CTO

Office of the
CHIEF TECHNOLOGY OFFICER

LEVERAGING DATA FOR THE NATION’S HEALTH

A vision for inter-agency data sharing for the U.S. Department of Health and Human Services

December 2019

The Data Initiative Team
Office of the Chief Technology Officer
U.S. Department of Health and Human Services

,

TABLE OF CONTENTS

EXECUTIVE SUMMARY

We are facing a fundamental economic shift. Building on the technological advancements of the third industrial revolution, the continuing transformation is characterized by unparalleled innovation and generation of data.[1] As these exponential changes disrupt almost every industry, there are emerging challenges for institutions and governments in their approach to the adoption of these opportunities, the governance of entirely new domains, and the management of systems that can initially be disruptive but which have the potential to provide enormous returns with sustainable investments. In the context of unprecedented processing power, storage capacity, and access to knowledge, the data being generated has changed the business models of entire industries and our collective norms and expectations around the use of data in providing better and more personalized services.

This report presents a vision and approach for data sharing that is consistent with the 2019-2020 Federal Data Strategy Action Plan[2] released by the Office of Management and Budget (OMB) in June 2019. It supports key aspects of the Foundations for Evidence-Based Policymaking Act of 2018,[3] which mandates a "systematic rethinking of government data management to better facilitate access for evidence-building activities,"[4] and the President's Management Agenda, which includes the Cross-Agency Priority Goal of "Leveraging Data as a Strategic Asset."[5] It also responds to the ***American AI Initiative,***[6] which *"directs agencies to make Federal data, models, and computing resources more available to America's AI R&D*

[1] Davis, Nicholas. January 19, 2016. What is the fourth industrial revolution? *World Economic Forum*. Available from: https://www.weforum.org/agenda/2016/01/what-is-the-fourth-industrial-revolution/.

[2] *DRAFT 2019-2020 Federal Data Strategy Action Plan*. 2019. Executive Office of the President of the United States. Available from: https://strategy.data.gov/assets/docs/draft-2019-2020-federal-data-strategy-action-plan.pdf.

[3] Foundations for Evidence-Based Policymaking Act of 2018, H.R. 4174, 115th Cong. (2018). Available from: https://www.congress.gov/bill/115th-congress/house-bill/4174.

[4] Vought, Russell T. July 10, 2019. *Phase 1 Implementation of the Foundations for Evidence-Based Policymaking Act of 2018: Learning Agendas, Personnel, and Planning Guidance*. (M-19-23). Available from: https://www.whitehouse.gov/wp-content/uploads/2019/07/M-19-23.pdf.

[5] Leveraging Data as a Strategic Asset. *Cross-Agency Priority Goals*. General Services Administration & the Office of Management and Budget. Available from: https://www.performance.gov/CAP/leveragingdata/.

[6] Accelerating America's Leadership in Artificial Intelligence. February 11, 2019. The White House Office of Science and Technology Policy. Available from: https://www.whitehouse.gov/articles/accelerating-americas-leadership-in-artificial-intelligence/

experts, researchers, and industries to foster public trust and increase the value of these resources to AI R&D experts, while maintaining the safety, security, civil liberties, privacy, and confidentiality protections we all expect".

The United States Department of Health and Human Services (HHS) has been at the forefront of recognizing the value of data and engages in this economy by fostering data-driven innovations. The Department's 29 distinct agencies and offices produce vast amounts of data on the state of the nation's health and well-being. HHS has undertaken several data-driven decision-making, service delivery, and business process modernization efforts within, and in some cases across, its agencies. In the past two decades, the Centers for Disease Control and Prevention (CDC) has led public health's evolution from monitoring infectious diseases to using data to track the occurrence of many noninfectious conditions, such as injuries, birth defects, chronic conditions, behavioral health conditions, illicit drug use, and environmental and occupational exposures. Among a range of other efforts, the Centers for Medicare & Medicaid Services (CMS) is committed to supporting Open Payments, a national disclosure program that promotes a more transparent and accountable health care system by making the financial relationships among manufacturers, group purchasing organizations, and health care providers available to the public. While HHS has been a leader in the federal government in promoting the use of data, we now stand at the brink of true transformation in the ability to leverage data in ways that can profoundly shape the Department's operations and its approach to managing the delivery of programs.

This document outlines a vision for a fundamental transformation in the way HHS ***internally*** shares, analyzes, and derives new insights by leveraging data across HHS agencies to improve the delivery of its programs and further its mission of providing effective health and human services to the nation. Serving to complement the many ongoing data-sharing efficiencies and initiatives at HHS, this document lays out a structure and process, and the legal and technological considerations, for achieving this future state. It includes strategies for closing gaps in data governance, robust data-sharing and use agreement frameworks, the necessary functional attributes of an overall technical platform that can support data sharing at HHS, and approaches for fostering a culture that values data sharing. Given the advent of modern data management tools and technologies, as well as advances made within HHS related to data governance, a federated approach that enables agencies to share data with each other more efficiently and effectively—while continuing to address data privacy, provenance, and security—is essential. By formalizing legal, organizational, analytical, and technical frameworks for data sharing, the Department will expand the breadth of insights attainable from its data and achieve greater efficiencies, across its 29 agencies and offices.

HHS authored a report[7] detailing the current state of data sharing across the enterprise in September 2018. The report summarized findings from discovery sessions on current inter-agency data-sharing practices with agencies from across the Department. A group of agencies was identified during that activity to outline potential solutions. Those agencies: the Agency for Healthcare Research and Quality (AHRQ), the CDC, CMS, Health Resources & Services Administration (HRSA), and the Substance Abuse and Mental Health Services Administration (SAMHSA) provided additional detail about their needs and played a vital role in forming the vision for the future state of data sharing detailed in this document.

[7] *The State of Data Sharing at the U.S. Department of Health and Human Services.* September 2018. Office of the Chief Technology Officer, U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf

INTERNAL DATA SHARING: BACKGROUND AND INTRODUCTION

HHS has a strong foundation of data capabilities to address mission-specific goals. In some cases, these capabilities have existed and operated for decades. The vision laid out in this document is designed to leverage these capabilities— including existing models of governance and data councils—and engage thought leaders and stakeholders.

The 2018 State of Data Sharing report documented existing data-sharing processes, challenges and best practices in considerable detail. Based on this work, the report identified the following major issues that must be addressed to increase the effectiveness and efficiency of data sharing within HHS:

Statutory and **Regulatory Environment:** Numerous statutes, regulations, and policies govern, and sometimes limit, the collection of and access to data. Some of these limits are designed to protect the privacy and confidentiality of individuals. Implementation of the vision for data sharing described in this report will be conducted within exisiting regulatory and statutory frameworks, while exploring opportunities for updating regulations, guidance, or policies where appropriate.

Risk of Disclosure: The risk of identifying individuals through indirect identifiers, and violating their privacy, increases as more variables and more granular data are collected and shared. Restricting access to microdata reduces this risk but is sometimes applied too broadly. The vision for data sharing laid out in this report promotes access to data at the right levels of scale, scope, and granularity to mitigate the risks of disclosure while simultaneously facilitating data sharing.

Norms: Some HHS data stewards may not see a demand for sharing restricted and nonpublic data. Others believe publicly available data is sufficient for most analyses, or they view data-sharing requests as an ad-hoc or special activity.[8] Finally, some data stewards have concerns about potential misrepresentation of government data.

Resource Constraints: Many agencies lack sufficient resources to promote available data or process requests for data. In agencies where data sharing is not a primary function, it may be a secondary consideration.[9]

Process for Data Access: While some HHS agencies leverage their own data extensively and some have internal processes for data sharing, others (and by extension, the Department) lack consistent and standardized processes for one agency to request data from another. In addition, few avenues exist to respond to or remedy a denied request or resolve delays in fulfilling a request.

This report addresses these issues in an effort to improve inter-HHS agency data sharing in a way that leverages existing capabilities and considers mission focus and efficiency. The report is organized into the following primary focus areas:

- **Data-Sharing Culture**
- **Data-Sharing Processes**
- **Data-Sharing Enabling Technologies**
- **Data-Sharing Regulations and Privacy**
- **Data-Sharing Organizational Structure**

[8] *The State of Data Sharing at the U.S. Department of Health and Human Services.* September 2018. Office of the Chief Technology Officer, U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf

[9] *The State of Data Sharing at the U.S. Department of Health and Human Services.* September 2018. Office of the Chief Technology Officer, U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf

DATA-SHARING CULTURE

Across its 29 agencies and offices, HHS has a mission-based culture driven by shared responsibility for improving the health and well-being of the American people. HHS recognizes the potential of the transformative power of data, and internal HHS data sharing is an essential element of many of the Department's core functions. For example, the CDC and the U.S. Food and Drug Administration (FDA), both HHS agencies, share information to ensure the safety of the U.S. food supply. CDC has primary responsibility for surveillance of outbreaks of foodborne illness, while the FDA acts on CDC's surveillance data to manage outbreak response.[10] These agencies also share data with the National Institutes of Health's (NIH) National Library of Medicine (NLM), which hosts an analytic platform that enables these agencies to compare pathogen sequence data with samples taken from food, the environment, and human patients. The platform quickly reports similarities to public health investigators, so they can trace outbreaks back to their source.

