COLERIDGE INITIATIVE

STEWARDSHIP MODULE

Contents

1. Motivation 3
2. Context 3
3. Current Approach 3
3.1 Integration with DF Admin 4
3.2 Monitoring and Reporting on Projects 4
3.3 Managing User Access 5
3.4 Refine Data Documentation 6
4. The Future 6
4.1 Overview of Customizations 7
4.2 Scenario 1 8
4.3 Scenario 2 8

1. Motivation

It is important to establish a set of data policies and procedures for managing data access, approvals and reporting for data providers. The ADRF has a Data Stewardship module designed to address such needs. This document provides the following:

(i) a summary of the existing ADRF approach,
(ii) features in development, and
(iii) a discussion of how those tools might be customized and built upon.

2. Context

The goal of the stewardship module is to facilitate the data access approval process, reduce administrative time and improve resource utilization. It provides automated and rich workflows to support the data access request and approval process, user-generated metadata moderation, data access control, and reporting. The main features are (i) management of dataset policies and data stewards, (ii) management of data access requests and approval workflows, (iii) management of user-generated metadata approval workflows, and (iv) reporting on dataset usage, listing projects, and user-generated metadata.

3. Current Approach

The ADRF Stewardship module is a web application under development for the ADRF that provides a user-friendly interface for data owners, providers, and stewards to effectively manage all stages of accessing and using restricted datasets in the evidence-based policy research and analysis lifecycle. It allows them to monitor and track usage of what data by which users for which projects, giving them better insight into how others are using their data and for reporting back to their agencies on that usage.

The module also allows them to field user questions about datasets, to accept or deny user requests for access to their data, and to administrate the export of appropriate research products after analysis is complete, giving data providers greater and more direct control over who has access to their data. Finally, the module makes it easy to moderate code recipes, user comments or annotations, and derived datasets connected to their datasets in order to ensure the quality and accuracy of information being tied to their datasets by dataset users. These functions are essential in that they provide easy to manage controls for access to datasets while also enabling straightforward answers to critical questions such as “which projects use my data?” or “how is my data being used and which byproducts were generated by whom?”

Currently ADRF administrative and technical staff complete access and export requests through separate tools and interfaces including Data Facility (DF) Admin (which is the dashboard built inside the ADRF to manage datasets, users, and projects), ADRF Explorer (which is the search and discovery tool built into ADRF), and Python and Shell scripts. The aim of the Data Stewardship module is to provide these capabilities to verified data providers or data owners

such as government employees or institutional data stewards through a single and easy to use web interface. The interface will tie directly into other tools in the ADRF such as DF Admin and ADRF Explorer via a single, comprehensive API that enforces the required controls used system-wide so that it may reliably and securely allow access to up to date information about data usage and access in the ADRF.

3.1 Integration with DF Admin

The Stewardship module control panel will interact directly with DF Admin. This approach is crucial as DF Admin is intended to be the single source for all actions essential to a fully functioning and secure data facility. DF Admin ties together restricted data, research projects, and people using the facility.

Through DF Admin, the Data Stewardship module will provide stewards with access to and control over important data and data agreement details. This includes the terms laid out in the agreement, who may access the data, and when access expires. It identifies the relevant data providers and stewards associated with a dataset, and for how long a dataset may be kept in the data facility. Stewards will also be able to review important details about prospective projects requesting dataset access, such as a description of the project's purpose, when it will start and when it will end, methodology and expected outcomes, whether approval is required for the project from an Institutional Review Board, and which users are members of the project. Finally, stewards will have access to details about users so that they may best evaluate whether dataset access is permissible or not. The level of detail on users includes their job title and organizational affiliation, whether they have reviewed and signed all relevant data usage agreements, and whether they have completed the training necessary for access to certain datasets.

3.2 Monitoring and Reporting on Projects

One key feature of the Data Stewardship module is allowing data providers, owners, and stewards to monitor usage of their datasets in ADRF projects at a glance and with the information that is important to them (see **Figure 1**). With access to all of the relevant details provided in DF Admin as described above, it will allow data stewards to see how many projects are using their datasets, for what purposes (e.g., by what code they are developing around the dataset, and what datasets they are deriving from, linking to their datasets, and annotating). The interface will provide easy-to-understand visualizations that will allow stewards to see the larger picture of how many projects are using their data, for what purposes, and when, thus facilitating data access reports back to the agency which provided the data.

