ScHARe Repository
Introduction

October 16, 2024

Deborah Duran, PhD · NIMHD
Luca Calzoni, MD MS PhD Cand. · NIMHD
Elif Dede Yildirim, PhD · NIMHD

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Outline

15'	ScHARe Overview
5'	Repository Background
5'	Getting Started
15'	Uploading your first Data Set
15'	HANDS ON: Uploading Data
15'	CDE Mapping and Dataviews
15'	HANDS ON: Dataviews for CDE Mapping
5'	Sharing Data
10'	Data Aggregation and Analysis - Overview
15'	Conclusion and Q&A

Experience poll

Please check your level of experience with the following:

	None	Some	Proficient	Expert
Python	☐	☐	☐	☐
R	☐	☐	☐	☐
Cloud computing	☐	☐	☐	☐
Terra	☐	☐	☐	☐
Health disparities research	☐	☐	☐	☐
Health outcomes research	☐	☐	☐	☐
Algorithmic bias mitigation	☐	☐	☐	☐

Interest poll

I am interested in (check all that apply):

- ☐ Learning about Health Disparities and Health Outcomes research to apply my data science skills
- ☐ Conducting my own research using AI/cloud computing and publishing papers
- ☐ Connecting with new collaborators to conduct research using AI/cloud computing and publish papers
- ☐ Learning to use AI tools and cloud computing to gain new skills for research using Big Data
- ☐ Learning cloud computing resources to implement my own cloud
- ☐ Developing bias mitigation and ethical AI strategies
- ☐ Other

ScHARe

What is ScHARe?

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe

Science
collaborative for
Health disparities and
Artificial intelligence bias
Reduction

Register: nimhd.nih.gov/schare

ScHARe is a cloud-based population science data platform designed to accelerate research in health disparities, health and healthcare delivery outcomes, and artificial intelligence (AI) bias mitigation strategies

ScHARe aims to fill five critical gaps:

- Increase participation of **women & underrepresented populations with health disparities** in data science through data science skills training, cross-discipline mentoring, and multi-career level collaborating on research
- Leverage population science, SDoH, and behavioral Big Data and cloud computing tools to foster a **paradigm shift** in health disparity and healthcare delivery outcomes research
- **Advance AI bias mitigation and ethical inquiry** by developing innovative strategies and securing diverse perspectives
- Provide a **data science cloud computing resource** for community colleges and low resource minority serving institutions and organizations
- Offer a **project data repository** centered on core common data elements for enhanced data interoperability and compliance with NIH Data Management and Sharing Policy

ScHARe

Google Platform
Terra Interface

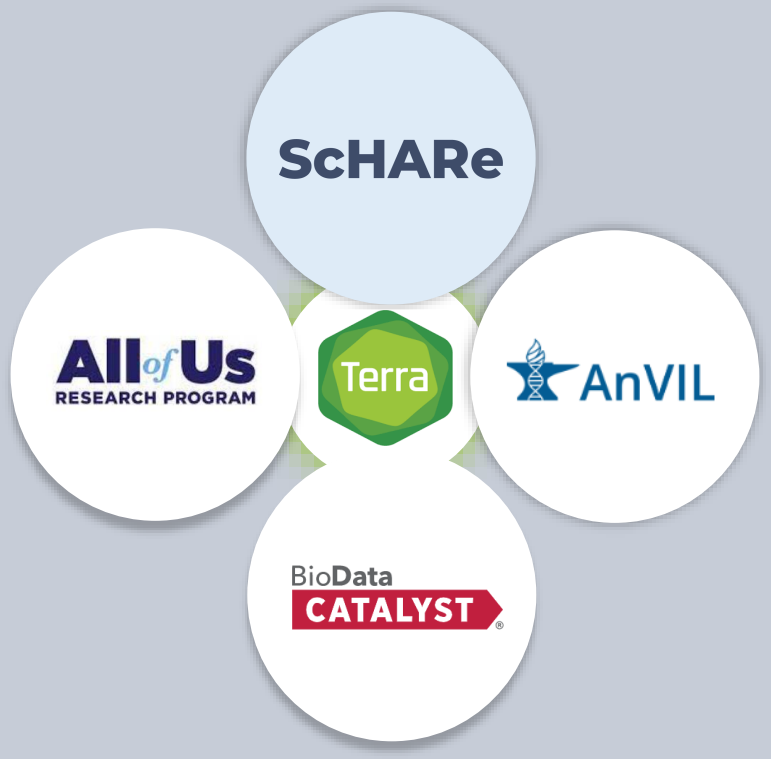
- Secure workspaces
- Data storage
- Computational resources
- Tutorials (how to)
- Copy-and-paste code in Python and R
- Learning Terra on ScHARe prepares you to use other NIH platforms

Terra recommends using **Chrome**

Must have a **Gmail** friendly account

PREPARING FOR AI RESEARCH AND HEALTHCARE USING BIG DATA

Mapping across cloud platforms with Terra interface for collaborative research



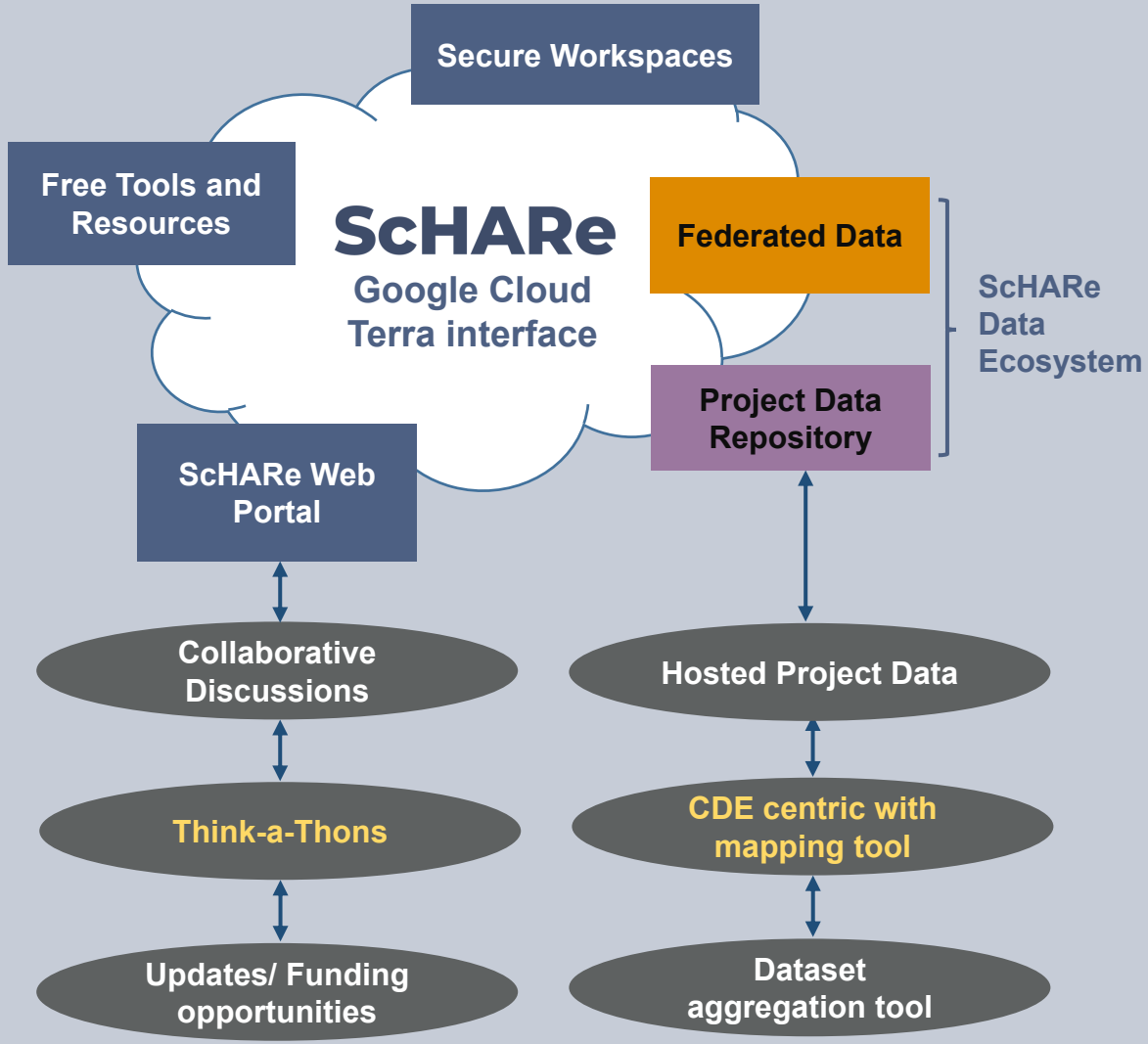
BE A PART OF THE FUTURE OF KNOWLEDGE GENERATION