HHS is well-positioned to build on this value and enhance its efforts to promote a culture that places a premium on data sharing within and across the Department. This report focuses on creating a broader foundation for data use by leveraging existing assets and avoiding redundancies and duplication in processes, resources, and technology. At the same time, the report aims to balance transparency and repeatability of processes across HHS with the agility and technology customization necessary to meet agency requirements.

Effective implementation of this report's vision for data sharing will require significant stakeholder engagement and feedback to balance needs and ensure outcomes are sustainable across the Department. Execution of a change management plan that supports HHS agencies during implementation will further ensure sustainability.

Change Management Plan

A Change Management Plan will include the following four components:

- **Communications:** A plan for communicating a data-sharing vision and strategy across the enterprise.

[10] *CDC and the Food Safety Modernization Act*. November 3, 2015. Centers for Disease Control and Prevention, National Center for Emerging and Zoonotic Infectious Diseases. Available from: https://www.cdc.gov/foodsafety/fsma/index.html

- **Training:** A plan for training relevant members across the various HHS agencies that are entrusted with data management and data-sharing vision.
- **Programming:** A plan for implementing internal code-a-thons, workshops, and other events that enable agencies to work collaboratively on specific data-related challenges, raise awareness of available data, and promote the value of data sharing. These events also provide an opportunity to solicit input on improving the data-sharing process and addressing obstacles and challenges.
- **Strategic Growth:** A plan to develop content around the success stories resulting from data sharing, curating them in a central location, and communicating them out across the enterprise to support the continual growth of the data-sharing culture.

Communications Plan

The communications plan will explain the overarching vision and goals for the future of data sharing within HHS. Responsibility for implementing this plan will reside with a HHS wide Data-Sharing Steering Committee that includes experts who serve on the HHS Data Council, agency data workgroups, the CIO Council, the Office of the CTO and other HHS groups whose work can inform its efforts.

The overall objective of the communications plan is to encourage stakeholder buy-in to the vision laid out this report. It will do this in several ways, such as providing guidance on soliciting feedback from stakeholders to identify areas requiring additional stakeholder engagement and identifying any persistent concerns.

Training Plan

Building on the success of the HHS Data Science CoLab[11], the Data-Sharing Steering Committee will also produce a training plan that identifies skills and capabilities Department staff need to govern and contribute to data sharing. The training plan will also provide opportunities for meeting these needs through training of HHS staff.

Programming Plan

Virtual or in-person programming will be essential to fostering a culture that values data sharing within HHS. To this end, the Data-Sharing Steering Committee will develop a programming plan that outlines events that can be especially effective for this purpose and for facilitating greater collaboration and greater awareness of HHS's data assets.

The focus of these data-centered events will be specific, high-priority issues that involve multiple agencies and that because of legal, technical, or operational limitations, require the use of data

[11] https://www.hhs.gov/idealab/2017/10/04/data-science-colab-connecting-communities-across-hhs/

that is often difficult for those agencies to share in an effective manner. By focusing attention on a singular priority issue at a time, and conducting events iteratively as new issues are identified, these events can generate significant and immediate progress.

Strategic Growth Plan

The Data-Sharing Steering Committee will develop a strategic growth plan that identifies specific internal audiences for targeted engagement and specific channels and approaches for highlighting successes in data sharing. The strategic growth plan will also support efforts to quantify the return on investment (ROI) on data sharing. Sharing success stories and measuring ROI will spur the next generation of data stewards and provide positive reinforcement for agency-level staff.

DATA-SHARING PROCESS

The 2018 State of Data Sharing Report[12] showed that a cohesive, well-defined HHS-wide strategy for facilitating inter-agency data sharing is essential to a long-term data-driven vision of HHS. The data-sharing process described below is a key part of this strategy. Although it already exists in some HHS agencies, it does not exist for the Department as a whole. Furthermore, many of the agency-level processes for data sharing are largely manual, so an opportunity exists to implement new practices that will reduce the time required to fulfill data-sharing requests.

It is important to note that the process described here (Figure 1) is not intended to replace all existing data request and sharing processes. Agencies and divisions that have effective data request and data-sharing processes can continue to use them, and they will be encouraged to share those practices as part of requirements-gathering activities taking place during the implementation of this report's work plan. It is also important to note that HHS agencies, in collaboration with the Data-Sharing Steering Committee, will determine the types and sources of data shared through this process.

Note: the technical components and working groups referred to below are described in detail in subsequent sections of this report.

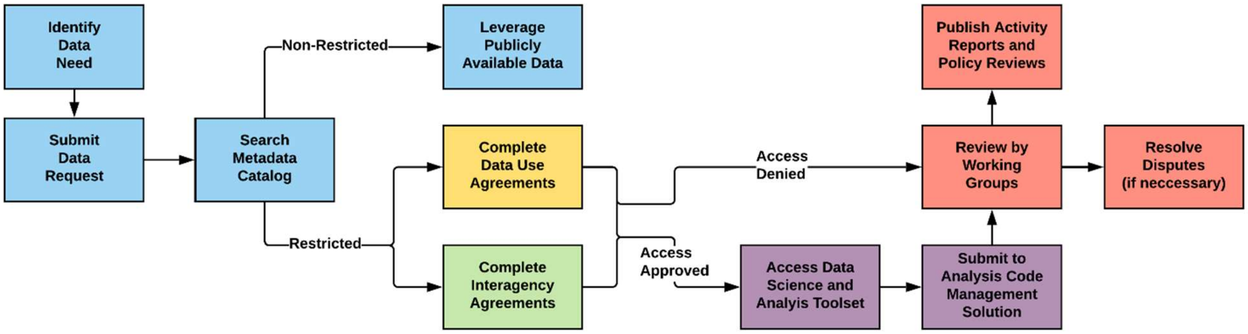
Identify Data Need
Submit Data Request
Search Metadata Catalog
Non-Restricted
Leverage Publicly Available Data
Restricted
Complete Data Use Agreements
Complete Interagency Agreements
Access Denied
Access Approved
Access Data Science and Analyis Toolset
Submit to Analysis Code Management Solution
Review by Working Groups
Publish Activity Reports and Policy Reviews
Resolve Disputes (if neccessary)

Figure 1: Data Request Process

[12] *The State of Data Sharing at the U.S. Department of Health and Human Services.* September 2018. Office of the Chief Technology Officer, U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf

Promote Data Discoverability

A robust environment of inter-agency data sharing is indicative of a strong, collaborative cross-HHS effort to improve health and social services for all Americans. A failure to generate requests for data, conversely, can indicate a lack of awareness of the data available or the perception that the burden of acquiring data is greater than the benefit to be derived.[13] It will be important, therefore, to foster an environment across the enterprise that promotes collaboration, assists in identifying needed data, and results in a strong, continuous demand for inter-agency data sharing.

Submit Data Request

As a data need is identified, it should be recorded in a manner that is detailed and structured. Once established, the Data Use Authorization Management process will capture information regarding a data request, potential restrictions on uses of the data, the time of response, and any denials of access that might result. Through this process, HHS will gain a more comprehensive view of the data of greatest need and most significant value, as well as the obstacles to accessing that data across the entire enterprise. Subsequently, more effective and strategic actions can be taken to address any problems and better allocate resources.

Search Metadata Catalog

With tens of thousands of datasets maintained across 29 agencies and offices, it is currently impossible for any single HHS staff member who needs data to know the existence, location, and characteristics of all datasets that agencies have agreed to share. A centralized, comprehensive, and searchable catalog of metadata will be developed over time to ensure all relevant and available data can be identified. By searching the Metadata Catalog, authorized users will be able to quickly retrieve summary information about all relevant datasets managed and made shareable by HHS agencies and offices.

In addition to reducing the amount of time spent searching for data, this Metadata Catalog may capture data use agreements, inter-agency agreements, statutes, and regulations restricting a dataset's use. While each agency will retain control of the data it shares, this enhancement will save data users' time spent researching statutes, regulations, and agreements.

[13] *The State of Data Sharing at the U.S. Department of Health and Human Services.* September 2018. Office of the Chief Technology Officer, U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf

Leverage Publicly Available Data

Working together with agencies that administer public data portals, a comprehensive public data catalog shall be developed that includes all data available for use without restrictions, for example, through HealthData.gov and other agency-operated sites.

Encouraging HHS staff to explore this public data catalogue first will likely enable some users to obtain needed data more quickly and easily. This will also reduce the amount of time, effort, and money required to operate and maintain the internal data-sharing platform described in detail later in this report.

Complete Data Use Agreements

Many data use agreements have been created within HHS, and some can serve as models for developing—in collaboration with the Office of General Counsel—more broadly applicable agreements that can be housed in the internal data-sharing platform. These agreements, together with widely accepted interpretations of laws, regulations, and policies that have not been incorporated into agreements, will be available to users of the platform.

Complete Inter-agency Agreements

In addition to developing more broadly applicable data use agreements, HHS will develop an approach for managing the time and money agencies spend in sharing data with other agencies. For example, in some cases, the inter-agency data sharing agreement process may be automated, an approach the data-sharing platform is ideally suited to accommodate.

