Data Stewardship in the Administrative Data Research Facility

ADRF Project Overview

Administrative Data Research Facility (ADRF) provides a secure, cloud-based computing environment to analyze confidential micro-data for evidence-based policymaking, built atop Amazon GovCloud. All data access within the ADRF framework is based on the use of a *data stewardship* module which provides agency officials with an effective, streamlined workflow for authorizing access to sensitive data, monitoring appropriate uses of that data, then reporting on data usage in projects, while collecting the metadata generated by users.

Agencies need to be able to share data across state and jurisdictional lines in order to respond to many social problems. For example, examining the impact of access to jobs and neighborhood characteristics on the earnings and employment outcomes of ex-offenders and social benefit recipients on their subsequent recidivism or retention on welfare requires data from at least four different agencies (Corrections, Human Services, Labor and Housing) – ideally from multiple states. The same holds true for describing the earnings and employment outcomes of different education pathways, since students may get jobs in multiple states. To make that research possible, data from multiple organizations must be linked – with authorizations for data access based on requests at an individual, per-person level.

The design of the ADRF makes the necessary linkage of sensitive data possible, within a secure environment that maintains FedRAMP "moderate" certification. ADRF also streamlines the data sharing process, clarifying the relevant steps both for those who request data access and for those who determine whether to grant or reject that access. These functions are essential in that they provide necessary controls while also enabling straightforward answers to critical questions such as "Which projects use my data?" or "How is my data being used and which by products were generated by whom?"

To date, the ADRF platform has provided secure access for approved projects to over 50 confidential government datasets from over 20 different agencies at all levels of government.

Data Stewardship Requirements

A variety of data governance structures and legal requirements for use of sensitive data have grown in place at all levels of government. The general requirements for data stewardship are specified in Title III of the Evidence-Based Policymaking Act of 2018, "PART D – ACCESS TO DATA FOR EVIDENCE, § 3583. Application to access data assets for developing evidence, (a)

Standard Application Process" which requires each statistical agency to follow a process that includes:

> *(6) Standards for transparency, including requirements to make the following information publicly available: (A) Each application received; (B) The status of each application; (C) The determination made for each application; and (D) Any other information, as appropriate, to ensure full transparency of the process established under this subsection.*

The NIST Special Publication 800-53 Revision 4 standards document – "Security and Privacy Controls for Federal Information Systems and Organizations" – defines specific requirements for the role of an *information steward* as:

> *An agency official with statutory or operational authority for specified information and responsibility for establishing the controls for its generation, collection, processing, dissemination, and disposal.*
>
> *Individuals with information security implementation and operational responsibilities (e.g., mission/business owners, information system owners, common control providers, information owners/stewards, system administrators, information system security officers);*
>
> *Organizations consider mandating specific architectural solutions when required to enforce specific security policies. Enforcement includes, for example: (i) prohibiting information transfers between interconnected systems (i.e., allowing access only); (ii) employing hardware mechanisms to enforce one-way information flows; and (iii) implementing trustworthy regrading mechanisms to reassign security attributes and security labels.*
>
> *These individuals have the collective knowledge to understand organizational priorities, the importance of organizational operations and assets, and the importance of the information systems that support those operations/assets. The senior leaders are also in the best position to select the common controls for each security control baseline and assign specific responsibilities for developing, implementing, assessing, authorizing, and monitoring those controls.*

Those requirements are also defined in the "Committee on National Security Systems Glossary" (CNSS No. 4009) which establishes common terminology among federal agencies.

Our review of the literature, as well as our work with state and federal agencies and integrated data systems managers, has shown that many agencies that collect data do not have streamlined workflows either for data management or for approving data usage for research purposes. For

example, the process of creating data use agreements is often handled through email by sending draft documents between parties, sometimes requiring multiple iterations to capture all of the required information, signed documents, and so on. In addition, a paper trail makes it difficult to monitor or report data usage, and describe how the data use address agency needs. A major reason is that existing software technologies are not designed to address the core data stewardship functionalities: meeting the information security requirements and operational responsibilities of data stewards, streamlining the data request and approval process, and monitoring and reporting about the usage of sensitive data. The ADRF framework has been built specifically to address these needs..