ScHARe Components

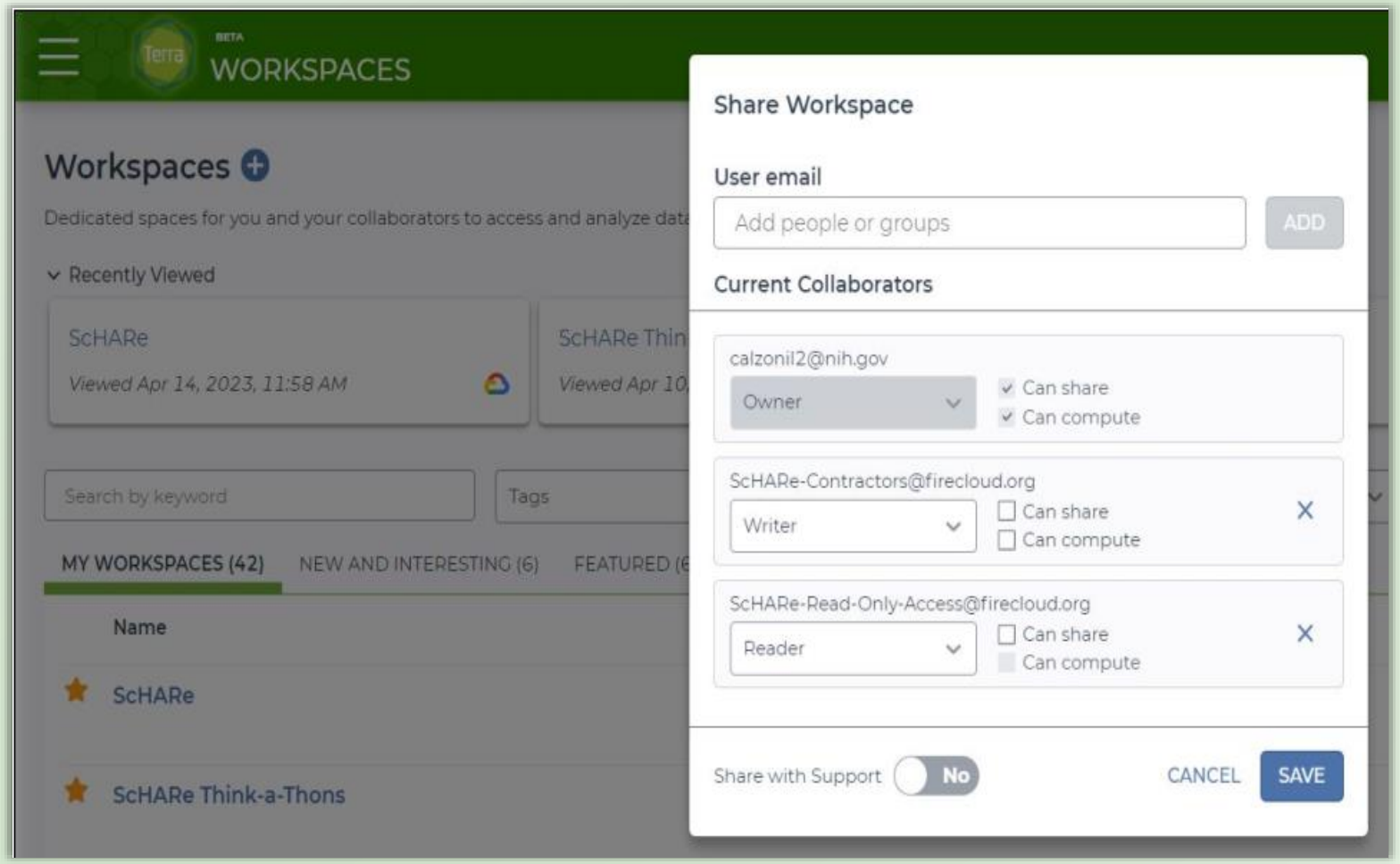
ScHARe co-localizes within the cloud:

1. **Datasets** (including social determinants of health and social science data) relevant to minority health, health disparities, and healthcare outcomes research
2. **CDE-focused data repository** to comply with the required hosting and sharing of data from NIMHD-/NINR-funded programs
3. **User-friendly computational capabilities and secure, collaborative workspaces** for students and all career level researchers
4. **Tools for collaboratively evaluating and mitigating biases** associated with datasets and algorithms utilized to inform healthcare and policy decisions (*upcoming*)