Access Data Science and Analysis Toolset

After a request for data is approved, the user will receive access to the requested dataset(s) via the Data Science and Analysis Toolset interface. This component of the data platform will facilitate desired dataset configuration and make datasets available for analysis in a controlled, secured environment. To promote security and verify the authenticity of the user, a secure login protocol will be adopted.

Submit to Analysis Code Management Solution

All data configurations will be logged by the Toolset interface. Depending on the sensitivity of the data, the code for configuring data will be stored in one of two locations: a platform-wide repository or a project-specific repository. The Analysis Code Management Solution serves to capture software code used for analysis, so it can be reused for similar purposes.

Review by Working Groups

A working group described later in this report will monitor the outcomes of data requests and support the requirements of applicable data use agreements. A working group will also recommend workshops or other types of training when problems with data use are identified.

Resolve Differences

Pre-defined roles can facilitate resolution of differences, if they arise, or other issues regarding data-sharing requests. These will be logged and will include the level of escalation, to whom, by whom, and the nature of the data guidance required. The following table outlines possible roles for staff, data-sharing working groups, and the data-sharing steering committee.

Who Escalates	To Whom	Type of Issue
Project-level staff	Working Group associated with the type of issue	● Data: incongruous data definitions ● Privacy and Security: Disputes about use agreement terms ● Technical: Unusable data delivery format ● Operations: Unexpected cost, risk, etc.
Working Group	Data-Sharing Steering Committee	● Major or repeated project issues ● Agency complaints ● Requests for changes to data governance processes
Data-Sharing Steering Committee	Deputy Secretary of Health & Human Services	● Major high-impact issues that need immediate resolution and could result in termination of the project and damage to an agency or Department

Table 1: Escalation Assignments

Policy Reviews

Four working groups described later in this report will conduct an annual data-sharing activity and policy review and submit findings from this review to the Data-Sharing Steering Committee for consideration. These reviews will provide the Data-Sharing Steering Committee with information on best practices, lessons learned, and common challenges. These reviews will also recommend ways to build on successes and overcome challenges.

ENABLING TECHNOLOGIES FOR DATA SHARING

This section depicts the envisioned data-sharing environment from physical, logical, conceptual, and functional viewpoints. It identifies potential technical and analytical approaches, which when implemented, will enhance the ability of HHS staff to derive new insights from HHS data — insights that may improve the delivery of HHS programs and further the HHS mission.

In the envisioned environment, each agency will continue to host, maintain, and own the data it collects. Each agency will also continue to determine who may gain access to its data. Inter-agency data sharing is enabled by having agencies make certain elements of their data available in a secure data-sharing platform where the data are transferred, when needed, in accordance with applicable legal requirements. Data in the platform will be prepared by metadata extraction, format conversion, cross-linking to other data, aggregation, anonymization, and indexing. Data-sharing agreements, analysis tools, governance processes, representations of common fields, and analytical code libraries stored and managed in the platform will enable staff from other HHS agencies to access and use the data.

Salient benefits of the proposed data-sharing platform include:

- **Security**: Appropriate security levels will be assigned to users, as defined by the Data-Sharing Steering Committee in partnership with agencies, making the right data accessible to the right users and user groups.
- **Cost Effectiveness**: Content in the platform will be prepared “as needed,” reducing preparation costs for upfront processing. A “big data” computing fabric will make it possible to scale data processing.
- **Flexibility**: Users from different agencies within HHS (across the country, or around the world at a later time) can have rapid, secure access to data from anywhere. This increases the consumption of data and helps drive optimized business decisions.
- **Knowledge Distribution**: Content from various sources can be collected, organized, and processed using big data analytics approaches. HHS enterprise-wide information available in the platform should enable employees to obtain insights and solve challenges.

The platform environment will leverage modern technology to enable the use of AI/machine learning techniques to unlock new insights while ensuring adherence and compliance with privacy and security statutes and regulations. The environment will enable seamless data sharing and collaborative analysis of both aggregated and sensitive microdata across HHS agencies. This necessitates an environment that enables robust tracking and data governance in order to meet and exceed the rules and regulations surrounding the use and disclosure of data maintained by HHS.

Functional components for the environment are detailed in this section. In addition, a notional technical schematic is depicted. Descriptions of each component, provided below, reflect an initial vision for the major capabilities of the proposed platform, but are subject to revisions based on user experiences and feedback.

Functional Components

To address the key challenges for HHS and its agencies outlined in "The State of Data Sharing at the U.S. Department of Health and Human Services,"[14] the platform for data sharing must add value across four functional attributes: foundational, data governance, data security, and technical usability. The following sections depict the attributes of the data-sharing platform with respect to these categories.

Key Transformational Concepts

- **The platform will allow for transition from the current *point-to-point* to a *hub-and-spoke* model**: The proposed model requires a transition from the current model where inter-agency data sharing requires one-to-one or several one-to-many agreements and transactions, essentially a point-to-point system, to a hub-and-spoke system for data sharing and analysis. Data-sharing agreements, analysis tools, governance processes, representations of common fields, and analytical code libraries will be stored and managed in the platform. A high degree of importance will be placed on meeting user needs and ensuring data privacy and security.
- **The platform will allow for ready use of external code libraries.** The level of new code creation and increasing sophistication of code libraries can dramatically increase analyst efficiency and reduce errors. The platform will facilitate the ready use of external code libraries and the integration of external code. Note that governance and security concerns must be addressed to accomplish this without introducing undue risk.

[14] *The State of Data Sharing at the U.S. Department of Health and Human Services.* September 2018. Office of the Chief Technology Officer, U.S. Department of Health and Human Services. Available from: https://www.hhs.gov/sites/default/files/HHS_StateofDataSharing_0915.pdf

- **The platform will prioritize code sharing.** Collaboration or code sharing across agencies is not widely utilized despite a significant degree of similarity in work. The platform will enable users to share analysis and other code for re-use, enhancement, and modification. Traceability, promotion, approval, and other considerations should be promoted. Code sprawl should be avoided.
- **The platform will facilitate data quality correction.** Agencies use internal and external resources to clean and correct data. Over time, the goal is to build capability for the platform to facilitate these functions when possible for the purposes of consistency and efficiency. This vision includes the ability for users to flag data quality concerns and for agency data stewards to review flagged items.
- **The platform will allow for eventual integration of non-HHS foundational data for internal data sharing.** Many agencies use information and data from other sources (e.g., U.S. Census Bureau, the Social Security Administration, and the U.S. Departments of Agriculture, Education, Labor, and Housing and Urban Development). The platform should accommodate the inclusion of data from these sources and accommodate data-sharing agreements, segregation of access, and other requirements. The platform will also support possible sharing of internal HHS data with these and other federal agencies that rely on HHS data to inform their internal research, evaluation, and programmatic and policy activities.
- **The platform will allow for a common, centrally managed user authentication.** Rules that enable role-based access will need to be developed. Over time, the platform should enable users to authenticate with familiar/existing credentials. This authentication should be passed through to the different components of the system to enable robust and efficient management of users and permissions. HHS uses and has gained substantial adoption to Access Management System[15] (AMS).
- **The platform will optimize storage of external data for internal data sharing.** HHS uses or buys a significant amount of external data. Where feasible and desired, the platform should centralize storage and dissemination of this data to optimize data transformation efforts and reduce the need for duplicate purchasing of external data by different agencies.

Data Use Agreement Attributes

- **The platform will enable automated tracking of contents of data use agreements.** Many data use agreements have common requirements (e.g., background checks, limit user access, limit granularity, focus use). The platform shall be capable of collecting and maintaining this information.

[15] https://ams.hhs.gov/

- **The platform will enable standardization and mapping of these agreements.** Language and terms and conditions in data use agreements vary widely across the Department. To the extent possible, the platform shall standardize language in data use agreements. To increase visibility, traceability, and auditability of data use agreements, the platform shall make all data use agreements searchable. Where needed, any new data use agreements shall be automatically generated by the system where the request/need is deemed low risk from a privacy and security perspective.
- **The platform will store data use agreements for long-term use.** Data use agreements are presently stored in various formats and means across the Department. The desired platform will store inter-agency data use agreements and enable accounting for disclosures.
- **The platform will offer standardized processes and rules for data anonymization.** The platform shall offer standardized processes for data anonymization, including masking and de-identification, based on a risk assessment and legal requirements. The ability to streamline and standardize these techniques across HHS will save staff time and ensure anonymization is implemented appropriately where necessary.

Data Security Attributes

- **The platform will prevent unauthorized use of data and appropriately segregate access to data.** To preserve data integrity and security, the platform shall prevent unauthorized use of data. This shall include unauthorized export of data from the platform. In as automated a fashion as possible, the platform shall segregate access on a basis consistent with data use agreements.
- **The platform will be able to track and log use.** For various purposes including supporting auditing, ensuring compliance with data use agreements, and preventing unauthorized activity, the platform shall track and log user activity in a secure manner. This shall be accomplished both at the system and user levels.

Technical Usability Attributes

The platform shall enhance transparency and awareness of available data, thereby increasing opportunities to use and benefit from the Department's data resources. A vast diversity and amount of data are maintained across the Department. Data and documentation are largely segregated. The platform shall unify and increase search ability and discoverability of documentation surrounding available data content to increase users' ability to find and derive value from data and archive older datasets.