The initial step in implementing a data governance framework involves defining the owners or custodians of data assets within an agency, in a process called *data stewardship*. Processes and workflows must be defined to formalize how the data will be stored, archived, backed up, and protected from mishaps, theft or attacks. A set of standards and procedures must be developed that defines how data is to be used by authorized personnel. Controls and audit procedures must be put into place to ensure ongoing compliance with internal data policies and external government regulations, to guarantee that data get used in a consistent manner across multiple applications.

Definitions

While notions of a "data steward" role have been popularized by vendors in the IT industry[1], those are generally tied to the priorities of corporate settings and commercial product offerings. In practice, those definitions do not align well with the privacy, security, and operational requirements of handling sensitive data in government agencies.

Instead, the following definitions clarify a system of controls within the ADRF framework over the entire dataset lifecycle:

- *data steward*: an ADRF user, nominated by the *data provider*, who is responsible for reviewing project requests for the *dataset* and exports from projects using it; who also provides descriptions of columns in the tables within the *dataset*.
- *data steward organization*: the organization of the ADRF user who is the designated *data steward* for a given *dataset*.
- *data provider*: an agency or other organization which owns a dataset available in the ADRF.

[1] For example, see "Designing a data transformation that delivers value right from the start" by McKinsey & Company.

- *analyst*: staff from within an agency who make use of a *dataset* – starting from a topic of interest, trying to pull together related publications, data, tools, and examples of analysis to plan and specify the work they need to perform.
- *researcher*: similar to *analysts* although from outside an agency, typically from academia.
- *dataset*: the entity which is the central point of information about data, including information about data use agreements, annotations, data transfer requests, responsible parties, etc.
- *dataset lifecycle*: the major processes in the lifecycle of a *dataset,* including ingestion, access request/approval, data preparation, archival, and destruction.

These definitions extend beyond typical IT notions, offering instead a rich context of linked data. That approach can be used across agencies, based on tiered access strategy that balances data quality with compliance for privacy and security requirements. The rich context in turn allows for use of the metadata for constructing knowledge graphs. Further along in the ADRF project roadmap, AI applications based on those knowledge graphs will assist analysts and researchers to identify appropriate datasets to use, and recommend how to structure their analysis.

Usage

The Data Stewardship module in the ADRF framework provides the necessary controls within a streamlined workflow that benefits both the data stewards as well as the analysts and researchers who use the data. The following goals have been established for its usage:

- Ensure that dataset policies get followed, including all legal and operational requirements.
- Streamline the data access approval workflow.
- Refine the data catalog and other data documentation.
- Reduce administrative time while improving resource utilization.
- Generate automatic reports about dataset access and usage.
- Provide workflows for approval of user-generated metadata.
- Provide workflows for approval for sharing and export of notebooks that summarize analysis.

The Data Stewardship module is implemented as a web-portal which can be accessed by approved users. A user submits a project proposal using the Project Request workflow; the proposal includes

the datasets to be used, the project members, and other information such as start and end dates. Before a request gets submitted to data stewards, members of the project must sign and upload any required *non-disclosure agreements* (NDAs) for their requested datasets. The request is then routed to the designated data stewards for evaluation. If approved, ADRF staff ensure that each user has completed the required security trainings, then the ADRF staff activate the project. Once a project is active, the Data Stewardship module includes an additional workflow for Monitoring and Reporting. These monitoring tools give data providers visibility into how their data is being used. Currently the ADRF platform logs all data access so that data owners can request to see how many people, on which projects, have accessed their data over a given period of time.