Intramural and Extramural Resource



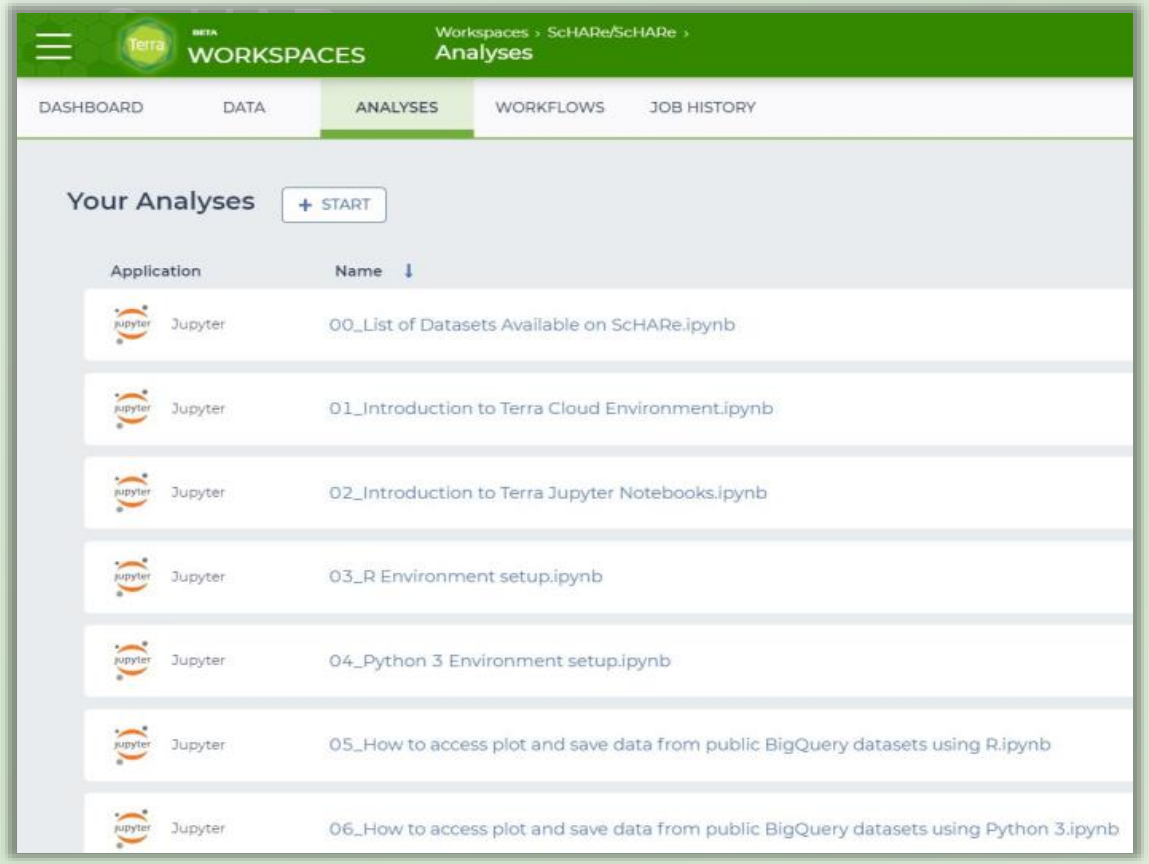
ScHARe Terra interface: secure workspace



- Secure workspace **for self or collaborative research**
- **Assign roles**: review or admin
- **Host own data and code**

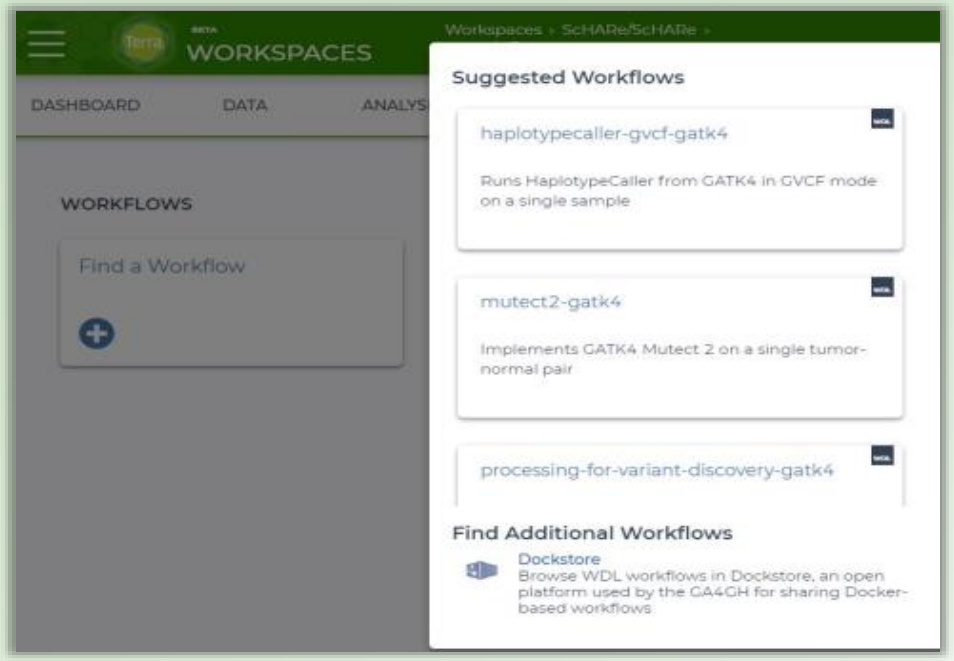
ScHARe Terra interface: analyses

Notebooks for analytics and tutorials



Modular codes

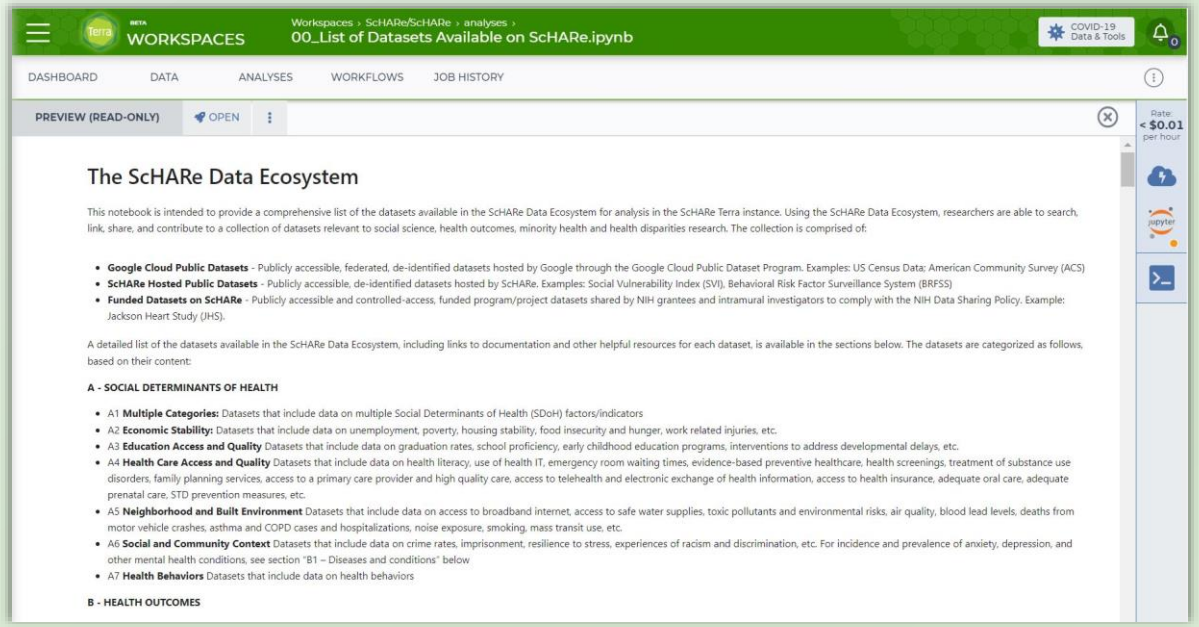
- **Easy-to-use copy-and-paste analytics**