The platform shall be able to dynamically flex to meet user load and computational needs. As the platform gains adoption and users throughout the Department, the platform shall seamlessly scale to meet user loads. This scalability shall span all aspects of the platform including user account maintenance, data use agreements, data integration, etc. As the sophistication of analysis techniques applied increases, so will the computational load (the size and quantity of operations performed in the database and on machines hosting statistical and analysis tools) on the system. The platform shall seamlessly flex in near real-time to effectively and efficiently accomplish computational tasks. For example, algorithms or analysis that can be parallelized shall be expanded to “n processors” to balance computational time and system complexity.

Technical Architecture

Overview

The Data-Sharing Platform will be designed to facilitate streamlined interaction between people and process. The platform will be built around five key components:

- Data Use Authorization Management System
- Metadata Catalog
- Data Science and Analysis Toolset
- Analysis Code Management Solution
- Inter-agency Data Hub

Two primary sets of users are anticipated: analyst users and system owners. Analyst users are expected to primarily interact with the Metadata Catalog, Data Science and Analysis Toolset, Inter-agency Hub, and Analysis Code Management Solution, based on analytical needs. System owners will likely be primarily involved in administrative tasks associated with data use and interact with the Data Use Authorization Management System.

All users will likely initially interact with the platform via the Data Use Authorization Management System to view the Metadata Catalog to review available data within the Data System. Analyst users may then engage the Data Science and Analysis Toolset housed within the User and Analysis Interface to explore, analyze, and visualize data stored in the Data System. Analysis code will be housed and maintained by the Analysis Code Management Solution.

To facilitate ongoing active and ad-hoc auditing of user and system activity, interaction with the system will be tracked at a detailed level by a system and user activity logging capability. This capability will include tracking from all system logs, database logs, network logs, and other user behavior. The exact specifications of and requirements for the monitoring system will be determined in collaboration with users and HHS’ privacy and security teams.

Analysts and non-technical end users will access the system through a graphical user interface. This interface will serve two primary functions: facilitating the use of the Data Use Authorization Management System and serving as the access point to analysis tools. It may be desirable to customize the user interface or point of entry into the system for each user group.

Figure 2 depicts the envisioned details of the platform. The entire platform will reside within a cloud environment. All users and sensitive HHS systems are also presumed to reside within this environment.

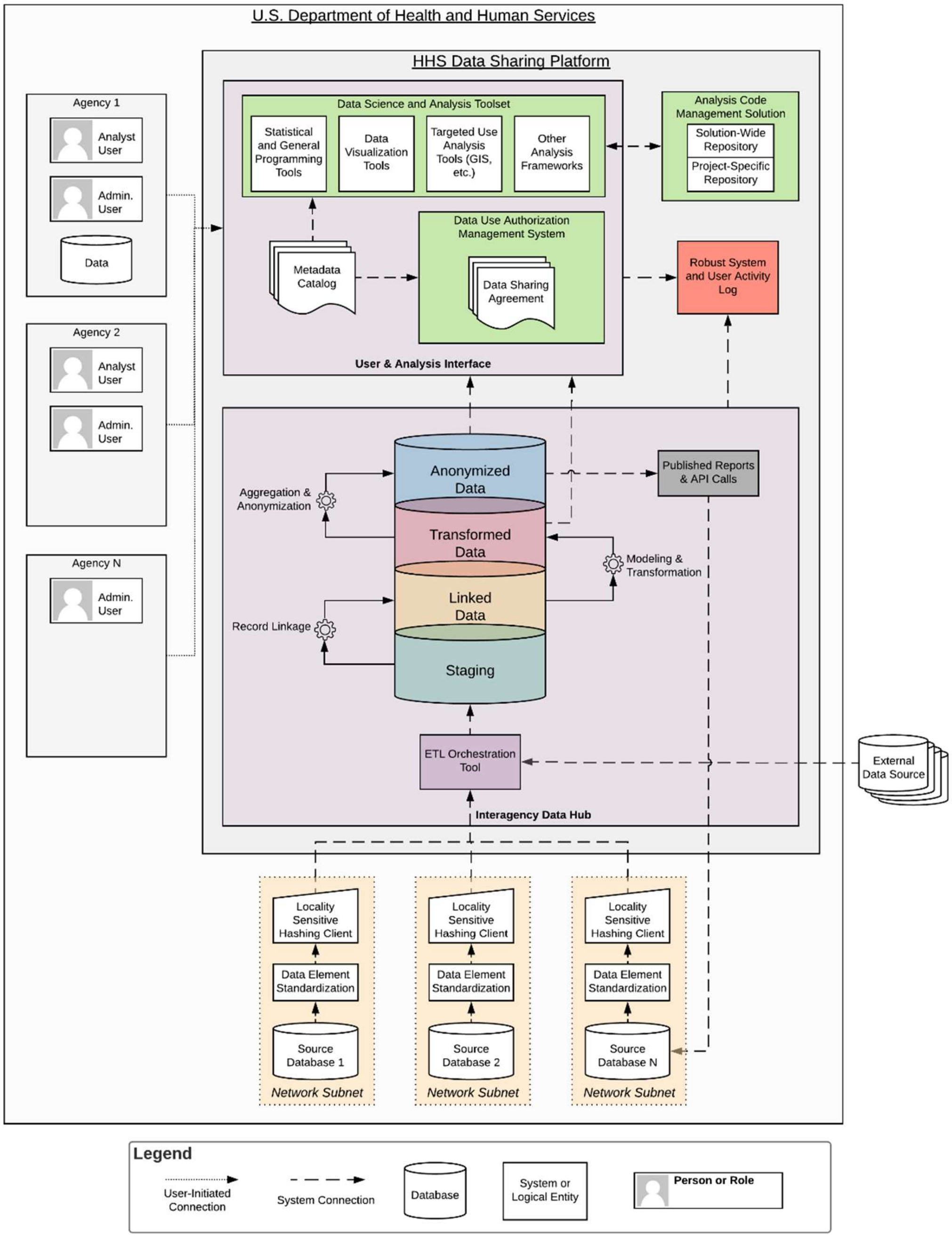
U.S. Department of Health and Human Services
HHS Data Sharing Platform
Agency 1
Analyst User
Admin. User
Data
Agency 2
Analyst User
Admin. User
Agency N
Admin. User
Data Science and Analysis Toolset
Statistical and General Programming Tools
Data Visualization Tools
Targeted Use Analysis Tools (GIS, etc.)
Other Analysis Frameworks
Analysis Code Management Solution
Solution-Wide Repository
Project-Specific Repository
Metadata Catalog
Data Use Authorization Management System
Data Sharing Agreement
Robust System and User Activity Log
User & Analysis Interface
Anonymized Data
Transformed Data
Linked Data
Staging
Aggregation & Anonymization
Record Linkage
Modeling & Transformation
Published Reports & API Calls
ETL Orchestration Tool
External Data Source
Interagency Data Hub
Locality Sensitive Hashing Client
Data Element Standardization
Source Database 1
Network Subnet
Locality Sensitive Hashing Client
Data Element Standardization
Source Database 2
Network Subnet
Locality Sensitive Hashing Client
Data Element Standardization
Source Database N
Network Subnet
Legend
User-Initiated Connection
System Connection
Database
System or Logical Entity
Person or Role

Figure 2. Technical Schematic of the HHS Data-Sharing Platform

The following sections detail each of the platform's primary components (Data Use Authorization Management System, Metadata Catalog, Data Science and Analysis Toolset, Analysis Code Management Solution, Inter-agency Data Hub) and their interaction with one another.

Data Use Authorization Management System

The Data Use Authorization Management System will create and track data use and access requests initiated based upon data available in the Metadata Catalog. The Data Use Authorization Management System will also facilitate a workflow for the disposition and approval of each data request.

Users shall be able to select individual or multiple datasets to request. For each dataset, the user shall be able to see the following:

- Descriptive metadata about the dataset (located in the Metadata Catalog)
- Important statutes, regulations and agreements that apply to the use or sharing of data, both from a data governance perspective and from a data application perspective
- Any fees associated with the use of data, if applicable

When developing a request, the user shall be required to submit the following:

- Overall purpose of the request
- Datasets requested for use
- Rationale for requesting each dataset
- Duration of use required (e.g., one-time use, year-long research project, monthly report, annual report)
- Level of access requested

Metadata Catalog

A Metadata Catalog will include information about internal datasets available for analysis by analyst users across HHS agencies. A Data Working Group (described in the "Inter-agency Data-Sharing Structure" section) will unify standards for metadata based on recommendations and input from all HHS agencies. It will draw from widely used metadata standards that align and fit with HHS's needs.

HHS currently maintains an Enterprise Data Inventory[16] (EDI), an inventory of agency data resources including public, restricted public, and non-public datasets, with descriptive information about each. This information will be a valuable resource during the development of

[16] https://healthdata.gov/dataset/hhs-enterprise-data-inventory

the Metadata Catalog, focusing on systems rather than data extracts, and provide more descriptive and granular information about the information content of the system.

The Metadata Catalog will be human- and machine-readable and accessible programmatically from both the Data Use Authorization Management System and all analyst environments.

