuming Washington County does not have the expertise, AISP will support their work to develop a governance structure that will enable the translation of policy into practice, assess possible programmatic shortcomings, and convert these initial findings into data-driven RFPs to which evaluators can respond.

Because the development of AISP Innovation was financed by a foundation, Washington County receives the DataHub at little to no cost. The DataHub is a standalone server that does not integrate directly with the existing IDS. This means that legacy systems are not part of the DataHub. Instead, with technical assistance optionally provided by AISP, the County database administrator builds separate databases by adopting the AISP Innovation data standards.[5] These databases are cross-sectional and include data from multiple agencies spanning multiple domains. These data are imported to the DataHub and configured through the Clearinghouse, which ensures that the County's data are formatted to a set standard.

Having joined the network, Washington County will continue to receive technical assistance as requested.[6] Immediately, the County begins to communicate with a pool of talented evaluators and statisticians who have registered with AISP Innovation. To begin, Washington County would like to evaluate the effect of an early childhood intervention on the health outcomes of a select cohort of third graders.

The County posts an RFP to the Clearinghouse, and three research teams respond with research designs, a budget, and a timeline. The County selects "Evidence Research Inc." as its evaluator based on their prior experience and enters into an agreement with the evaluator using the standardized contractual and data use agreements available on the Clearinghouse.

B. Dr. Jane's Workflow

Using the administrator panel on the Clearinghouse, Washington County specifies the appropriate access control and credentials to select users. Now Dr. Jane from Evidence Research Inc. can query and analyze data by way of the Analysis Console on the Clearinghouse API (*Figure 2*). The API enables Dr. Jane to issue queries and analyze data on the Washington County DataHub in several industry-standard statistical languages.

5 These data standards are defined by the AISP Innovation Data Standards expert panel. See Wulczyn et al., 2017.
6 The extent and feasibility of funding ongoing technical support is discussed in the AISP Innovation business plan, which is being developed concurrent with this report.

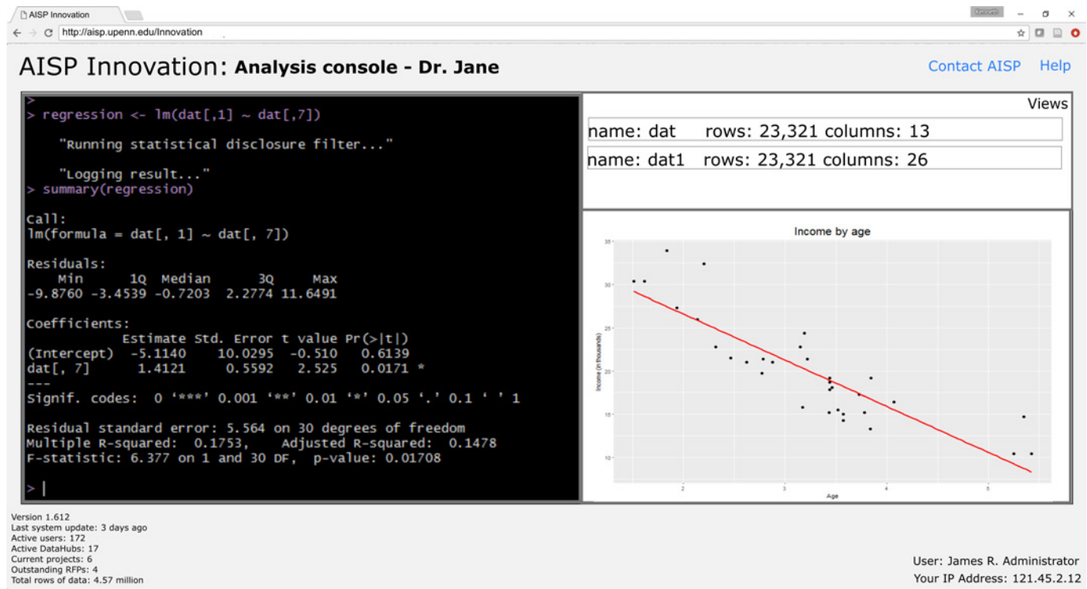


Figure 2: The Analysis Console as part of the Clearinghouse.

She begins by creating a data view of all third graders, their demographic information, data about their home lives, three select health outcomes, and whether or not they participated in the early childhood program. Dr. Jane's data access is limited by the specific statistical disclosure restrictions imposed specifically for this project by Washington County administrators. At no point is Dr. Jane able to download the data from the Washington County DataHub.