- Modular codes developed for reuse
- **Adding SAS**

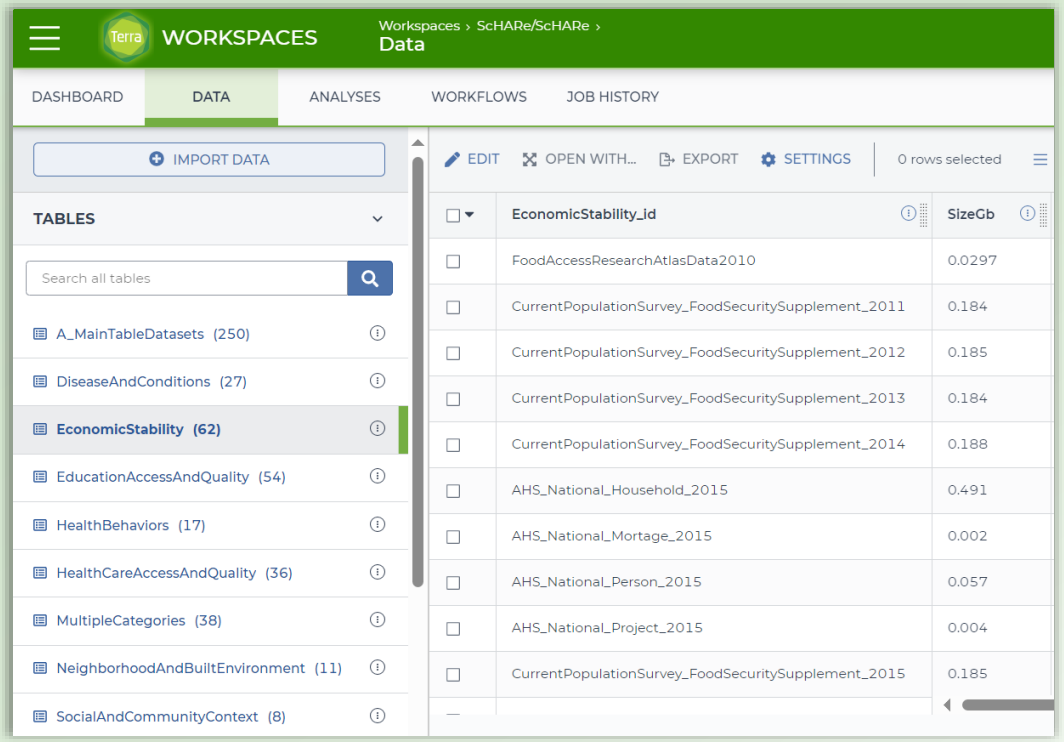
ScHARe Terra interface: access to datasets

What data?



In the **Analyses tab**, the notebook **00_List of Datasets Available on ScHARe** lists all datasets

Where?



In the **Data tab**, data tables help access data

ScHARe Ecosystem structure

Researchers can access, link, analyze, and export **a wealth of SDoH and population science related datasets** within and across platforms relevant to research about health disparities, health care delivery, health outcomes and bias mitigation, including:

250+
FEDERATED PUBLIC DATASETS

Public datasets

Publicly accessible, federated, de-identified datasets hosted by ScHARe or hosted by Google through the Google Cloud Public Dataset Program

ScHARe **e.g.:** *Behavioral Risk Factor Surveillance System (BRFSS)*

Google **e.g.:** *American Community Survey (ACS)*

CDE
FOCUSED REPOSITORY

Funded datasets

Publicly accessible and controlled-access, funded program/project datasets using Common Data Elements shared by NIH grantees and intramural investigators to comply with the NIH Data Sharing Policy

e.g.: *Jackson Heart Study (JHS)*
Extramural Grant Data
Intramural Project Data

Innovative Approach:
CDE Concept Codes
Uniform Resource Identifier (**URI**)

ScHARe Ecosystem

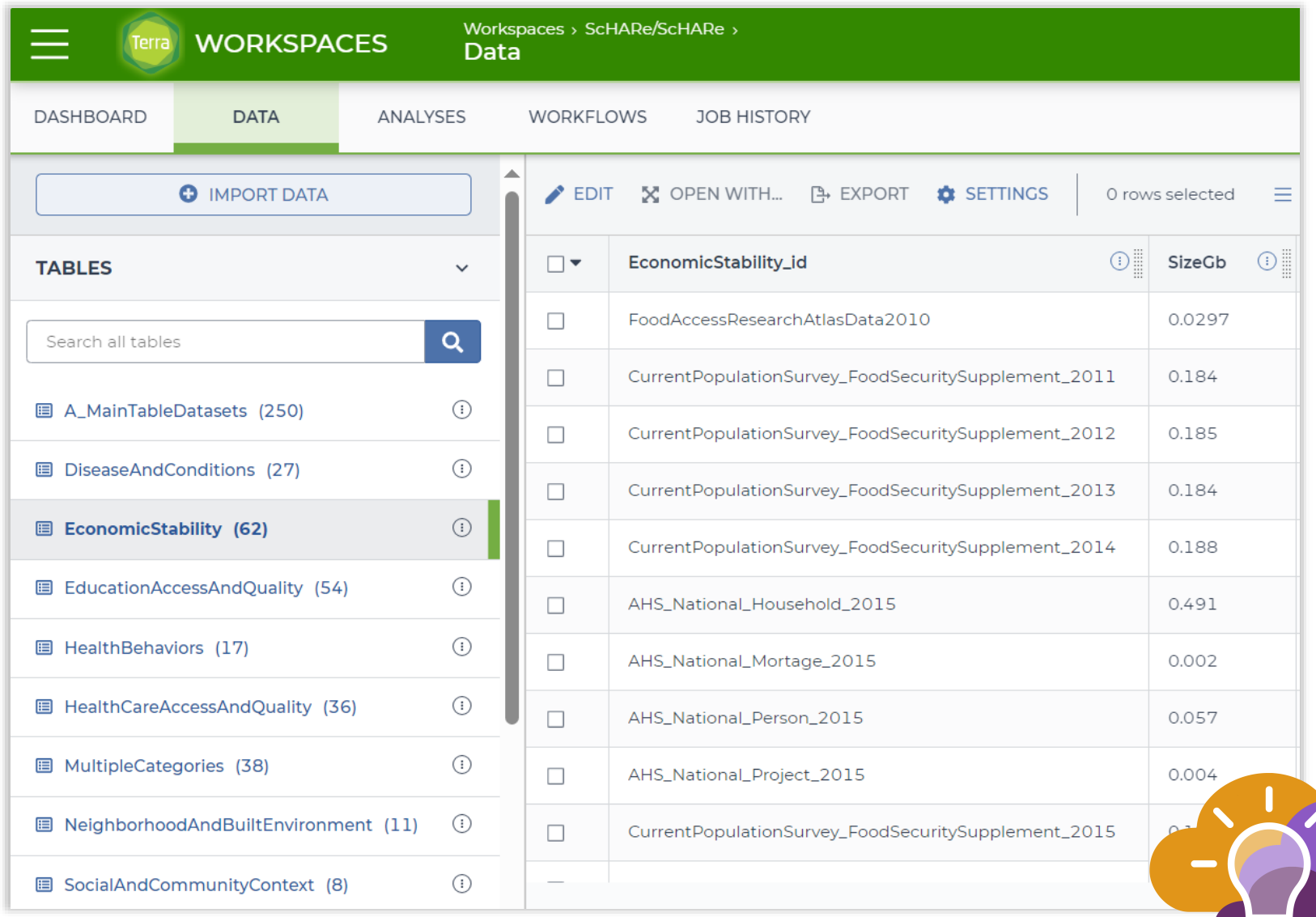
OVER 260 DATA SETS CENTRALIZED

Datasets are categorized by content based on the CDC **Social Determinants of Health categories**:

1. Economic Stability
2. Education Access and Quality
3. Health Care Access and Quality
4. Neighborhood and Built Environment
5. Social and Community Context

with the addition of:

- **Health Behaviors**
- **Diseases and Conditions**



ScHARe Ecosystem: ScHARe hosted datasets

Organized based on the **CDC SDoH categories**, with the addition of *Health Behaviors* and *Diseases and Conditions*:

260+ datasets

- **What are the Social Determinants of Health?**

Social determinants of health (SDoH) are the **nonmedical factors that influence health outcomes**

They are the **conditions in which people are born, grow, work, live, and age, and the wider set of forces and systems shaping the conditions of daily life**

https://www.cdc.gov/about/priorities/social-determinants-of-health-at-cdc.html?CDC_AAref_Val=https://www.cdc.gov/about/sdoh/index.html

ScHARe Ecosystem: ScHARe hosted datasets

Education access and quality

Data on graduation rates, school proficiency, early childhood education programs, interventions to address developmental delays, etc.