Data Science and Analysis Toolset

After a request for data is approved, the user will receive access to the dataset(s). Analyst users will then access data through a virtual environment contained within the overall Data-Sharing Platform. This virtual environment will contain a mix of approved open source and commercial analysis software widely used in HHS. All analysis is able to occur within this environment, in addition to the native environments that analysts work with today.

The Data Science and Analysis toolset will directly access data from the Inter-agency Data Hub, have programmatic connectivity to the Metadata Catalog, and the Analysis Code Management Solution.

Key to the long-term viability of the platform in aggregate and the effectiveness of analysts is the scalability of the analysis environment. Each analysis environment shall be made scalable to meet the computational needs of the analysis. It will be necessary to strike the appropriate balance between system complexity and manual configuration. Data movement will also be a key consideration in the execution speed and memory considerations for the system.

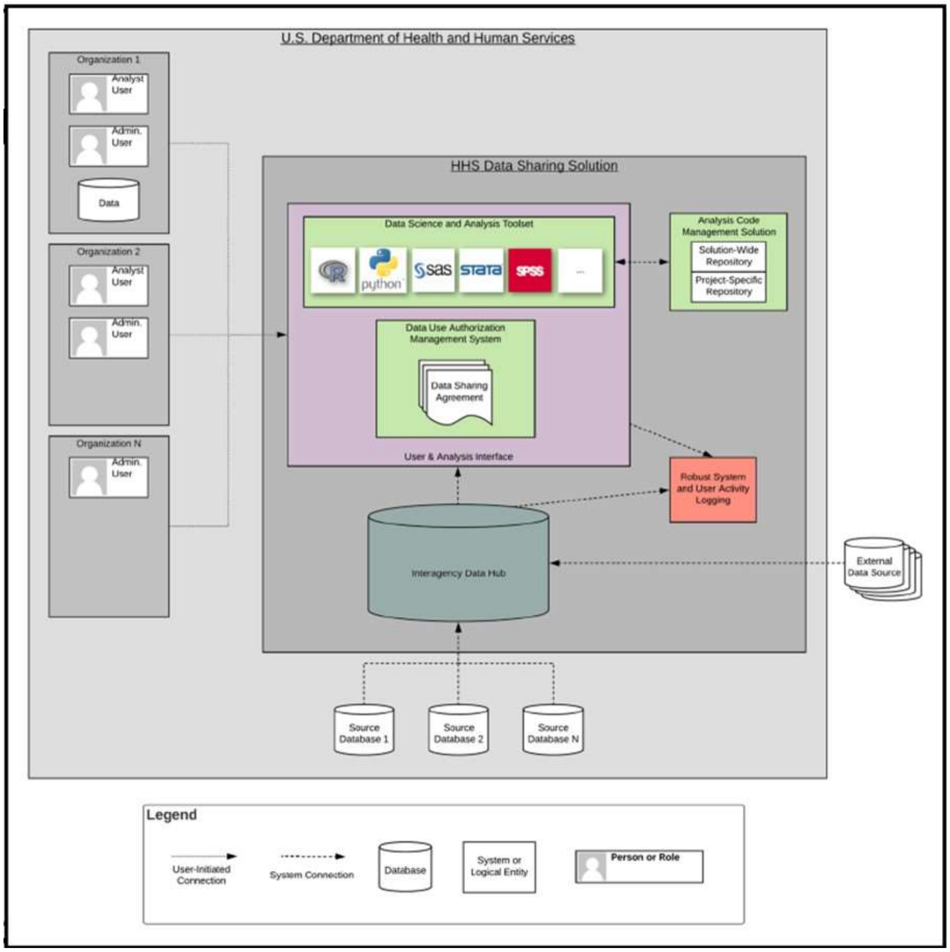

Figure 3. Potential Data Science & Analytical Toolsets on the HHS Platform

Analysis Code Management Solution

Effective analysis code management is a vital portion of the Data-Sharing Platform. A repository of analysis code is envisioned that is accessible to all users of the Platform and contains code contributed by analysts across the Department. Such a repository will encourage cross-departmental collaboration and standardization in analysis approaches (where appropriate). It is likely to require a system of governance.

A project-specific repository will contain code that does not have broader applicability. It will only be accessible to project staff.

The Analysis Code Management Solution will segregate access based on the intended use and potential utility of the code. No sensitive information or data will be stored in the Analysis Code Management Solution.

Inter-agency Data Hub

The Inter-agency Data Hub is the heart of the overall platform. The primary function is to provide HHS staff with access to data from other HHS agencies more quickly and easily. Initially, the focus will be on high value, commonly requested data sets to prove the model and then continually expand to include a wide range of applicable agency data. It is planned that the platform will enable role-based access and technical processes to allow users to access datasets from the platform's computing environment once a data request is approved.

As depicted in Figure 2, multiple data layers will facilitate transformation of the data from its raw ingested state into a form readily consumable by analysts. First, raw data will be ingested and stored[17] in the Staging layer. Next, disparate datasets will be linked on physical attributes of interest and stored in the Linked Data layer. Next, data will be transformed in a manner that facilitates analysis, and stored in the Transformed Data layer. Note that significant effort is anticipated to be needed to define and refine a data model that balances simplicity and completeness for analysis purposes. Lastly, data are anonymized, either through aggregation or other means, such as the introduction of noise, and stored in the Anonymized data layer. End users, depending on their analysis needs, will be able to access data in the Transformed Data and/or Anonymized Data layers.

The Inter-agency Data Hub's purpose is to leverage existing data resources and make data more readily understandable and analyzable by analysts from across the Department. It is not intended to replace agency systems that serve reporting functions across the Department.

[17] Note that the term "stored" in this section is intended to refer to the functional representation of data either through physical or virtual means.

Other Platform Components

Data Integration

To support data integration activities within HHS, a clear dataset and metadata ingestion process will be needed. It is critical that the metadata be uniform across all datasets to facilitate the creation of a uniform data catalog and overall management of datasets, both from a functional and security perspective. No unallowable[18] PII will be integrated into the Data-Sharing Platform.

A locally installed hashing client will mask all PII within the source system environment. This will, in turn, necessitate some standardization and cleansing of PII prior to integration into the Inter-agency Data Hub. The hashing client will employ a locality sensitive hashing (LSH) technique to preserve the ability to use fuzzy matching techniques. Note that it may be necessary or desirable to salt (i.e., apply a shared, sufficiently complex string) the input to further reduce the risk of potential re-identification. The data sourcing process is depicted in Figure 4.

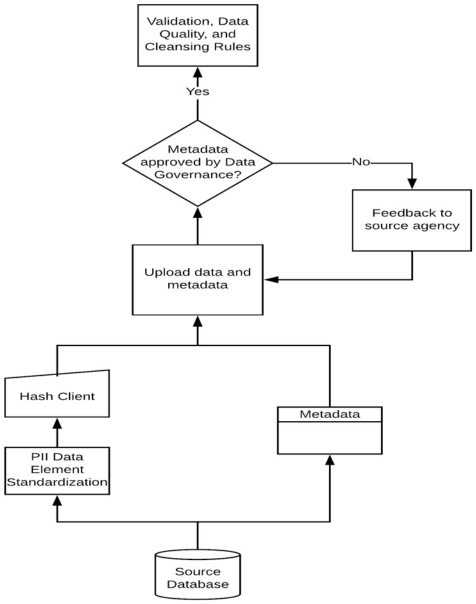
Validation, Data Quality, and Cleansing Rules
Yes
Metadata approved by Data Governance?
No
Feedback to source agency
Upload data and metadata
Hash Client
Metadata
PII Data Element Standardization
Source Database

Figure 4: Dataset and Metadata Integration Flow Diagram

[18] Data maintained by HHS agencies is governed by a variety of laws, statutes, and rules. These regulations describe how the data can be used and must be maintained. In certain circumstances, it is allowable to use PII for analysis and other purposes—that of physicians, for example.

Data Repository Design

The Inter-agency Data Hub will support data storage in different storage mechanisms and data types, including relational, NoSQL, geospatial, image, etc., with a focus on data harmonization. As many of HHS's agency systems are designed to capture and store data for in-system reporting, creating a comprehensive data model to facilitate cross-dataset analysis is vital. This will help to avoid duplication of efforts among analysts.

Hosting and Authentication Considerations

A platform that meets the functional requirements described herein will come with several hosting and authentication infrastructure issues that must be taken into consideration. Making use of open source software to the extent possible will serve to keep these costs low and make the most efficient use of government resources. Cost drivers with respect to hosting and authentication may include but are not limited to:

- The platform which shall be built in a cloud environment and which has achieved a Federal Risk and Authorization Management Program[19] (FedRAMP) certification.
- Cloud servers that can be scaled for high-intensity workloads along with the capability of shutting down servers in periods of minimal work.
- The platform which shall integrate with the HHS AMS following the Identity, Credential, and Access Management services framework.
- The architecture which shall be configured in three tiers: Presentation, Application and Data. Internal segmentation of the system architecture will help to promote strong access and authorization controls.
- Network and server logging shall be enabled and configured on key assets in order to create and store key logs for review shall an incident occur. Third-party monitoring tools may be required depending on software availability through HHS.
- Technical and non-technical staff resources for detailed design, development, and maintenance and operations of the platform.