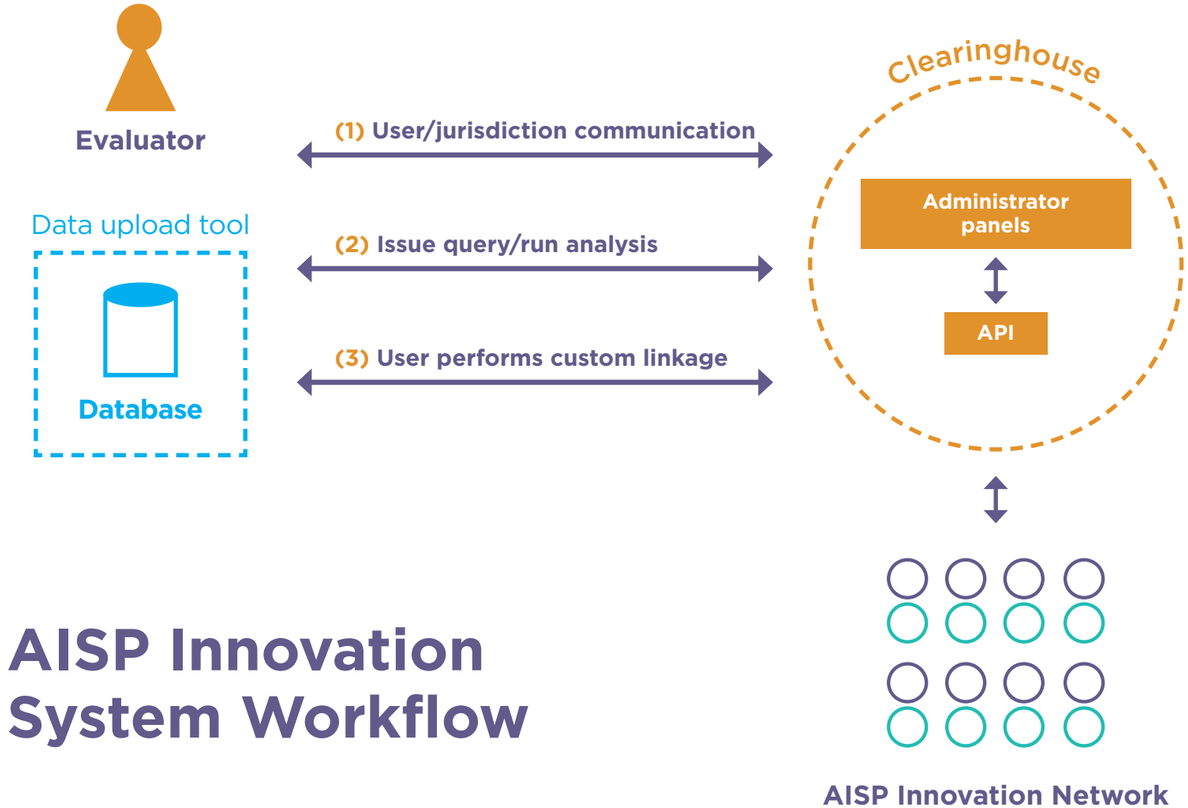
Meanwhile, AISP Innovation platform has logged every query and line of code that Dr. Jane has submitted to the system for audit purposes, if such a review is deemed necessary.

Dr. Jane is satisfied with her cohort, and issues code to the Washington County DataHub to run a regression estimating the causal effect of the early childhood intervention on health outcomes. The DataHub estimates the regression and returns the results to Dr. Jane via the API. She then writes up her findings in a report that is shared with Washington County. Further communication follows, including the fulfillment of contractual obligations. In the meantime, Washington County can communicate the results of the analysis to policy stakeholders and politicians, who can then adopt needed changes to the program to make it more effective.

Figure 3 displays the technical workflow of the system. In the case of Washington County, the jurisdiction took the initiative to engage with evaluators. However, Use Cases 1 and 2 suggest the opposite is possible, that evaluators would request and pay jurisdictions for access to their data.

Figure 3 also illustrates the technology behind Use Case 2. Here we can suppose that Dr. Jane has collected data on a randomized control trial she administered at Washington County's School District. She has outcomes of the experiment and would like to link these results to demographic data on the County's DataHub. She can use the encrypted "Data Upload Tool" to link her data with data on the DataHub and then analyze both data sets jointly via the Clearinghouse API.

The ability for AISP Innovation to support query and analysis of data without ever actually having the data will be a crucial innovation and will clear a significant technology hurdle. The next section details the specific technological suggestions that have been developed by the AISP Innovation Technology and Data Security Expert Panel to achieve this goal.