Economic stability

Data on unemployment, poverty, housing stability, food insecurity and hunger, work related injuries, etc.

Health care access and quality

Data on health literacy, use of health IT, preventive healthcare, access to health insurance, etc.

* Health behaviors

Data on health-related practices that can directly affect health outcomes.

Neighborhood and built environment

Data on access to safe water supplies, toxic pollutants and environmental risks, air quality, blood lead levels, noise exposure, smoking, mass transit use, etc.

* Diseases and conditions

Data on incidence and prevalence of specific diseases and health conditions.

Social and community context

Data on crime rates, imprisonment, resilience to stress, experiences of racism and discrimination, etc.

** Not Social Determinants of Health*

ScHARe Ecosystem: Google hosted datasets

Examples of interesting datasets include:

- **American Community Survey** (U.S. Census Bureau)
- **US Census Data** (U.S. Census Bureau)
- **Area Deprivation Index** (BroadStreet)
- **GDP and Income by County** (Bureau of Economic Analysis)
- **US Inflation and Unemployment** (U.S. Bureau of Labor Statistics)
- **Quarterly Census of Employment and Wages** (U.S. Bureau of Labor Statistics)
- **Point-in-Time Homelessness Count** (U.S. Dept. of Housing and Urban Development)
- **Low Income Housing Tax Credit Program** (U.S. Dept. of Housing and Urban Development)
- **US Residential Real Estate Data** (House Canary)
- **Center for Medicare and Medicaid Services - Dual Enrollment** (U.S. Dept. of Health & Human Services)
- **Medicare** (U.S. Dept. of Health & Human Services)
- **Health Professional Shortage Areas** (U.S. Dept. of Health & Human Services)
- **CDC Births Data Summary** (Centers for Disease Control)
- **COVID-19 Data Repository by CSSE at JHU** (Johns Hopkins University)
- **COVID-19 Mobility Impact** (Geotab)
- **COVID-19 Open Data** (Google BigQuery Public Datasets Program)
- **COVID-19 Vaccination Access** (Google BigQuery Public Datasets Program)

How to access Google hosted datasets

Big Query

The Google public datasets are **available for access on Terra using BigQuery**

- BigQuery is the Google Cloud storage solution for structured data
- It is easy to use, works with large amounts of data and offers fast data retrieval and analysis
- Our **instructional notebooks in the Analyses tab** provide code and instructions on using Big Query to access Google datasets

ScHARe Datasets PDF list

ScHARe

The ScHARe Data Ecosystem

This document is intended to provide a comprehensive list of the datasets available in the ScHARe Data Ecosystem for analysis in the ScHARe Terra instance. Using the ScHARe Data Ecosystem, researchers are able to search, link, share, and contribute to a collection of datasets relevant to social science, health outcomes, minority health and health disparities research.

The collection is comprised of:

- **Google-hosted Public Datasets** - Publicly accessible, federated, de-identified datasets hosted by Google through the Google Cloud Public Dataset Program. Examples: US Census Data; American Community Survey (ACS)
- **ScHARe-hosted Public Datasets** - Publicly accessible, de-identified datasets hosted by ScHARe. Examples: Social Vulnerability Index (SVI), Behavioral Risk Factor Surveillance System (BRFSS)
- **ScHARe-hosted Project Datasets** - Publicly accessible and controlled-access, funded program/project datasets shared by NIH grantees and intramural investigators to comply with the [illegible] Jackson Heart Study (JHS)

Scan me

bit.ly/ScHARe-datasets

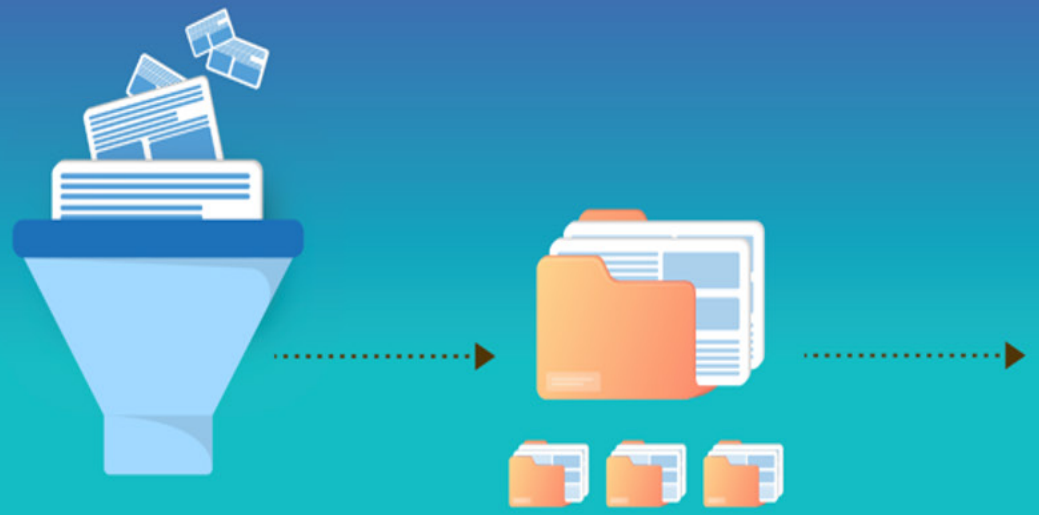
CDE benefits:

- Faster start-up for project
- Better data aggregation across projects
- Shared meaning
- Concept-focused to allow questions/answers variations
- Coding enables an URI approach for better data interoperability

A **Common Data Element** (CDE) is a standardized, precisely defined question, paired with a set of allowable responses, used systematically across different sites, studies, or clinical trials to ensure consistent data collection

Because Researchers use CDEs…

they can more quickly share data and get results faster, which ultimately can help make a **meaningful difference to our nation's health.**



For more information about how CDEs accelerate research discoveries, visit: cde.nlm.nih.gov/resources

ScHARe Core CDEs

PhenX Toolkit

NIH Endorsed

- Age
- Birthplace
- Zip Code
- Race and Ethnicity
- Sex
- Gender
- Sexual Orientation
- Marital Status
- Education
- Annual Household Income
- Household Size

- English Proficiency
- Disabilities
- Health Insurance
- Employment Status
- Usual Place of Health Care
- Financial Security / Social Needs
- Self-Reported Health
- Health Conditions (and Associated Medications/Treatments)