[19] https://www.fedramp.gov/

DATA SHARING: REGULATION AND PRIVACY

A brief overview of key statutes and policies that must be accounted for as part of a regulation inventory with respect to data use agreements is noted below. Please note the following list of regulations, a substantial portion of which are summarized in “The State of Data Sharing at the U.S. Department of Health and Human Services,” are not all-inclusive (it does not, for example, include all agency-level policies in place). A more rigorous inventory shall be completed under the direction of the Privacy and Legal Working Group and in close consultation with the Office of General Counsel and aligned with other related efforts.

- Privacy Act of 1974 – The law strives to balance the government’s need to maintain these records with the individual’s right to be protected from unwarranted invasions of personal privacy. The Privacy Act limits agencies’ maintenance and disclosure of information about individuals. Agencies may maintain "only such information about an individual as is relevant and necessary to accomplish a purpose of the agency required to be accomplished by statute or Executive Order of the President" and permits disclosure only in limited circumstances.
- The Confidential Information Protection and Statistical Efficiency Act (CIPSEA) of 2002 – At HHS, there are two entities covered under CIPSEA: The National Center for Health Statistics, the federal health statistical agency, and the Center for Behavioral Health Statistics and Quality, a designated statistical unit. CIPSEA restricts the use of information exclusively to statistical purposes only.
- Title 42 of the Code of Federal Regulations (CFR) Part 2: Confidentiality of Substance Use Disorder Patient Records – The confidentiality provisions of 42 CFR Part 2 restrict the disclosure and use of substance use disorder patient records. Consequently, the ability of researchers and policymakers to more fully characterize population health issues related to substance use disorders has been limited.
- Health Insurance Portability and Accountability Act (HIPAA) privacy and security regulations – This provided for the continuation of healthcare information for workers across changes in jobs, strove to reduce healthcare fraud, required industry-wide standards for electronic healthcare transactions, and created regulations surrounding the privacy, security, and handling of protected health information (PHI).
- FOIA (5 U.S.C. § 552), The Trade Secrets Act (18 U.S.C. § 1905) and section 1106 of the Social Security Act - These also play fairly substantial roles in some agencies’ assessments of data disclosure policies and procedures. Data acquired or purchased

from entities outside of HHS may be subject to interpretation of laws beyond the scope of HHS' authority.

Due to privacy and security regulations, HHS agencies enter into a substantial volume of data use agreements with sister agencies. Given the sensitive nature of the data maintained across HHS, the breadth of information, and the volume of potential uses and users, developing a robust and automated data use agreement generation process with respect to privacy and security is paramount. This could yield standard pre-approved data use agreement elements tailored to the category of user and type of data.

The automated process to generate a data use agreement detailed in Figure 5 will begin with the requesting user entering in key information into the Data Use Authorization Management System. The requestor will be required, with the exception of the Office of the Inspector General, to submit at least the following pieces of key information:

- Overall purpose of the request
- Name, role, and division of all individuals requesting access
- Datasets requested for use
- Existing data to which the requesting agency already has access
- Rationale for requesting each dataset
- Duration of use required (e.g., one-time use, year-long research project, monthly report, annual report)

This initial data gathering process will seek to understand the type of data being requested, the purpose and intended use. Utilizing the metadata tags of the requested datasets, a lookup will be performed to the in-scope regulations to understand what governance, privacy, and legal considerations need to be included within the data use agreement. The results of this analysis will lead to a risk-based decision on whether a standard automated process can be followed to generate the data use agreement, or if—because of the risk posed by the request—a more intensive manual process must be followed to generate the data use agreement.

Using the regulations inventory as guidance, an ordering and assignment of potential data-sharing risks will be formed. These designations will be used to automate the data use agreement process such that for standard requests (deemed low risk), the requestor's data request needs and inputs will be fed into an automated script to produce the customized data use agreement. The newly generated agreement will then be distributed to both parties for execution with electronic signature.

For those requests which are deemed medium or high risk, or for those which could lead to a re-identification of individual PII/PHI datasets, the requestor will be required to meet with a member of the Privacy and Legal Working Group (described in the section “Inter-agency Data-Sharing Structure”) in order to fully document the request details and the existing data which creates the potential for re-identification.

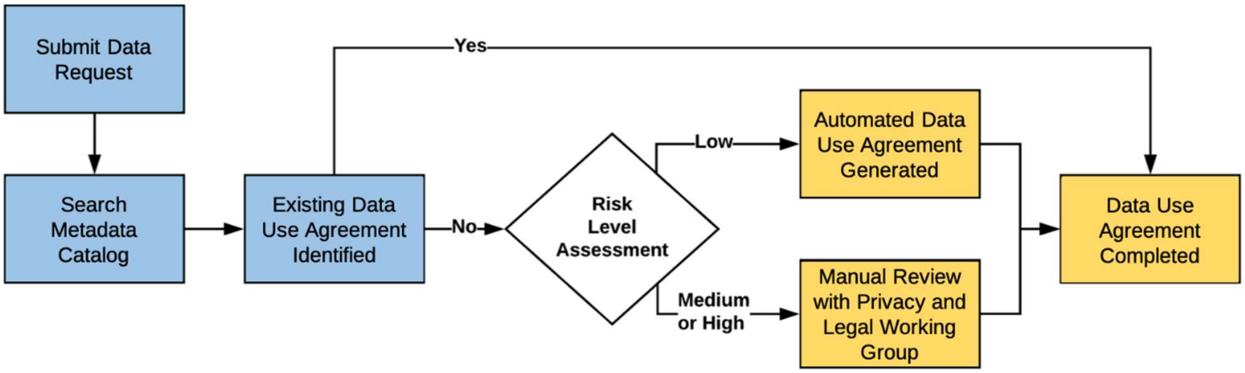
Submit Data Request
Search Metadata Catalog
Existing Data Use Agreement Identified
Yes
No
Risk Level Assessment
Low
Medium or High
Automated Data Use Agreement Generated
Manual Review with Privacy and Legal Working Group
Data Use Agreement Completed

Figure 5: Data Use Agreement Generation

INTER-AGENCY DATA-SHARING STRUCTURE

The development of a unifying data-sharing vision that recognizes the needs and challenges of each agency is best met through an HHS-wide steering committee that receives advice and support from HHS-wide working groups. Rather than duplicate roles, groups, and expertise already in place within HHS, this structure will be fashioned in ways that support and leverage these existing resources and will leverage existing committee structures where possible.

Inter-agency Steering Committee

Comprised of agency directors or their designees, a Data-Sharing Steering Committee will create the vision and strategy for sharing data across the enterprise and provide oversight and direction for the Chief Data Officer and Working Groups.

The Data-Sharing Steering Committee will institute the Working Groups, further delineate their roles, and describe how existing agency personnel will interface with the Working Group members to facilitate inter-agency data sharing. The Data-Sharing Steering Committee will report directly to the Deputy Secretary of HHS.

Prior to the Committee's launch, a charter will be written to formalize the Committee's mission, purpose, and scope; define its functions and responsibilities; and establish terms and a selection process for its members.

Inter-agency Working Groups

Inter-agency Working Groups are a proven mechanism for increasing communication and collaboration among agencies, informing and advising agency leaders, and sharing agency best practices and lessons learned.

Four Working Groups are envisioned: a Data Working Group, a Privacy and Legal Working Group, a Technology Working Group, and an Operations Working Group.

Facilitated by a member of the Data-Sharing Steering Committee, these working groups will be comprised of agency staff who are empowered to collaboratively manage the subject matter-specific elements of the new data-sharing environment.

All Working Groups will provide guidance and subject matter recommendations directly to the Data-Sharing Steering Committee. This relationship among Data-Sharing Steering Committee and the Working Groups is outlined in Figure 6.

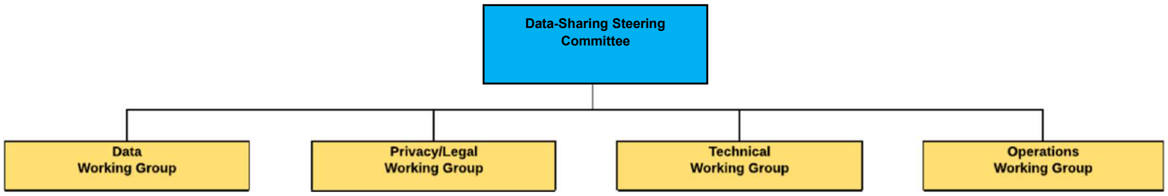

Figure 6: Relationships among the Steering Committee and Working Groups

Data Working Group

The Data Working Group will be responsible for maintaining a comprehensive Metadata Catalog, tracking inter-agency data requests, and addressing significant concerns or issues raised by data stewards and users. The Data Working Group will also:

- Identify and advance policy recommendations for consideration by the Data-Sharing Steering Committee
- Monitor data request trends
- Identify data sharing needs that are high priority
- Define and measure data-sharing metrics and tracking data sharing efficiency (e.g., time from data request and data receipt) within the Data Use Authorization Management System

Privacy and Legal Working Group

The Privacy and Legal Working Group will work closely with legal counsel to establish clear and consistent guidance for the use and interpretation of applicable legal requirements across HHS. The Privacy and Legal Working Group will make enterprise-wide recommendations to the Data-Sharing Steering Committee regarding the interpretation of regulations, policies, and laws governing data sharing and use.