AISP Innovation System Workflow

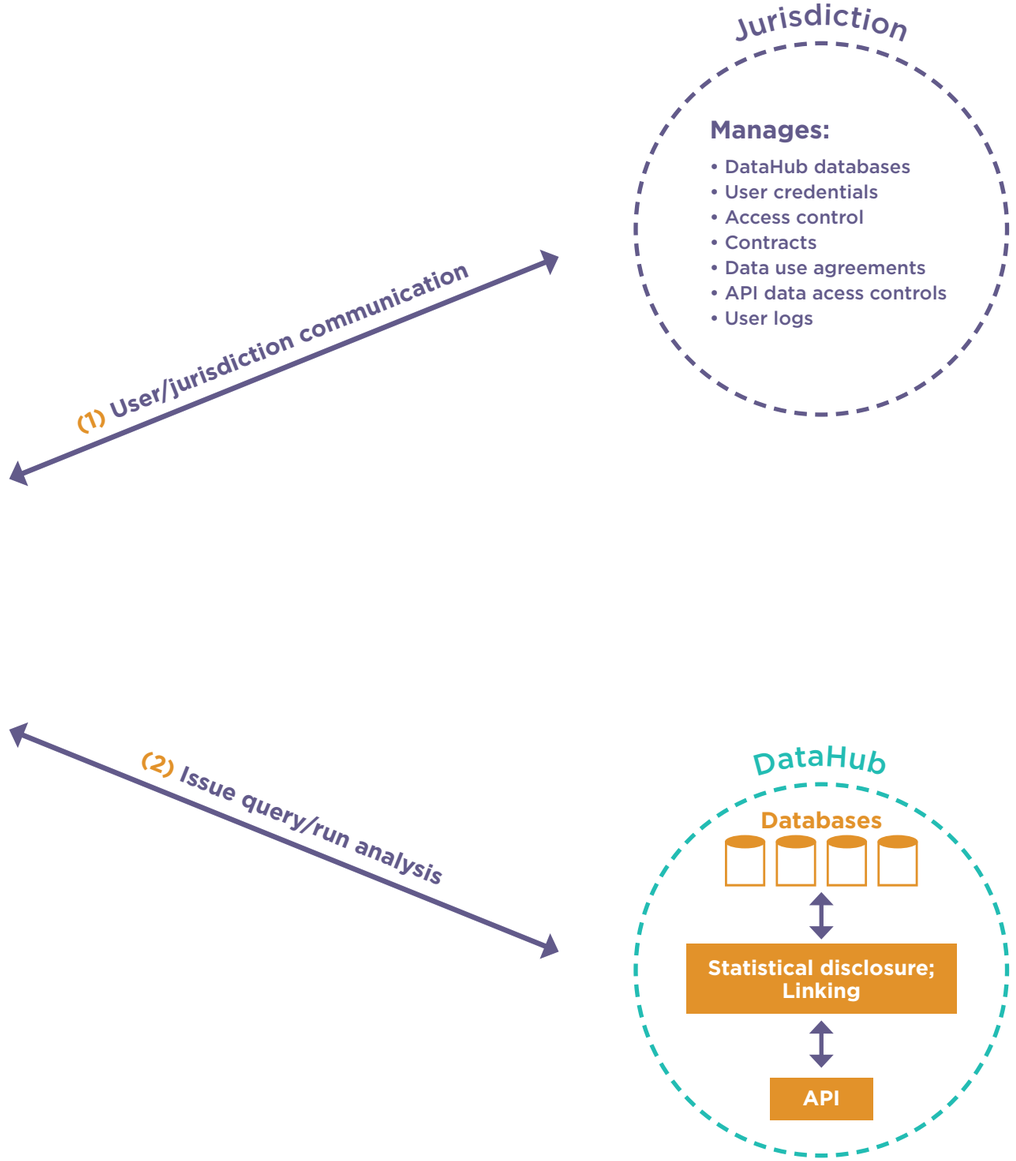


Figure 3: AISP Innovation system workflow.

VI. Design and Technical Components

This section provides some additional technical details on the major privacy and technological components of the AISP Innovation system. While this paper is written with some specificity, the goal is to work with an eventual developer to design each of the components in greater depth. Initially, AISP Innovation was conceptualized to be built entirely on open-source software, not only to keep costs down but to make it more efficient to deploy and maintain as more jurisdictions join the AISP Innovation network. The eventual developer may decide, however, that a commercial or hybrid solution is more flexible and cost-effective.

We divide the following section into security objectives, security features (which include administrator tools, encryption, insider threats, automatic updates, and intrusion detection and prevention), data linking, statistical disclosure control, and, finally, Clearinghouse and API specifications.

A. Security Objectives

We have three objectives when it comes to security: The first is *integrity*, aimed at protecting AISP DataHub instances and the AISP Clearinghouse from unauthorized modification. The second is *confidentiality*, protecting AISP DataHub instances and the AISP Clearinghouse from unauthorized data access. The third is *availability*, to maximize access to AISP DataHub and the AISP Clearinghouse for authorized users within the aforementioned constraints.

These objectives are aimed at overcoming a very specific threat model. The system will handle sensitive data, and an unauthorized user (hacker, etc.) might try to gain access to these data. There are three primary concerns: (a) An unauthorized user could breach the DataHub using some kind of vulnerability in the software or the operating system; (b) The unauthorized user could listen to the communication between the DataHubs; or (c) An individual could steal the credentials from an authorized user. These scenarios are addressed below.

We would like the eventual developer to consider the following security frameworks to counter these threats:

The system will utilize a “defense in depth” strategy by deploying controls in layers and leveraging best practice security architecture frameworks. Our baseline security frameworks draw upon the Information Security Management Standards published by the International Organization of Standards under publication ISO 27001. In addition to ISO 27001, which defines a set of security building blocks, several other ISO and federal standards may apply to different aspects of the system. In particular, the following additional ISO publications will provide the best practice guidelines for the effort:

- ISO/ IEC 27002 Code of practice for information security management
- ISO/ IEC 27003 Guideline for information security management implementation
- ISO/ IEC 27004 Guideline for information security management measurement and metrics framework
- ISO/ IEC 27005 Guideline for information security risk management
- ISO/IEC 27017:2015 Guideline for cloud computing
- ISO/IEC 27018:2014 Management of personal data in the cloud