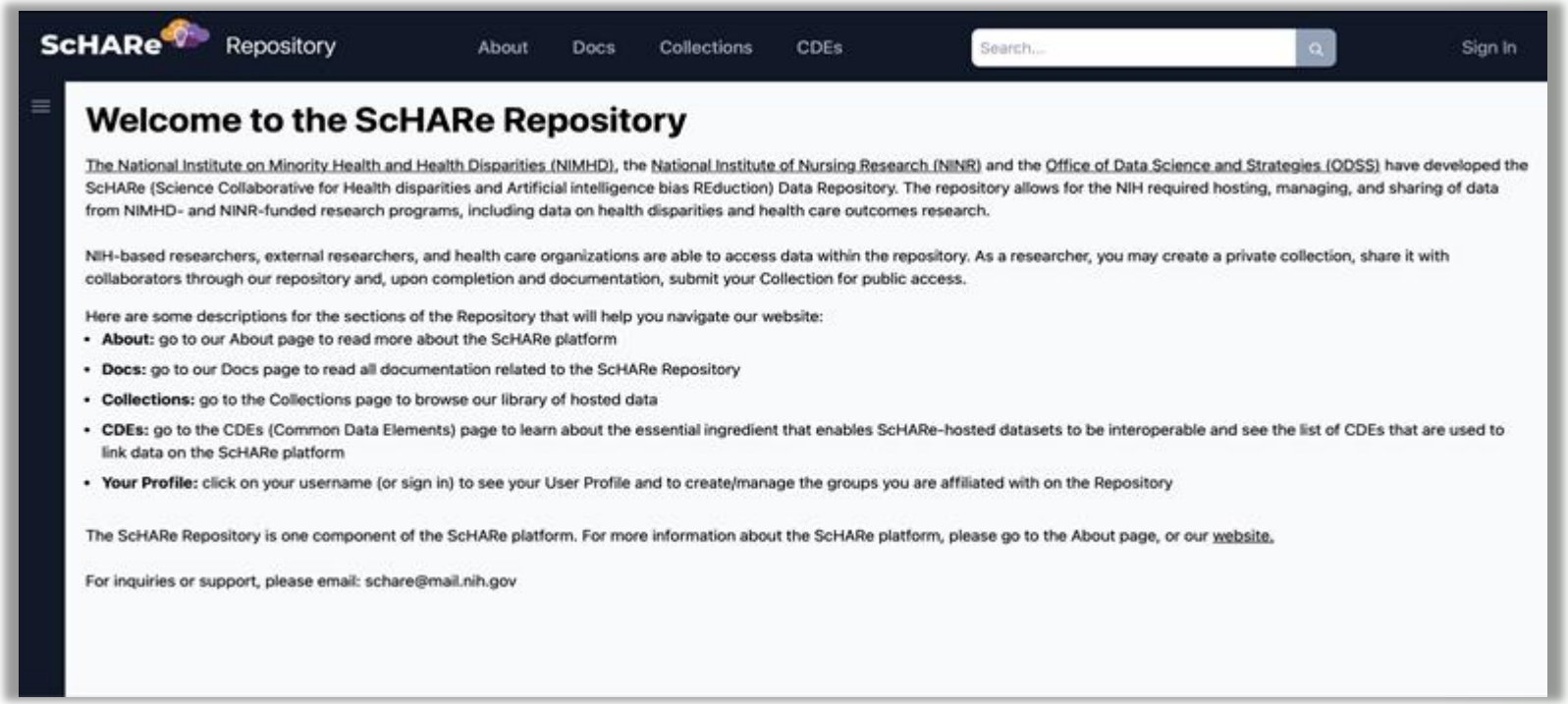
- **NIMHD Framework***
- **Health Disparity Outcomes***

* Project Level CDEs

ScHARe has developed **Common Data Elements** to ensure consistent data collection across studies, facilitate interoperability, and link data from different sources

NIH CDE Repository:
cde.nlm.nih.gov/home
PhenX Toolkit:
www.nimhd.nih.gov/resources/phenx/

ScHARe Repository



- Host your project data in a **safe space** with privacy levels, secure workspaces, collaboration platform
- Comply with **NIH Data Management and Data Sharing Policy**
- **Focus**: Social Science, SDoH, Health Disparities, Health Outcomes Research
- **CDE centric:** Map project CDEs or variables to ScHARe-PhenX CDEs
- **Link your data** with others and federated data

ScHARe

Research Think-a-Thons

- Novice **training webinars** for data science, cloud computing and research using Big Data
- **Target:** underrepresented populations, women, racial/ethnic and sexual gender minorities, rural and poor populations

Generational career & discipline exchange

Think-a-Thons

3rd Wednesday of every month 2 pm

Goals:

- Upskill underrepresented populations in data science and cloud computing
- Foster a research paradigm shift to use Big Data in health disparities/health outcomes research
- Promote use of Dark Data

1. TUTORIAL AND TARGETED THINK-A-THONS

- Monthly sessions (2 1/2 hours)
- Instructional/interactive
- Designed for new/experienced users
- Networking
- Mentoring and coaching
- Topics include:
  - Data Science 101
  - Terra
  - Social Determinants of Health analytics
  - Common Data Elements
  - AI readiness
  - Ethical and transparent AI
  - Bias mitigation

Launched April 2024

2. RESEARCH THINK-A-THONS

- Multi-career (students to senior investigators)
- Multi-discipline (data scientists and researchers)
- Featured datasets with guest experts leads
- Guest experts in topic areas, analytics, data sources etc. to provide guidance
- Generate research idea - decide design, datasets and analytics
- Learn Ethical AI
- Publications

Register:

bit.ly/think-a-thons

Think-a-Thon tutorials

bit.ly/think-a-thons

February	**Artificial Intelligence and Cloud Computing 101**
March	**ScHARe 1 – Accounts and Workspaces**
April	**ScHARe 2 – Terra Datasets**
May	**ScHARe 3 – Terra Google-hosted Datasets**
June	**ScHARe 4 – Terra ScHARe-hosted Datasets**
July	**An Introduction to Python for Data Science – Part 1**
August	**An Introduction to Python for Data Science – Part 2**
September	**ScHARe 5: A Review of the ScHARe Platform and Data Ecosystem**
October	**Preparing for AI 1: Common Data Elements and Data Aggregation**
November	**Preparing for AI 2: An Introduction to FAIR Data and AI-ready Datasets**
January	**Preparing for AI 3: Computational Data Science Strategies 101**
February/March	**Preparing for AI 4: Overview Prep for AI Summary with Transparency, Privacy, Ethics**
April	**Research Teams – SDoH and Health Disparities**
May	**Be a Part of the Future of Knowledge Generation 1: AI/Cloud Computing Basics and CDEs**
July	**Be a Part of the Future of Knowledge Generation 2: AI-Ready Datasets and Computations**

SPECIAL EVENTS

- ScHARe for **Educators** (Community Colleges and low-resource MSIs)
- ScHARe for **American Indian/ Alaska Native Researchers**
- ScHARe for **Coders and Programmers** to conduct research

Experience conducting ethical AI

Transparency

Public perception and understanding of how AI works

- **Technical documentation for duplication/re-use**

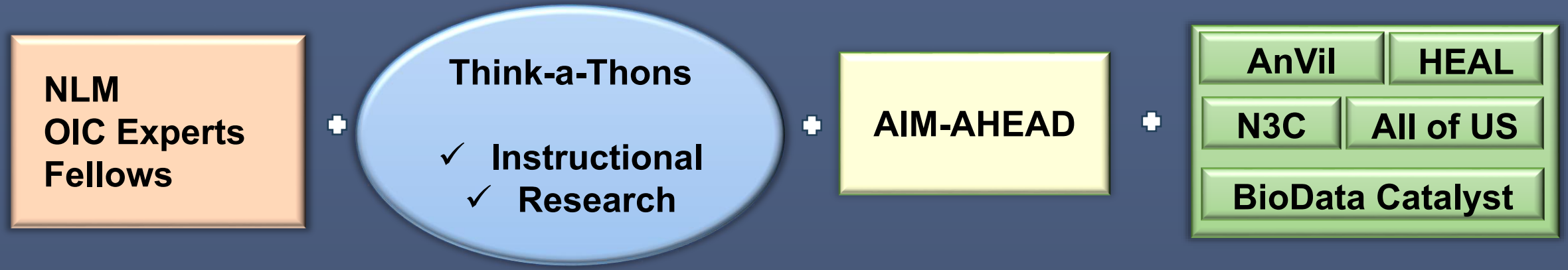
- **Tools:**
  - Data dictionary
  - Health sheet (Data sheet)
  - Model cards (capabilities and purpose of algorithms are openly and clearly communicated to relevant stakeholders)