Technology Working Group

The Technology Working Group will be responsible for enhancing and maintaining the functionality of the Data-Sharing Platform. Additionally, the Technology Working Group will be responsible for the sustainability and upkeep of all aspects of the technical architecture. This group will work closely with the Office of the Chief Information Officer, who will have primary responsibility for operating and maintaining the platform.

Operations Working Group

The Operations Working Group will seek to lower the financial and other resource-related barriers to data sharing. The Operations Working Group will also be responsible for a financial review of data sharing activity and, in collaboration with the Data Working Group, making recommendations to the Data-Sharing Steering Committee for increasing the efficiency of data sharing. In addition, this group will work closely with the Office of the Chief Information Officer.

NEXT STEPS

The following next steps build upon the foundation of data-sharing practices already in use across HHS while incorporating the technology, structure, and processes discussed herein in order to achieve the envisioned future state. The concepts described in the future state are based on current practices, accomplishments, lessons learned, and challenges identified by several HHS stakeholders and feedback from HHS agencies. As the next steps are actualized, it is critical that an iterative approach is adopted in which feedback is continually gathered throughout all stages of the project and changes are made as needed.

HHS and its agencies recognize a culture that fosters a more transparent and effective data-sharing practice is already beginning to take root throughout the organization. The following stages are outlined to allow the organization to build on its data-sharing culture in a way that is fully collaborative, transparent, and efficient. A high-level implementation timeline is depicted in Figure 7 below. All stages will be performed with a multi-stakeholder group in a coordinated approach to ensure the platform is both technically aligned and produced to meet the needs of HHS agencies. In addition, initial partnerships will be with smaller agencies to develop and test the platform using an agile methodology. Milestones will be developed in coordination with the multi-stakeholder group to track progress and determine next steps.

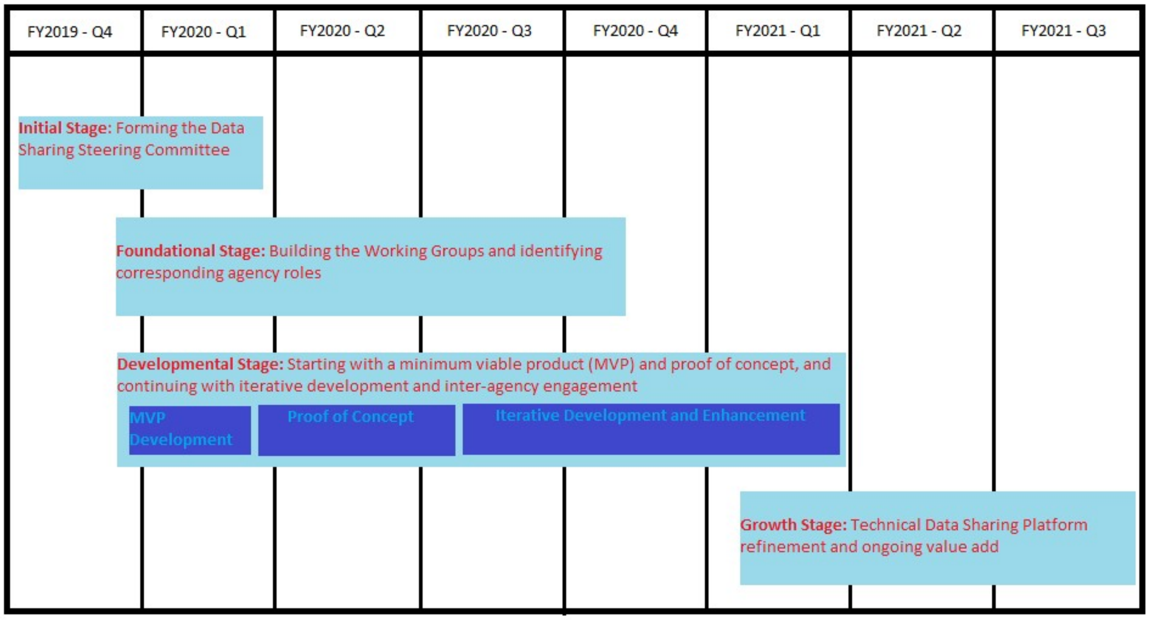
FY2019 - Q4
FY2020 - Q1
FY2020 - Q2
FY2020 - Q3
FY2020 - Q4
FY2021 - Q1
FY2021 - Q2
FY2021 - Q3
Initial Stage: Forming the Data Sharing Steering Committee
Foundational Stage: Building the Working Groups and identifying corresponding agency roles
Developmental Stage: Starting with a minimum viable product (MVP) and proof of concept, and continuing with iterative development and inter-agency engagement
MVP Development
Proof of Concept
Iterative Development and Enhancement
Growth Stage: Technical Data Sharing Platform refinement and ongoing value add

Figure 7: Data Sharing Stages and Timeline

Initial Stage: Forming the Data-Sharing Steering Committee

The initial stage will provide clear direction for the initiative and establish mechanisms for the development of priority use cases and for initial participating agencies to assist in the creation of technical and functional requirements.

This stage includes the formal establishment of the Data-Sharing Steering Committee, which will provide strategic and visionary direction for data sharing across HHS. Each participating agency will direct an existing agency-level representative to serve on the Data-Sharing Steering Committee. The agency nominees will orient themselves with the overall data-sharing culture and long-term vision of the committee.

The Data-Sharing Steering Committee will determine the priority use cases for improved data sharing between HHS agencies. The use cases will be reviewed for the technical and functional requirements needed for execution. Mapping the requirements for each selected use case will provide a roadmap for platform development that serves the current HHS priorities, with special attention given to the scalability of the platform as it pertains to future state goals.

The Change Management Plan will be crafted during this stage which will start immediately and is anticipated to be completed in 3-6 months.

Initial Stage Challenges

The main challenge faced at the early stage of creating this infrastructure entails continuing leadership support and collaboration across a multi-stakeholder group. Focus shall be paid at this stage to ensure existing resources are leveraged and a clear and consistent message is being spread about the benefits and long-term cost savings of internal data sharing, both for the individual agencies and HHS as a whole.

Foundational Stage: Building the Working Groups and Agency Roles

This stage includes the formation of the Working Groups and the identification of existing agency staff roles within each to best collaborate with the Data-Sharing Steering Committee. The Change Management Plan will serve as guidance during this stage.

With strategic direction and leadership from the Data-Sharing Steering Committee, the Working Groups will craft data management standards, begin working toward standardized interpretations of statutes and agreements, and provide technical and operational guidance related to data sharing in concert with a to-be-identified proof of concept.

Meanwhile, initial participating agencies will begin identifying staff to fill key data-sharing responsibilities to serve as points of contact with the counterpart roles identified in the Data-Sharing Steering Committee. Initially, participating agency staff will be identified to serve in Working Groups.

Following the finalization of the Data-Sharing Steering Committee and identification of participating agency staff, Internal Code-a-Thons and training sessions will be conducted to simultaneously build knowledge sharing and trust among Working Groups and between the newly formed Data-Sharing Steering Committee and each participating agency, with the primary objective of improving data/knowledge sharing culture.

This stage will start quickly, dependent on the formation of the initial Data-Sharing Steering Committee and is anticipated to be completed in 12 months.

Foundational Stage Challenges

The main challenge faced at this early stage will center on fostering coordination and collaboration. Therefore, it is important that the Data-Sharing Steering Committee, comprised of representatives from HHS agencies, is formed prior to the identification of roles within agencies to solidify buy-in and to promote diverse managerial ideas around each role within the group. This will enable the Data-Sharing Steering Committee to leverage current knowledge and expertise from the HHS Data Council, the CIO Council, and the Office of the CTO.

Other challenges at this stage include the inherent conflict that may arise from standardizing interpretation of statutes surrounding data sharing as well as those that may arise from standardizing the request/approval processes within the future state. The purpose of Internal Code-a-Thons and other inter-agency training and events is to promote breaking down communication silos to allow for standardization where allowable by law and where beneficial to the efficiency of the future state.

Developmental Stage: User Engagement and Demonstrating Value

The only way to effectively design a platform that meets the needs of end users is to actively solicit feedback from users and stakeholders and develop in an iterative fashion. The previous steps have established a technical and legal framework for securely using data in a way that directly connects agency staff to the new process. HHS can ensure a lasting impact with initial projects that demonstrate the quantifiable benefit of the Data-Sharing Platform and the Data-Sharing Steering Committee and Working Groups structure. Its purpose is not only to assist in the overall definition and validation of the platform but to act as a mechanism to build user-acceptance and buy-in through delivering value. HHS has expressed the following key goals for the development stage:

- Begin to positively transform the organizational culture and outlook toward data sharing with active collaboration with internal experts and early adopters—initial participating agencies that have already expressed an interest in supporting the effort
- Demonstrate the value of joint analysis of disparate data from across HHS
- Advance understanding about important health and human services issues in the United States through the development of new insights to guide more effective, proactive responses
- Tangibly demonstrate key components of the platform to users and stakeholders, and support initial utilization via training and technical support for users and stakeholders.
- Afford HHS the opportunity to further evaluate the functionality and capabilities of its platform

The Developmental Stage will start with the development and launch of a minimum viable product (MVP) to initial participating agencies. A proof of concept will be built in an iterative approach as use cases for initial inter-agency data sharing are defined and as more agencies are brought onboard. As new projects are initiated and as more HHS agencies begin working within this new framework, the number of users and the technical capabilities of the Data-Sharing Platform will broaden. It will be important for the Technical Working Group to facilitate and assist with testing and usability throughout this stage.