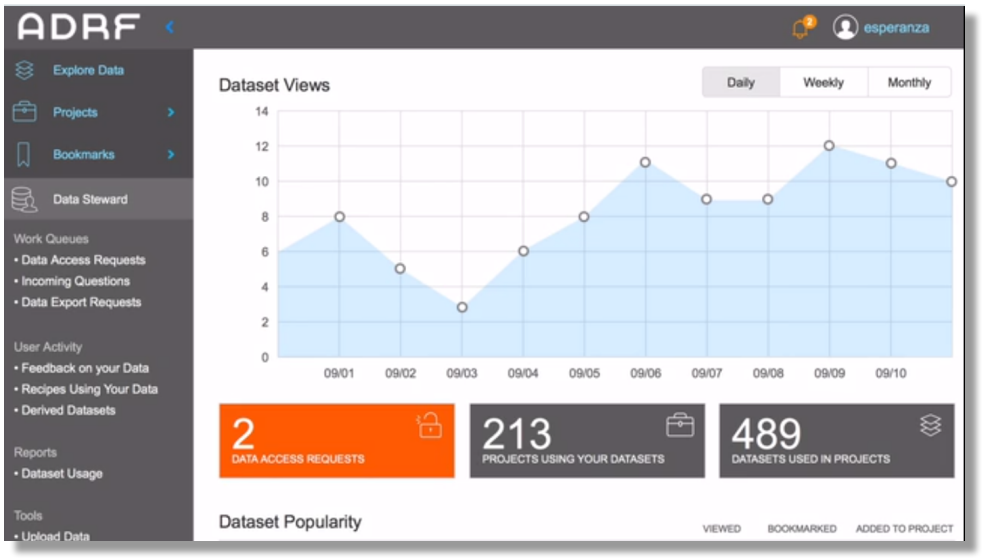

Figure 1: Sample wireframe

3.3 Managing User Access

The Data Stewardship module facilitates the data access approval process – reducing time and improving resource utilization – by providing automated, standardized, and transparent workflows to support the data access request and approval process, user-generated metadata moderation, and data access control. The main features are (i) management of dataset policies and data stewards, (ii) management of data access requests and approval workflows, and (iii) management of user generated metadata.

It provides a place where data stewards can review, respond to, and set access from requests for their data given the appropriate context they need to make decisions including details about who is requesting access to the data and for what purposes (see **Figure 2**). That is, data stewards will be able to see details about the requestor such as name, job title, organization with which requestor is affiliated as well as if those users have completed special training for handling certain types of data or if they have signed a specific data usage agreement, terms of usage agreement, memorandum of understanding, or non-disclosure agreement and if they have received approval from their institutions review board, if required. With these details, data stewards can be certain that their data are only accessed under the appropriate conditions and that users will abide by the terms of their usage.

Following a user's access to data for a particular purpose, data stewards may also view the elements of analysis which the user would like to export for purposes of publication or presenting results to external parties. In all of these interactions, data stewards can either approve or reject user requests and, where helpful, provide comments on why they may not access or export data and ways in which they might modify their request to more successfully gain access to the materials they need for their research.

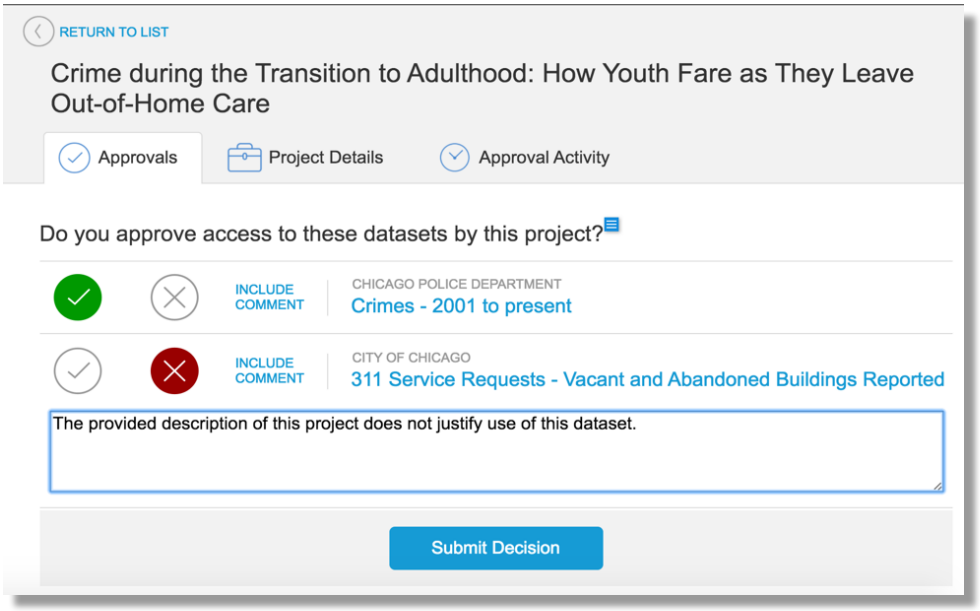

Figure 2. Wireframe demonstrating the approval and denial process for dataset requested by a specific project. Details would include additional information about the project, historical approval activity.

3.4 Refine Data Documentation

Finally, the data steward module will allow data stewards to update, improve, and manage annotations to the metadata documentation for the datasets over which they have control. In the case where the data steward is an employee of the organization that provides a dataset, they might have special insight into the meaning of a variable or value in an encoded column, shedding much needed light on elements of a dataset the might otherwise be unintelligible or obsolete and therefore might go unused by others. Allowing them to manage annotations set by other users ensures that only accurate information about the dataset is added for others to see.