- ISO/IEC 27031 Guideline for information and communications technology readiness for business continuity
- ISO/IEC 27033-1 Guideline for network security
- ISO 27799 Guideline for information security management in health organizations
- ISO/IEC 27034 Guideline for application security

In addition to following the security architecture best practices outlined above, the development will draw upon the security control objectives and requirements under the National Institute for Standards and Technology and Health Insurance Portability and Accountability Act guidelines:

- National Institute for Standards and Technology publication 800-53 outlining the controls required under Federal Information Security Management Act of 2002
- Security Standards for the Protection of Electronic Protected Health Information, 45 CFR Part 160 and Part 164, Subparts A and C, that were adopted to implement provisions of the Health Insurance Portability and Accountability Act of 1996
- State statutes that may apply to some kinds of data

The AISP DataHub would be implemented under different network configuration contexts at various sites. With guidance from AISP Network Sites and the AISP Innovation Technology and Data Security Expert Panel, the AISP Innovation platform developer will establish a standardized network threat management architecture for the AISP DataHub. The network and telecom security approach would address requirements related to:

- Separation of network segments for the AISP DataHub at the jurisdiction; a distinct network segment would allow tailored network access policies to be defined and implemented.
- A distinct security zone for the DataHub with dedicated firewalls to segregate the AISP DataHub from the rest of the network at the site; by limiting the support for certain network protocols, services, and firewall ports, a distinct security zone can enable a higher security posture of the AISP DataHub than of the rest of the jurisdiction’s network.
- Alternate firewall architectures that can be adopted by the jurisdiction to implement the AISP DataHub; the firewall architecture would include network layer, session layer, and application layer to screen incoming network traffic.
- Alternate proxy server configurations to disallow sources from establishing communication sessions with AISP servers directly.
- Specific protocol support for establishing remote server connections with the AISP DataHub.
- Network security for the AISP Clearinghouse.

In addition to defining physical access controls for the AISP DataHub, the AISP platform development team will conduct security risk analysis and document requirements for physical safeguards for the AISP Clearinghouse.

B. Security Features

With guidance from the AISP Innovation Expert Panel as well as network sites, the AISP platform system developers will consider alternative physical scenarios under which the AISP DataHub is likely to be implemented at various sites and the associated risks and potential countermeasures. As part of this implementation, the system will consider the following components:

1. Administrator Tools

The system should provide two separate administrative interfaces—one for DataHub administrators and one for Clearinghouse administrators—both accessible via the Clearinghouse. The former is used to manage data access and to configure individual DataHubs; the latter is used to manage user accounts and for monitoring.

When a new jurisdiction wishes to join the system, the new DataHub administrator creates a cryptographic key pair and sends the public key to the Clearinghouse. Alternatively, for a less sophisticated administrator, the keys could be generated by the Clearinghouse and the private key mailed to the DataHub administrator, e.g., on a USB key. The private key is installed in the DataHub and protected by a passphrase. Thus, if the Clearinghouse is compromised by an external breach or an insider, the unauthorized user does not gain access to the DataHub. Similarly, an unauthorized user who gets physical access to a DataHub still cannot access the private key.

The administrator of a DataHub can connect to an administrative interface for the DataHub. Requiring the administrator to connect to the DataHub via the Clearinghouse should limit possible vulnerabilities. The login is the passphrase of the administrator cryptographic key. Once logged in, the administrator can perform the following actions (and potentially others):

- Manage local data sets (e.g., add new data, delete existing data)
- View data access requests by remote users (which the DataHub would periodically download from the Clearinghouse) and potentially approve them by signing them with a private key
- View active access permissions and potentially modify or delete them
- Create other local users and delegate some or all of the administrator privileges to them (internally this would cause the creation of additional key pairs, which would be endorsed by a signature with the original administrator's private key)
- View audit logs that contain all queries by approved remote users, as well as all actions by local users (more on this below)
- Flag data access requests or queries for further investigation by the Clearinghouse (perhaps the administrator can enter a brief description of the concern)
- Configure key parameters of the DataHub, such as network access (internet protocol, gateway, etc.)

- Set parameters for local anomaly detection (limits on the number of queries per user, the number of accesses per data set, the requested cell size, etc.)
- Obtain access to a local terminal for diagnostics maintenance (if necessary)

The DataHub records all queries by approved remote users, as well as all actions by local administrators; the log records are cryptographically signed with the key of the person who initiated the corresponding action, so the actions are undeniable. Users and administrators should be informed that the system logs all of their actions, and should have to agree to this audit function as part of their user agreement. Since the logs contain only human-initiated actions (e.g., "User Alice requested access to data set XYZ with parameters A=5 and B=9" or "User Bob ran query ABC on data set XYZ," and *not* fine-grained system-level events, such as "The DHCP server assigned the DataHub IP address 1.2.3.4"), the records should be small enough to be retained indefinitely.

The Clearinghouse administrator can connect to an administrative interface on the Clearinghouse using a similar process as above. (If the interface is accessible over the network, two-factor authentication should be used.) The administrator should be able to perform the following actions (and potentially others):

- Add new DataHub sites into the system, or remove existing sites;
- Add new users (after the vetting process has been completed), or remove existing users (e.g., after inactivity or violations of the user agreement);
- View summaries of current activity, such as active queries, a list of DataHubs and the last time they connected to the system, etc.;
- Delegate privileges to other Clearinghouse administrators, or revoke such delegations;
- View flagged requests or queries as reported by the DataHub for further investigation;
- Configure the parameters of the Clearinghouse-level anomaly detection (perhaps number of queries per user per day, number of DataHubs contacted, number of requests made, etc.); and
- View detailed logs of user and DataHub activity.