It will be necessary to define the parameters and data for the proof of concept in a manner that facilitates the accomplishment of HHS objectives and allows HHS to complete the work within the desired timeline. This will be accomplished through two parallel work streams. The first will focus on making incremental progress on the platform, with the primary objective of accomplishing the first three goals in the above list. The second provides partners with suitable data to demonstrate the platform component's capabilities, with the objective of accomplishing the last two goals in the above list.

The Developmental Stage will be completed in two parallel work streams across three main phases shown in Figure 7: MVP Development, Proof of Concept, and Iterative Development and Enhancement. Phases will be completed as the solution gains maturity—MVP Development will be complete when the platform launches to initial participating agencies, Proof of Concept will be complete when the first use case(s) are complete, and Iterative Development and Enhancement will continue until a fully-featured product is launched. The details for each work stream are detailed in the following sections.

Work stream 1 will be composed of the following:

- **Use Case Selection:** To initiate buy-in from agencies and further define the data that would be beneficial in completing the proof of concept, guided discovery sessions will be held with agencies and data groups such as the HHS Data Council, the CIO Council, and the CTO to refine a list of use cases for the proof of concept. This activity will yield a (or a set of) refined research question(s) for the proof of concept that can be answered/accomplished with the data to be used in the proof of concept.
- **Data Sourcing:** This initiative will prepare data for internal use. The result will be joined data from across agencies modeled in a manner that readily facilitates analysis. A detailed, granular dataset composed of rich information content surrounding each use case, developed in coordination with agencies, will be a key deliverable. A data dictionary would accompany the data itself.
- **Collaborative, Iterative Analysis:** The key to transforming the organizational culture around data sharing is proving first-hand the tremendous value that can come from the analysis of previously disconnected data. This activity will focus on iterative (1-week sprints) analysis performed with an integrated team from agencies.

Note that Work stream 1 would require data use agreements with agencies, a committed level of participation from resources in the agency, and a development/test environment to perform analysis. It is anticipated that analysis would be performed using tools already available at HHS (i.e., database, analytics tools, BI, and visualization tools.)

Work stream 2 will be composed of the following:

- **Data Anonymization**: This activity would be focused on anonymizing data in a manner that achieves full compliance with data-sharing statutes and other policies applicable to the data at the lowest possible level of granularity in order to maximize the possibility of attaining new, actionable insights. Data would be anonymized/aggregated at a level that would functionally prevent re-identification.
- **Public Use Data Selection**: HHS has already made a wealth of data available to the public[20]. Navigating this data can sometimes be challenging. This activity would be focused on compiling a short list of data to be used for vendor capability demonstrations.

This stage will be considered completed (after 12-16 months) only after a number of iterative projects are conducted across identified agencies, and only after all HHS agencies have representation within the Working Groups.

Developmental Stage Challenges

It is important that the developmental stage begins at the outset of this future state transition via a proof of concept and continues with iterative work throughout the lifecycle of every stage. It is important not just to show value to each agency, but to identify and show the value of data sharing early in the process and to continue collaboratively refining throughout. The members of the Working Groups within each agency as well as the HHS Data Council, the CIO Council, and the Office of the CTO will be key partners in accomplishing this. The risk of postponing or failing to iterate on a proof of concept is a lack of full buy-in.

Beyond the need to show value early and often, another challenge at this stage will be the sensitivity that must be shown to data originators/sources as data anonymization standards are formed and implemented. Care must be taken with respect to these standards necessitating that the collaborative and iterative process start early during the initial and foundational stages.

Growth Stage: Data-Sharing Platform Refinement and Ongoing Value-Add

With the successful completion of iterative projects and buy-in with agencies across HHS, the Data-Sharing Steering Committee will be able to support several multi-agency initiatives while also providing a scalable platform for direct data collaboration efforts.

[20] Examples: www.healthdata.gov/ ; www.cdc.gov/datastatistics/index.html ; wonder.cdc.gov/ ; www.acf.hhs.gov/cb/research-data-technology/statistics-research

The Data-Sharing Steering Committee will determine the priorities. They will also set the roadmap for incorporating all HHS agencies and new or modified functionalities based on the potential impact of identified use cases. This will allow the platform to expand and change over time in a way that ensures direct value to HHS.

Momentum for Growth

As the infrastructure is being built, ensuring that momentum continues from a People, Process, and Technology standpoint will be an ongoing challenge. The advocacy for data sharing and the surrounding culture must be sustained not just in the systems and processes implemented, but at the long-term executive leadership-level within HHS and its agencies.

CONCLUSION

The United States health care system is the most expensive in the world with an annual spend of $3 trillion or $9,523 per person, accounting for 17.5 percent of the gross domestic product. Developing solutions that are most effective in promoting health and human services requires a data-driven approach and a fundamental transformation in how data is leveraged and connected across HHS, its agencies, and eventually, external stakeholders.

This document describes the culture, technology, processes, and structure necessary to fulfill enterprise-wide data-sharing opportunities and leverage existing resources to more effectively share and analyze data among HHS agencies. This initiative will create the value necessary to address present-day challenges by encouraging a collaborative data-sharing culture, embracing best practices and standardizing processes, removing inefficiencies across the Department, and setting the course for a future state at HHS where inter-agency data sharing is the norm.

This document presents a vision for transitioning from ad-hoc, decentralized, and point-to-point data-sharing practices to a well-defined, federated, hub-and-spoke data management model, inclusive of all HHS agencies. As owners and stewards of data, agency leaders realize the transformational potential of data. This power, however, can only be harnessed when data is shared across agencies in a secure, responsible, and repeatable manner. Achieving this is essential in promoting a government that is data-driven, evidence-based, efficient, and effective in the delivery of health and human services, enabling improved services and health outcomes for the nation.

APPENDIX

Key Definitions

Technology

Data-Sharing Platform: An umbrella term encompassing the five technological components of inter-agency data sharing:

Data Use Authorization Management System: A centralized customer relationship management (CRM) system that captures and houses documents and details surrounding all inter-agency data requests.

Metadata Catalog: An inventory of descriptive information necessary to facilitate an understanding of and the use of data housed in the Inter-agency Data Hub. This term is intended to include other documentation surrounding the data, including data dictionaries.

Data Science and Analysis Toolset: The mechanism by which end users analyze, view, and otherwise use data available in the Inter-agency Data Hub.

Inter-agency Data Hub: A centralized data repository to facilitate the use and analysis of data from sources across HHS by HHS users.

Analysis Code Management Solution: A centralized log of data transformations occurring because of inter-agency data requests. The repository serves to capture code so that it can be re-used by the same user or other authenticated users.

Big Data: Data that contain greater variety, arriving in increasing volumes, at ever-increasing velocity. These datasets are too large to be managed using traditional data-processing software and can enable the exploration of new business and research questions.[21]

Big Data Analytics: The application of advanced analytic techniques to big datasets of structured and unstructured data to develop insights that can drive better and faster decision-making.[22]

[21] Adapted from: https://www.oracle.com/big-data/guide/what-is-big-data.html

[22] Adapted from: https://www.ibm.com/analytics/hadoop/big-data-analytics

Hashing: A data security technique that involves using a mathematical function to convert text into numbers, which must be decrypted to view the original text.[23]

Fuzzy matching: A method to improve word-based matching of sentences or phrases by using matches that are not exact, but fall above a defined matching percentage threshold.[24]

Locality sensitive hashing (LSH): A technique for data clustering or data searching that groups similar items using a probabilistic algorithm.[25]

Culture and Structure

Data-Sharing Steering Committee: A group managed by the Chief Data Officer, within the Office of the Secretary, for facilitating data sharing through four Working Groups: Data, Privacy and Legal, Technology, and Operations.

Internal Code-a-Thon: Inter-agency engagement events aimed at tackling cross-departmental initiatives and promote collaboration and data sharing—serving to promote and support a data-sharing culture.

Initial Participating Agencies: Initial HHS agency users of the Inter-agency Data-Sharing Platform. These agencies include any HHS groups using the platform to share data and conduct analysis during an identified proof of concept.

Working Group: Four inter-agency groups (Data, Privacy and Legal, Technology, and Operations) comprised of agency-level designees and facilitated by the Data-Sharing Steering Committee that provide reporting and policy recommendations to improve enterprise-wide data-sharing efforts.

Process, Regulation, and Privacy

Inter-agency Agreement Governed Data: Data requiring financial or in-kind compensation.

Priority Use Designation: Data identified by the Data-Sharing Steering Committee as being critical for the mission of the enterprise and warranting of additional resources for streamlined sharing.

Use Agreement Governed Data: Data restricted for a specific transactional use.

[23] Adapted from: https://www.techopedia.com/definition/14316/hashing
[24] Adapted from: https://www.techopedia.com/definition/24183/fuzzy-matching
[25] Adapted from: https://en.wikipedia.org/wiki/Locality-sensitive_hashing