Figure 1 shows how the Data Stewardship module fits with other modules and workflows in the ADRF framework:

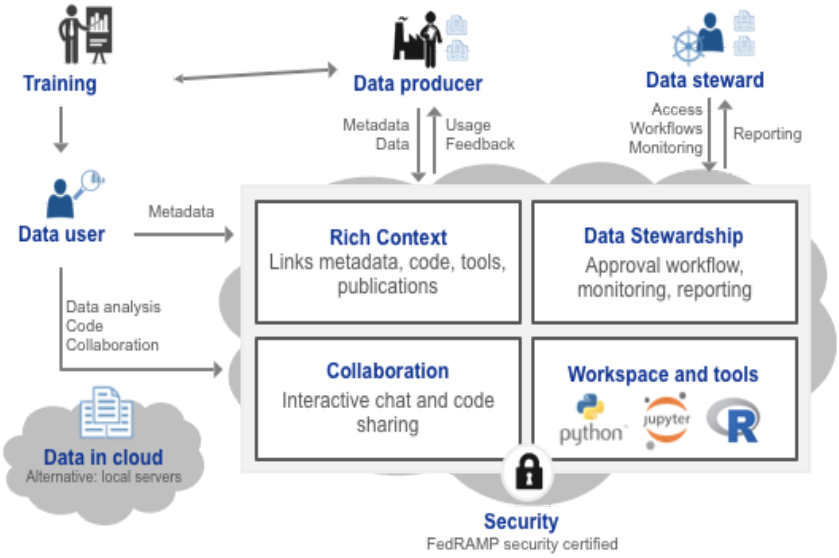


Figure 1: The ADRF Framework

The following figures provide sample screenshots of different parts of the workflow. The first set shows the workflower for users. Figure 2 shows how the Data Explorer functionality helps users find the datasets available for use; Figure 3 also provides a description of the datasets so that users can determine fitness for use. Figure 4 shows how datasets can be added to projects so that their use can be approved. The next three figures show the functionality for data stewards. Figures 5 and 6 demonstrate how the data stewardship dashboard can be used to track use and provide detail on the most used datasets, as well as the most frequent users and their institutions.