As with the DataHub, the Clearinghouse should maintain logs of all activity, including all user queries and all interactions with the DataHubs, and these logs should be kept indefinitely, if possible.

The Clearinghouse administrator should periodically run audits, and basic safety checks should be in place (e.g., a tool that alerts an administrator if some user asks an unusually large number of queries within a short time; if a user has too many failed login attempts; if a user tries to access data sets he or she isn't allowed to see, etc.). The logs should be archived; thus, in the event that some vulnerability exists but is discovered much later, it remains possible to assess the impact of the problem. We also anticipate a series of tools that will allow the system to undergo regular backups and secondary storage disposal. If a jurisdiction leaves the network, it is possible, perhaps at their request, that these logs could be purged.

2. Encryption

We envision several forms of encryption, depending on the operation.

a) Data at rest

When a new data set is imported into a DataHub, the DataHub should immediately generate a new cryptographic key, encrypt the data set with this key, and store the data only in encrypted form. (For efficiency, this step should use symmetric encryption, e.g., AES.) The key itself should then be encrypted with the private key(s) of the administrator(s) who can approve access to this data set. (Note that this would require storing multiple encrypted keys, one for each administrator, but the data themselves have to be stored only once.) Thus, an unauthorized user who gains physical access to the DataHub cannot get access to the data—the symmetric key is encrypted with the administrator's private keys, and the administrator's keys themselves are protected by passphrases, which the unauthorized user does not know. Even if the unauthorized user colludes with an administrator, the user can gain access only to the data sets that this administrator personally has access to.

b) Data in transit

When the data need to be sent to the Clearinghouse, they are sent only in encrypted form. Once the transfer is complete, the DataHub encrypts the relevant symmetric key, along with a hash of the data, using the Clearinghouse's public key, and it signs the result with the administrator's private key. Thus, the Clearinghouse can (a) use the hash to verify that the data were not altered in transit, (b) use the signature to verify that the data are genuine, and (c) actually decrypt the data using the provided symmetric key. The Clearinghouse immediately erases the symmetric key after decrypting the data, and erases the data themselves once they are no longer needed. To the extent possible, the key and the data should be kept in memory at all times at the Clearinghouse and should not be written to persistent storage. If the data are lost (e.g., after the failure of a Clearinghouse node), they can always be fetched again from the relevant DataHub.

c) Requests and approvals; other command traffic

All communication between DataHubs and the Clearinghouse, as well as between the Clearinghouse and the analysts, should be protected with strong encryption, and both endpoints should be required to authenticate. This ensures that an adversary cannot impersonate a DataHub or the Clearinghouse, and that an adversary cannot read the messages that are being exchanged (e.g., to learn what queries are being asked, or query results). As a fail-safe, if one of the endpoints cannot authenticate itself by presenting a valid cryptographic certificate, the other endpoint should terminate the connection.

Developers should not attempt to implement their own cryptographic protocols; experience indicates that it is very easy to make subtle mistakes that compromise security. Instead, developers should rely on widely used standards and common libraries whenever possible. One good candidate would be Transport Layer Security (TLS) with perfect forward secrecy, as implemented, for example, in the OpenSSL library.

3. Insider Threats

This section reviews the technical safeguards that should be put in place to address possible internal data security threats. But, it is important to note that there are multiple layers of security in place within an IDS, only one of which is technical. A strong work culture, careful hiring processes, and frequent audits are the most important safeguards to prevent threats from an insider. Prior to any interaction with identifiable data, a staff member should be required to sign a non-disclosure agreement (NDA) and undergo data security training that includes an overview of the legal repercussions of security incidents.

a) Malicious DataHub administrator

The administrator of a DataHub is able to access all data uploaded to that DataHub; this access is acceptable, since the administrator must have had access to the raw data in order to upload them. The administrator is also able to approve any and all access requests to these data. However, the administrator cannot access the data on another DataHub without signing up for a user account, submitting an access request, and getting the request approved by the administrator on the other DataHub. The administrator can see all of the requests and queries that are submitted by analysts registered with the DataHub in the administrator's charge, but is not able to impersonate an analyst (since the administrator does not have access to the analyst's private key). If the DataHub participates in an advanced linking query (via MPC or PSI-CA), the administrator can learn the result of the query, but cannot learn anything about the data on the remote DataHub that is not already implied by the query result.

b) Malicious Clearinghouse administrator

A malicious administrator at the AISP Clearinghouse could perform a denial-of-service attack on the system, by refusing to create accounts for analysts or by refusing to forward requests and approvals or to process queries. An administrator also potentially has access to any data product that is communicated to the Clearinghouse for analysis after the malicious administrator gains control over it. However, crucially, an administrator cannot impersonate another DataHub administrator, since each administrator only has access to his or her own private key, and cannot gain access to any data on a DataHub except through the normal request and approval process.

c) Malicious analyst

While all analysts must pass a thorough vetting process, in theory, a malicious analyst could request access to sensitive data or submit carefully crafted queries in order to compromise the privacy of individuals. However, all queries and requests would require approval of the corresponding DataHub administrators, and each action would be logged and undeniably linked to their credentials. Once the actions are discovered (either through anomaly detection or by an administrator audit of the query logs), it is possible to assess the damage from the query logs (for possible mitigation) and to take action against the analyst.