4. The Future

Many of the components required to employ good data stewardship inside of the ADRF already exist for technical administrators through Python or Shell scripts and through DF Admin. To support the work of non-system administrator data stewards, these components would need to be wrapped in a single web application, driven by a well-designed user experience that bundles up this functionality for these specific types of users.

At minimum, we recommend the installation of DF Admin for managing metadata and system access, an LDAP[1] solution for operationalizing system access, and the Data Stewardship module. While the data stewardship module has been designed to integrate with these other tools, it must still be built as an interface for which DF Admin is the backend. Development of the Data Stewardship module may also require additional customizations, as detailed below.

As an important note, it must be considered whether participants intend to incorporate the ADRF workspace and data storage components for storage and access to data, or if they prefer to keep the data stored with their current institutional infrastructure. If the latter, further work would need to be conducted to determine how the data stewardship module would integrate with the storage architecture outside of the ADRF.

4.1 Overview of Customizations

Overall, we see the following customizations necessary for developing the Data Stewardship module for the specific needs of participating institutions:

1. Develop interfaces for data stewards to work on top of DF Admin (i.e., separate from technical operators of the ADRF). To this end, it would be good to know, in addition to the features described above, whether there are additional data reporting or access requirements for participants such as:
    a. Are there other metrics you require in the interface around data usage and requests?
    b. Is there a specific process involved in approving access to datasets that should be accommodated for in the interface (i.e., what institutional roles must be involved in the approval process, what steps must be completed in what order to fulfill a data access request, or what baseline set of policies must be reviewed and agreed to before there is access to any of the data)?
    c. Are there any legal requirements for signing and keeping documents pertaining to data access and usage?
2. Build out functionality for compiling data usage metrics for reports and day-to-day monitoring. As such, we recommend participants compile a set of required metrics they might need to see or are legally required to gather in relation to the use of restricted data as these metrics would take top priority in further development of the module.
3. Refine existing data export procedures to be incorporated into the Data Stewardship module. Currently data exports are handled by a set of python scripts which send alerts to system administrators and policy experts to review data export requests and approve or deny the export of research data products from the system in accordance with relevant, laws and terms of use. That functionality would need to be integrated more closely with DF Admin and the Data Stewardship module. To this end, it would be

[1] https://en.wikipedia.org/wiki/Lightweight_Directory_Access_Protocol

beneficial to know what general policy there exists within individual institutions around which products of data analysis (e.g., charts and visualizations, aggregate data summary tables, or results of computations and prediction algorithms) are permissible to be exported and used in publications and public presentations.

4. As mentioned above, if participants prefer the storage of the actual data to remain at their own institutions – only allowing DF Admin to store metadata about these datasets – additional work would need to be conducted to design useful integrations with the individual institutions' infrastructures.

Below we detail the implementation of the ADRF Data Stewardship module and the above customizations specifically within the two separate scenarios of 1) each individual institution using their own stand-alone versions of the ADRF Data Stewardship module that then connects with installations at partner institutions and 2) one instance of the ADRF Data Stewardship module implemented in one place for the usage of all partner institutions centrally.

4.2 Scenario 1

In the scenario where each institution installed their own instance of the ADRF Data Stewardship module, the benefit would be the ability to customize workflows around management of data access and usage at the potential expense of not being able to harmonize these workflows and policies across the different institutions as easily. Similarly, this approach provides for more customized metrics for reporting and monitoring data usage, with a possible drawback that there would be less commonality between the measurements used across institutions to track and audit data access and usage. Finally, the potential for more customization across different installations might require some items to be implemented over others as limited development resources might not be available to fulfill all feature requests across all sites at once or such development would come at a greater expense.

Benefits

1. Greater control by each institution over access requests and monitoring.
2. Potential for greater customization of features where resources for customization exist.

Considerations

1. Potentially less harmonization of practices, policies, procedures, and metrics across institutions.
2. Greater customization across various sites would increase the budget for development.
3. Each institution would be responsible for maintaining individual instances.

4.3 Scenario 2

In the scenario where one installation of the ADRF Data Stewardship module is run centrally for all institutions, there is the benefit of greater harmonization of practices, policies, procedures, and metrics at the expense of less customized features available for each participant. Over time, however, it is feasible that feature requests from individual institutions could be developed into the module should other institutions agree to and support their development. This would further create a new set of standardized practices and workflows that can referenced as an innovation in the management and sharing of restricted datasets.

Benefits

1. Engenders more standard and harmonized access policies for restricted data.
2. Potential for customization of features where other cooperative members agree such customization would benefit the community.
3. Shared resources would keep costs down.

Considerations

1. Less customization and control on an institution-by-institution level.
2. Institutions would have to agree on an organizational structure that conforms to regional law and policy for managing data access across institutions.