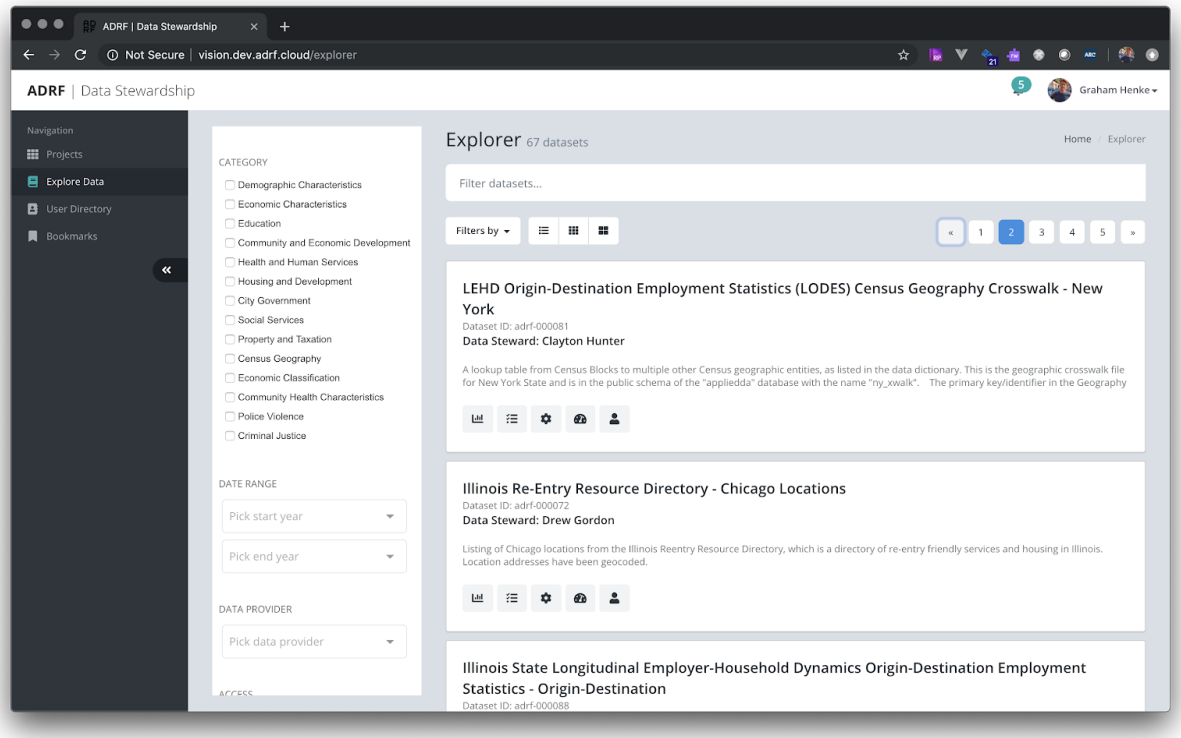


Figure 2: Data Explorer helps users find the datasets available for use

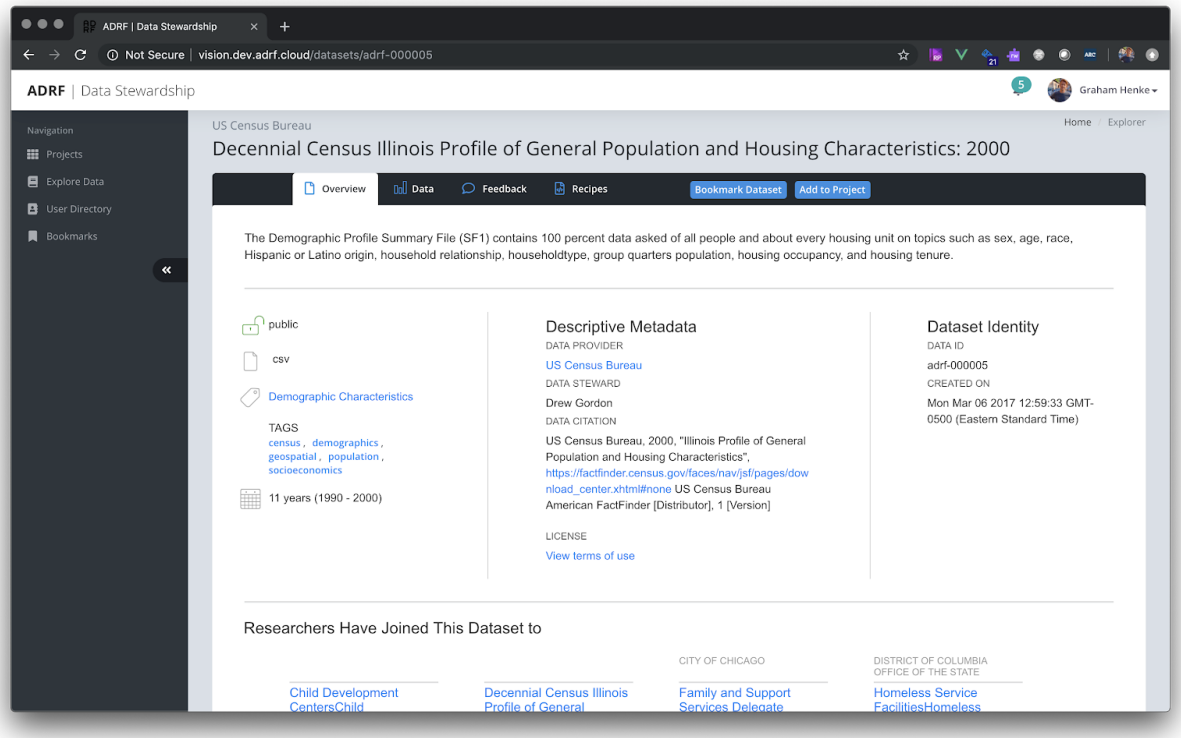


Figure 3:Data Explorer also provides a description of the datasets so that users can determine fitness for use