4. Automatic Updates

Software tends to "deteriorate" over time as vulnerabilities are discovered and repaired in new releases and as new protocols become available. Therefore, it is important to have a process in place to ensure that all the machines are up to date at all times, and to make it very difficult to operate a DataHub that is not secure.

We assume that the Clearinghouse will be administered by an information technology (IT) professional at all times, whereas the DataHub administrators may have a wide range of technical expertise. For this reason, we recommend that the DataHub software be centrally managed. In the event that a jurisdiction does not have this expertise, it would be provided as part of the technical support package offered to a jurisdiction when they join with AISP Innovation.

For instance, the IT professional at the Clearinghouse could, whenever security patches or new software versions are released, assemble a virtual machine (VM) image that contains all the software the DataHub will need—perhaps a minimal Linux installation and the DataHub software itself. The professional could then attach a timestamp and a version ID to the image and cryptographically sign it with a special private "software key" that is maintained at the Clearinghouse (offline, perhaps in a lockbox). The IT professional could then upload this image to a clearinghouse server, along with a list of DataHub IDs and the VM image ID each DataHub is supposed to use.

The DataHub nodes themselves could initially contain only a minimal operating system installation, a VM monitor, a DataHub ID, the public key of the Clearinghouse's software key, and a small updater tool. When the node starts, the updater could connect to the Clearinghouse, check the list to see which VM image it is supposed to use, and then, if it does not already have a copy of that image locally, download the image, verify the signature and timestamp, and (if the signature is valid, the image version is newer than that of the current image, and the timestamp is not older than, say, a few weeks) store and run the VM image. Periodically, the uploader could check the Clearinghouse for new images; if a new image is available, the uploader could download that image and then, perhaps at some point during the night, shut down the VM and replace the image.

With this approach, the DataHubs always run current software with the latest patches, without any special action by the DataHub administrator. They are also protected against unauthorized users who try to compromise the update process (because the DataHub would not accept an old or obsolete image or one that is not properly signed). The nightly updates would minimize disruptions to user queries, and the central mapping of machine IDs to VM images could be used to gradually roll out new images.

5. Intrusion Detection and Prevention

If an intrusion detection approach were deemed necessary, an effective system would go beyond logging to include automated processes to monitor events occurring in the AISP DataHub and the AISP Clearinghouse and to analyze them for signs of possible violations of security policies and standards. The prevention capability would alert system administrators of possible violation and would, in some cases, automatically modify firewall rules to stop the intrusion.

The intrusion detection and prevention protocols would include the following requirements:

- Signature-based detection—Signature-based detection would allow the AISP data infrastructure to detect threats that have already been identified by the security product vendors. This capability would excel at stopping known threats but would not be effective against new threats.

- Stateful protocol analysis-based detection—This capability would allow the AISP data infrastructure to monitor current data access patterns with predefined profiles of what constitutes "normal" access for each type of network protocol (HTTPS, SFTP, etc.). Any data access that does not meet the profile of normal access for a particular protocol is considered a potential threat. This capability can be effective in identifying new as well as known threats.

- Anomaly-based detection—Anomaly-based detection would allow AISP data infrastructure to use learning algorithms to determine what constitutes "normal" access patterns and flag any access pattern that does not seem normal. This approach would be able to detect unknown types of threats without a predefined profile. This is advanced functionality that could be saved for future releases.

It will be important to define and fine-tune the rules and thresholds for intrusion detection. A threshold that is too high would lead to some legitimate traffic being flagged as malicious, whereas too low of a threshold would lead to some malicious traffic passing by as legitimate. The team would define standard scenarios that could be used by all AISP DataHub sites to maintain a uniform intrusion detection posture.

C. Data Linking

1. Linking data across databases in a single DataHub

Each DataHub installation contains all of the components necessary for a jurisdiction to link and analyze their own data. The components allow a jurisdiction to:

- Authorize and manage users through the Clearinghouse;
- Validate conformity with the AISP data model;
- Register other data from their jurisdiction as a source "external" to the data model for special case analyses;
- Construct links and joins with the statistical properties they desire[7];
- Apply a variety of statistical disclosure controls, if needed; and
- Analyze their own data using their preferred statistical package through the DataHub API.

Many of the specifics of these features are discussed in detail below where the use cases demand more stringent protections.

2. Linking external data to data on the DataHub

Jurisdictions need a secure way to upload their data to the DataHub. In addition, Use Case 2 (i.e., evaluators who wish to use their own data) allows researchers to securely link their own custom cohorts to data on a DataHub at a given jurisdiction. This is the above instance where Dr. Jane has collected data on a randomized control trial and wishes to link these results to demographic data on Washington County's DataHub.