Fairness

Findable: *providing metadata, documentation, and clear identifiers*
Accessible: *wide audience*
Interoperable: *standardized formats and APIs enable seamless integration*
Reusable: *clear documentation, licensing, reduce redundancy*

- Metadata and data should be easy to find for both humans and computers
- Ensure that data represents relevant populations

Think-a-Thons training/mentoring pipeline



Using AI experts *to train and mentor novice AI users* *to upskill and mentor diverse perspectives in AI* *to increase diverse perspectives in biomedical research*

Goal: "Upskilling"

- ✓ Data science specialists into health disparities and health outcomes research
- ✓ Health disparities/outcomes researchers into using big data and cloud computing

Target Audience:

- ✓ Underrepresented populations (women, race/ethnic) users not trained in data science
- ✓ Data scientists with no or little research experience
- ✓ Resource and tool for Community Colleges and low-resource MSIs and organizations

ScHARe Repository
Introduction

October 16, 2024

Deborah Duran, PhD · NIMHD
Luca Calzoni, MD MS PhD Cand. · NIMHD
Elif Dede Yildirim, PhD · NIMHD

ScHARe

Repository Background

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

What is the ScHARe Repository?

The ScHARe Repository is a data repository for hosting and sharing population science data.

The repository also offers tools to explore, aggregate and manipulate that data.

Datasets Collection

Relevant to health disparities and health care outcomes research, including SDOH data.

Data Repository

Hosting, managing, and sharing of data from federated funded research programs across all communities

Bias Mitigation Tools + Resources

Manage bias associated with datasets and algorithms, informing ethical and inclusive decisions.

Collaborative Workspaces

Access and analyze public and controlled-access data in collaborative spaces.

What can you do with the Repository?

Upload your own data

Store collected data and annotate with a data dictionary. Align data to the ScHARe CDEs.

Harmonize data to CDEs

Map uploaded data to CDEs. Join your data with project or federated data via CDEs.

Browse for data

Find relevant federated national datasets or other project data.

Manipulate and aggregate data for analysis

Filter, sort, and select subsets for specific purposes. Link and aggregate datasets.

Control privacy levels and data sharing

Share confidential data among colleagues. Share public access data with the research community.

Data Analysis via ScHARe Terra or local analysis platform

Key Features of the ScHARe Repository

- Offers secure, long-term **storage** for project data
- Provides easy access to federated and other shared population science, Social Determinants of health (SDoH) **data sets**
- Leverages **CDEs** for organizing, aggregating, and harmonizing data
  - CDEs facilitate use of established semantic and coding systems
  - Aligns with ScHARe's focus on CDE centric data sharing
- Provides seamless tabular **data management**, aggregation, manipulation, and harmonization directly within the Repository
- Offers different **privacy levels** to meet individual data set requirements (public, restricted, confidential, private)
- Allows **sharing data** with other research collaborators