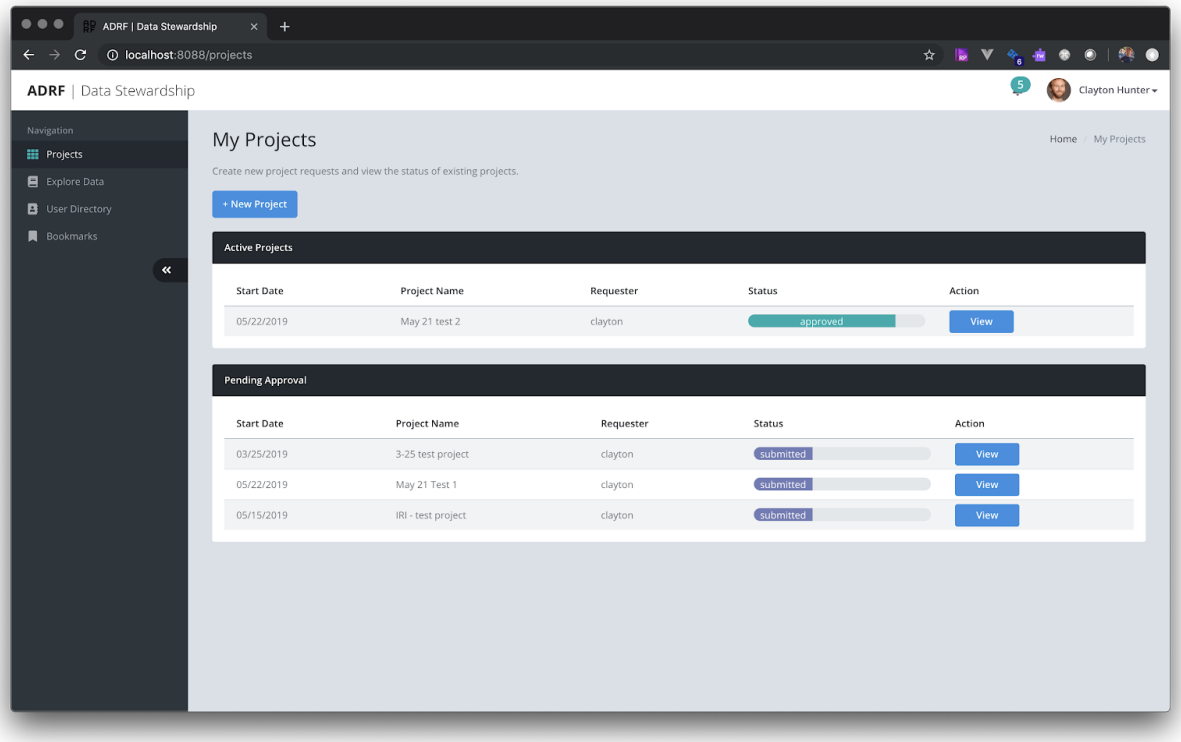


Figure 4: The module makes it easy to add datasets to projects so that their use can be approved