From a security perspective, it is best if the data linking mechanism is simple and has relatively few code paths. Adding special cases often leads to an exponential increase in the number of code paths that have to be tested. Simplicity will also reduce the development effort. This suggests that the system should offer only one general way to contribute a data set. This could work as follows:

- The data owner (institution or external researcher) downloads the software package from the AISP website to a machine he or she controls. In the case of an institution, this could be a dedicated workstation; in the case of a researcher, this could simply be the researcher's laptop.
- The data owner installs the software and enters the credentials (password, etc.) received from the AISP registration.
- The software connects to the Clearinghouse to verify the credentials.

7 Respondents to our request for proposals should present a plan for comparison of identifying variables in steps to develop an effective matching algorithm based on the available data from which cutoff scores can be derived and selected. All relevant statistics from the confusion matrix should be presented to aid in the decision. See Capuani et al., 2014. Assistance by participating jurisdictions will be provided as needed in evaluating the algorithm.

- The data owner copies his or her data to the DataHub.
- The data owner configures the software to tell it the location of each data set and a description of the data.

At this point, the data could (with proper authorization) be used to construct views, whether it is an external data set from a researcher or an institutional data set. The researcher would not usually authorize others to access his or her data, whereas the jurisdiction typically would. In the case above, Dr. Jane can then submit a request for authorization to access two data sets—her own and the one she wishes to link with. Once the jurisdiction (in this case, Washington County) approves, the data view is created as usual.

3. Linking across a distributed network

The ability to link multiple databases while maintaining privacy may seem unintuitive. We note that the technology for linking data in a completely distributed fashion already exists. One way to accomplish this is using secure multiparty computation (MPC). Briefly, MPC provides a way for N parties (say, different DataHubs) to collectively evaluate a function $f(D_1,D_2,...,D_N)$ on a set of private inputs ($D_1,D_2,...,D_N$, say, private data sets contributed by each party) such that no party learns anything about the inputs provided by the other parties other than what is implied by the output of the function. MPC is computationally expensive, but MPC technology has made substantial progress in recent years, and there are now practical solutions at scale. For instance, Hemenway et al. (2016) describe a system that uses MPC to detect possible collisions among communications satellites.

Another possible approach relies on specialized cryptographic primitives, such as private set intersection cardinality (PSI-CA). For instance, the DJoin system (Narayan et al., 2012) is able to process a variety of SQL-style JOIN queries over distributed data. Suppose DataHub A contains a set of medical records, and DataHub B contains a set of travel records; suppose further that a researcher would like to know how many patients have traveled to a certain country and have subsequently been treated for malaria. DJoin would use PSI-CA to answer this question in a completely distributed fashion (that is, unencrypted data would never leave either of the two DataHubs), and it would also maintain differential privacy on the results, should we decide to implement differential privacy in the future. Compared to MPC, this approach is less computationally expensive and could therefore be used with much larger data sets.

D. Statistical Disclosure Control

Since the data are sensitive, it is important to ensure that the responses to queries do not reveal private information about the individuals in the data set. One obvious threat comes from queries that directly ask about sensitive information (e.g., “Return information on persons with specific age, race, and gender, and a particular diagnosis”), and this could be handled by restricting the kinds of queries that are allowed. A more subtle threat comes from queries that reveal information indirectly (e.g., “How many persons with first name ‘Nathan’ live in the following zip code?”), and this could be handled with a concept like k-anonymity, e.g., by preventing queries that have a geographic area or grid “cell size” that is too small. But even k-anonymity does not completely prevent information disclosure, and it can be tricked, e.g., with intersection queries (“How many felons are in the following zip code?” “How many felons are in the following zip code who are not named ‘Nathan’?” and then compute the difference), or with auxiliary information (“How many felons are in the following zip code?” when the adversary knows that there are seven, but isn’t sure about “Nathan”).

Disclosure controls are intended to prevent the possibility of re-identification through linking of disclosed data with publicly available data. The Disclosure Limitation Rules Engine and the Statistical Disclosure Controller components would work together to provide some of the following functional capabilities as determined during requirements analysis:

- Functionality to register, define, and apply statistical controls to minimize the probability of re-identification from de-identified data
- Functionality to register specific information types as “sensitive information types”
- Functionality to apply “threshold rules” to determine the minimum frequency counts for safe disclosure (for each type of sensitive information)
- Functionality to apply cell suppression rules to result sets
- Functionality to alter result sets by adding “noise,” such as by using random rounding and rank swapping while retaining logical consistency between attributes; logical consistency would ensure that rank swapping does not create anomalous results
- Functionality to add “noise” by altering categorical data using Post Randomization Method
- Functionality to create synthetic result sets, which would include arbitrary data points on a model fitted to an actual cohort; the arbitrary data points on the fitted model would not belong to a real person

The use of each protection mechanism should require authorization; for instance, Dr. Jane could be allowed access to data sets X and Y with cell sizes up to 10, while user Bob could be allowed access to data set X with cell sizes up to 5, and user Charlie could be given direct access to data set Z (e.g., because he works for the jurisdiction that owns the data set). This format makes it possible to issue data in the most private form (person-level data, for instance) and the most public form (data aggregated to the census tract level, for instance).