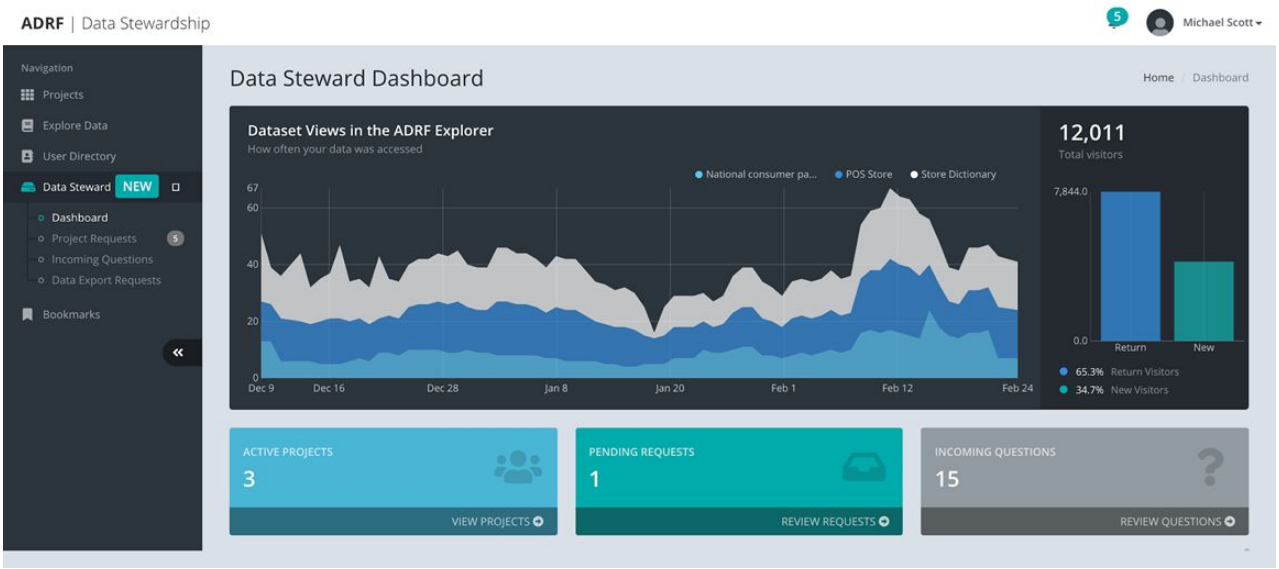
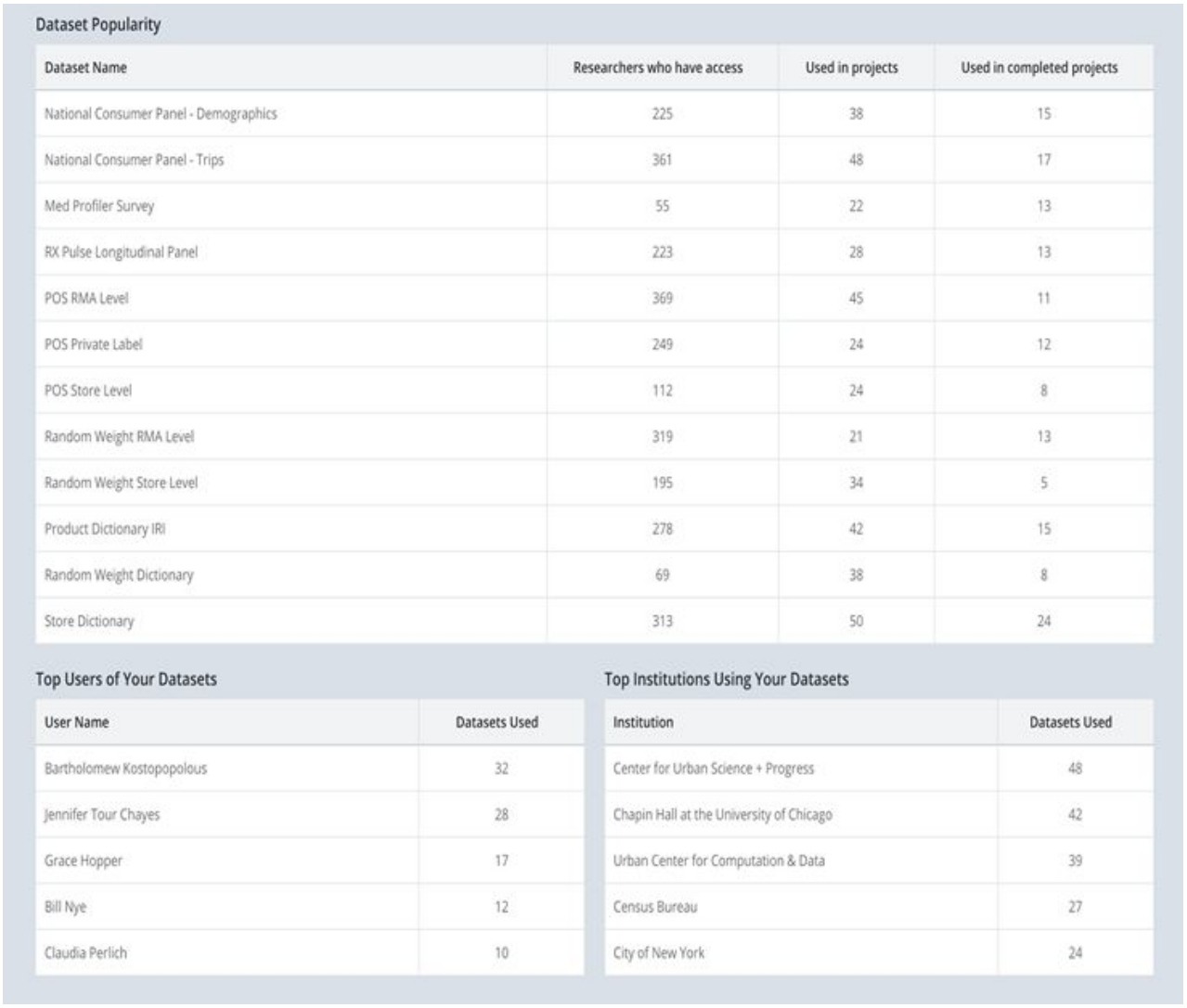