We also envision a person-level synthetic data set that allows users to test their queries. This data set would be standard across each DataHub. Imagine that a user wants to calculate a new view, such as “felony convictions by age.” We wish to allow this view to be generated while not giving access to the original person-level data that the user would otherwise use to verify that code ran as intended. Instead, the user would generate this view using a synthetic data set, which would allow code to be checked without viewing actual private data. We realize this functionality may not be perfect, particularly in instances where the researcher links external data to a DataHub.

The most comprehensive data protection is offered by a technique like differential privacy. While differential privacy is still in the research and development stage, we see it as the future of statistical disclosure, and hope to deploy these algorithms in future iterations of AISP Innovation. We would therefore like development to include “plug and play” privacy mechanisms, which would include differential privacy.

E. Clearinghouse and API Specifications

The Clearinghouse is the central component of the proposed technology infrastructure. We discuss it throughout this report, but reiterate more generally in this section. It is the standard by which all users, including evaluators and DataHub administrators, interact with the AISP Innovation system.

The consistency allows for the centralization of all privacy and security protocols and helps ensure that the entire system can be managed by an AISP system administrator. The Clearinghouse has three primary purposes.

First, it is a public-facing web presence and internal management information system. Public users can read about the latest evaluations and read about evaluation and planning projects posted by participating jurisdictions. Internal and external evaluators can view metadata for available data sets, apply for access, and respond to posted projects. Credentialed users can interact with a given jurisdiction about projects and negotiate contracts and data use agreements. They can also submit project write-ups via the Clearinghouse and, with permission, share results with the broader community.

The second critical function is to help jurisdictions manage their respective DataHubs. This is the functionality demonstrated in *Figure 1*. To maintain the integrity of the system, every DataHub must be configured to communicate with the Clearinghouse using the exact same protocols. Much of this is discussed above in detail. When a new DataHub comes into the system, the administrator can use the Clearinghouse to ensure that data are formatted correctly (field names, etc.). Once this has been accomplished, there should be no need for additional supervision of the data on the DataHub until the database is updated with new data. Data updates may happen annually, semi-annually, or more often. Finally, using the aforementioned administrator tools, DataHub administrators can manage user access and review user logs.

The last major Clearinghouse use case is analysis. Once an evaluator has been credentialed to access a given DataHub, he or she can begin to use the data. To do so, the evaluator accesses the Analysis Console via the Clearinghouse as illustrated in *Figure 2*, which allows queries to be issued, data views to be created, and data to be analyzed. We envision a series of APIs for bridging the DataHub and the Clearinghouse for these purposes.

The API would make it possible for evaluators to access statistical coding languages like R, SAS, and Stata.[8] When a data query is issued by the analyst interfacing with the Clearinghouse, the API makes a call to the DataHub, which compiles the correct data view and then verifies that the user is in fact credentialed to analyze data at the given resolution. If not, the API reports back to the user that the query has been denied. If so, the data view is created. The user is now free to analyze that data view via the Clearinghouse API. All calculations occur on the DataHub, and at no time are any data released to the analyst. The DataHub then issues a response that is communicated to the analyst via the Analysis Console. We leave it up to the eventual developer to suggest how this API should be structured and how the user interface would be designed. The hope is that the design resembles an integrated development environment, similar to R Studio.

VII. Conclusions

The technical challenges of developing an IDS are significant, but as this paper indicates, not insurmountable. IDS sites across the United States and the world have demonstrated that administrative data can be utilized to inform policy and practice utilizing carefully constructed design and technical components that emphasize security, administrator tools, encryption, processes to inhibit threats from unauthorized uses, advanced data linking methods that inhibit viewing identifiable data, and statistical disclosure control. The proposed AISP Innovation Clearinghouse and DataHub technology could provide access to these legal, procedural, and technical best practices in a cost-effective format for municipalities and agencies.

8 Statistical packages like SAS and Stata that are not open source would require additional licensing.

References

Capuani, Ligia, Ana Luiza Bierrenbach, Fatima Abreu, Pedro Losco Takecian, João Eduardo Ferreira, and Ester Cerdeira Sabino. (2014). Accuracy of a probabilistic record-linkage methodology used to track blood donors in the Mortality Information System database. *Cadernos de saude publica*, *30*(8), 1623-1632. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4863984/

Hemenway, Brett, Steve Lu, Rafail Ostrovsky, and William Welser IV. (2016). High-precision secure computation of satellite collision probabilities. In *Proceedings of the 10th International Conference on Security and Cryptography for Networks*, pp. 169-187. New York: Springer-Verlag. https://eprint.iacr.org/2016/319

Narayan, Arjun, and Andreas Haeberlen. (2012). DJoin: Differentially private join queries over distributed databases. In *OSDI '12 Proceedings of the 10th USENIX Conference on Operating Systems Design and Implementation,* pp. 149-162. Berkeley, CA: USENIX Association. http://www.cis.upenn.edu/~ahae/papers/djoin-osdi2012.pdf

Wulczyn, Fred, Richard Clinch, Claudia Coulton, Sallie Keller, James Moore, Clara Muschkin, Andrew Nicklin, Whitney LeBoeuf, and Katie Barghaus. (2017). *Establishing a Standard Data 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