Figure 5: Dashboard tracking dataset access and use



Dataset Popularity

Dataset Name	Researchers who have access	Used in projects	Used in completed projects
National Consumer Panel - Demographics	225	38	15
National Consumer Panel - Trips	361	48	17
Med Profiler Survey	55	22	13
RX Pulse Longitudinal Panel	223	28	13
POS RMA Level	369	45	11
POS Private Label	249	24	12
POS Store Level	112	24	8
Random Weight RMA Level	319	21	13
Random Weight Store Level	195	34	5
Product Dictionary IRI	278	42	15
Random Weight Dictionary	69	38	8
Store Dictionary	313	50	24

Top Users of Your Datasets

User Name	Datasets Used
Bartholomew Kostopopolous	32
Jennifer Tour Chayes	28
Grace Hopper	17
Bill Nye	12
Claudia Perlich	10

Top Institutions Using Your Datasets

Institution	Datasets Used
Center for Urban Science + Progress	48
Chapin Hall at the University of Chicago	42
Urban Center for Computation & Data	39
Census Bureau	27
City of New York	24

Figure 6: Dashboard providing details on access and use

Roadmap and Timeline

Planned releases for the Data Stewardship module throughout 2019-2020 are currently scheduled for:

		2019				2020	
Feature	Details	Q1	Q2	Q3	Q4	Q1	Q2
Data Stewardship (ADRF Portal)							

Approval workflow	Digital data use agreements workflow and governance	POC	a				
Telemetry	Access to data usage statistics from the platform			POC	a		
Agreement Center	Upload/Review MOUs and Data Use Agreements (DUA) --pdf only--			POC	a		
Templates	Templates/Forms for MOUs, DUAs			POC	a		
Electronic Signatures	Electronic Signatures for MOUs and DUAs			POC	a		
Reporting	View data usage and requests		POC	a			
Data Export Review	User initiated via Portal; Streamlined/secure export workflow			POC	a		

By Q3 2019, the ADRF is expected to have FedRAMP certification for its version 2.0 release and deployments in production environments. The initial Data Stewardship module features will be included in that release. Subsequent releases will provide a dashboard for metrics (rather than by request), followed by the Data Agreement Center, a portal for data stewards or their appointees to propose and negotiate new data use agreements.

All projects the ADRF framework have Jupyter notebooks integrated into the analysis workflows. The notebooks can only access files already located on the file system within project workspaces, so that project permissions are enforced. Notebooks allow project members to create well-commented and shareable code and analyses – given appropriate review by data stewards.

Looking ahead, the ADRF team has been collaborating with Project Jupyter on the development of "Rich Context" features support in Jupyter, both in open source implementation plus open standards. These include support for projects, real-time collaboration, data registry, metadata handling, commenting and annotation, and usage tracking. For more details, see the Jupyter documentation for three modules in development:

- data explorer
- metadata explorer
- commenting

These features are expected in Jupyter by Q4 2019, with subsequent integration and use by the later versions of the ADRF framework.

Similarly, Git and Gitlab provide means of sharing project work with other users, following appropriate export review process by data stewards who are familiar with the terms of use for a project's datasets.

On the one hand, these open source frameworks represent contemporary "cloud-native" best practices for leveraging computing resources on GovCloud. On the other hand, they are widely used throughout industry. Source code for these frameworks is therefore subject to extensive testing, code reviews, audits, and other scrutiny to identify and resolve flaws or potential security issues.