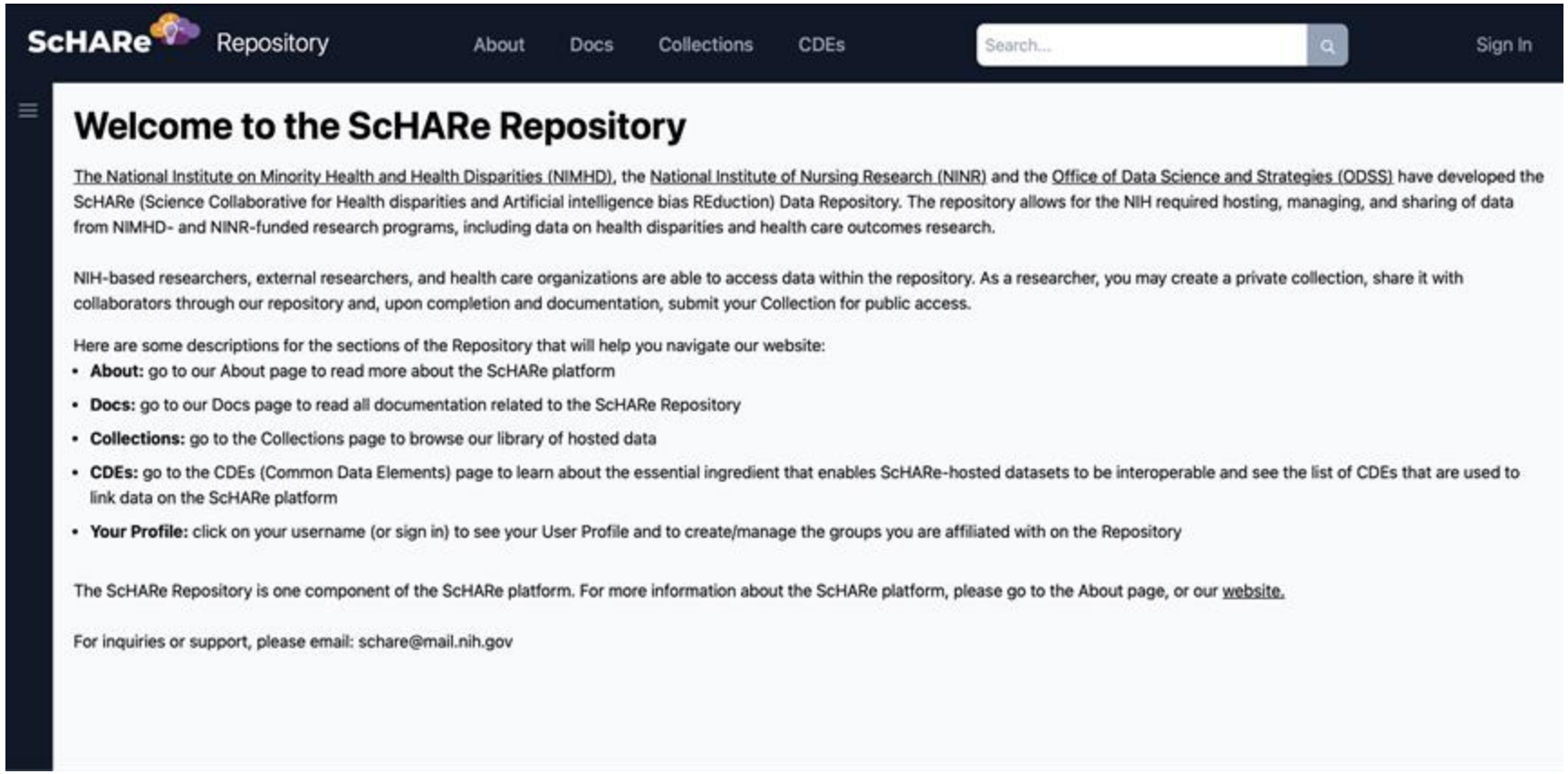
ScHARe

Getting Started

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

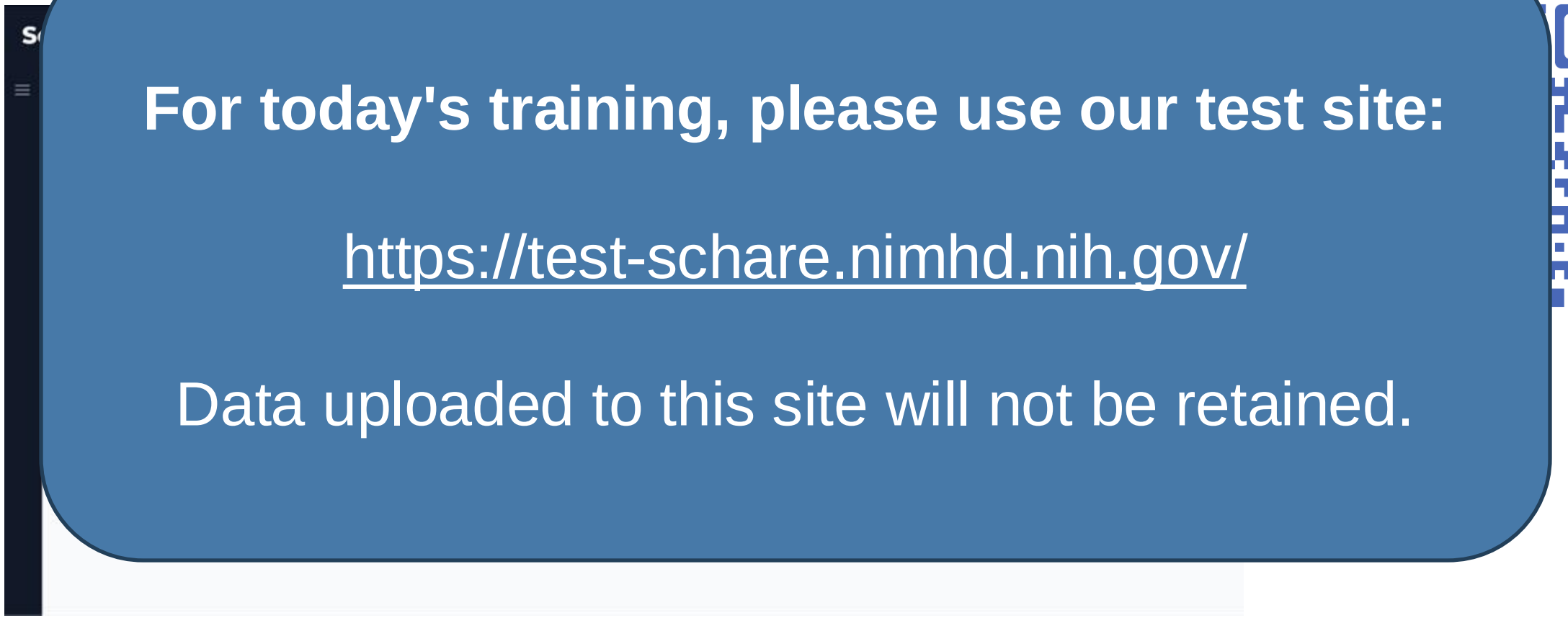
Visit the ScHARe Repository!

https://schare-repository.nimhd.nih.gov/



Visit the ScHARe Repository!

https://schare-repository.nimhd.nih.gov/

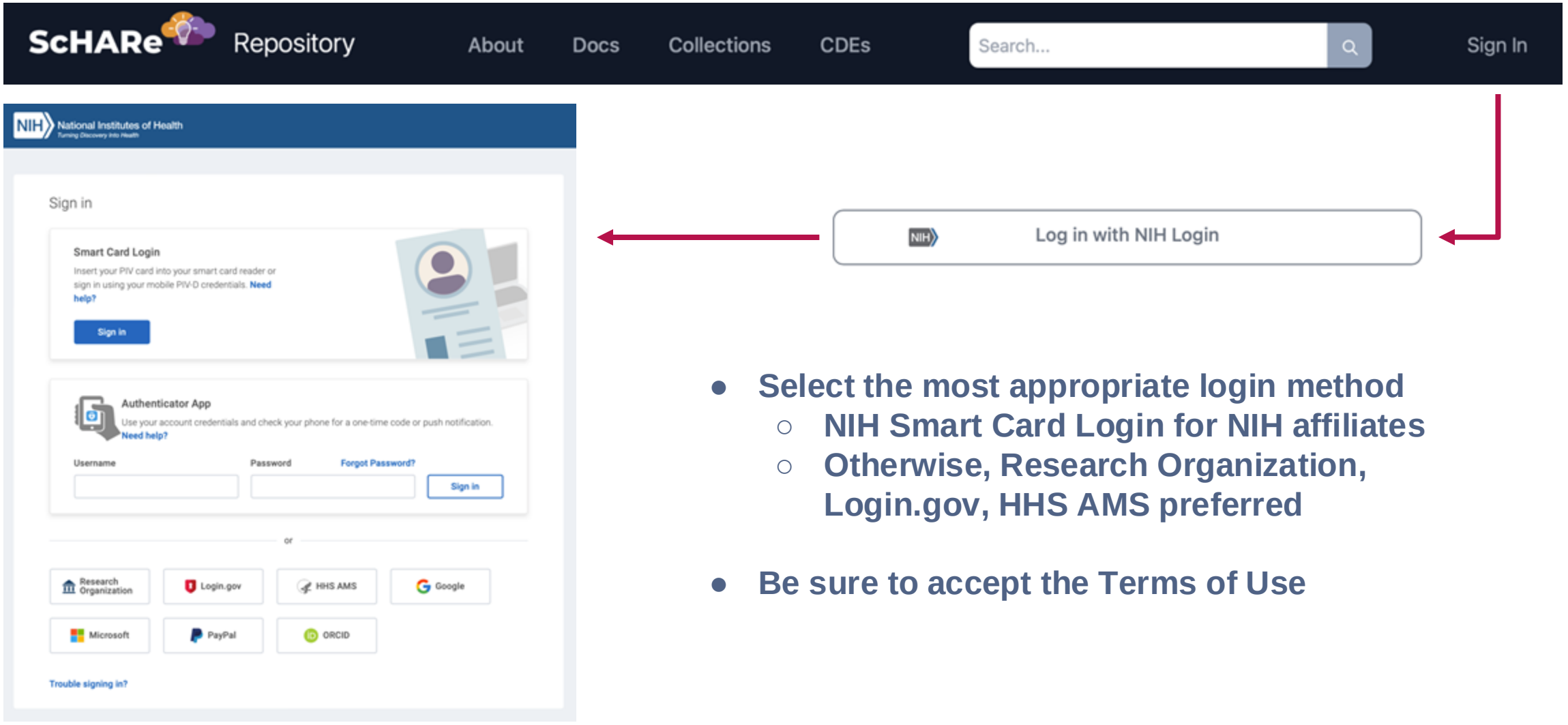


For today's training, please use our test site:

https://test-schare.nimhd.nih.gov/

Data uploaded to this site will not be retained.

Log In to the Repository



- Select the most appropriate login method
  - NIH Smart Card Login for NIH affiliates
  - Otherwise, Research Organization, Login.gov, HHS AMS preferred
- Be sure to accept the Terms of Use

A Quick Tour of the Site

Learn more about the site

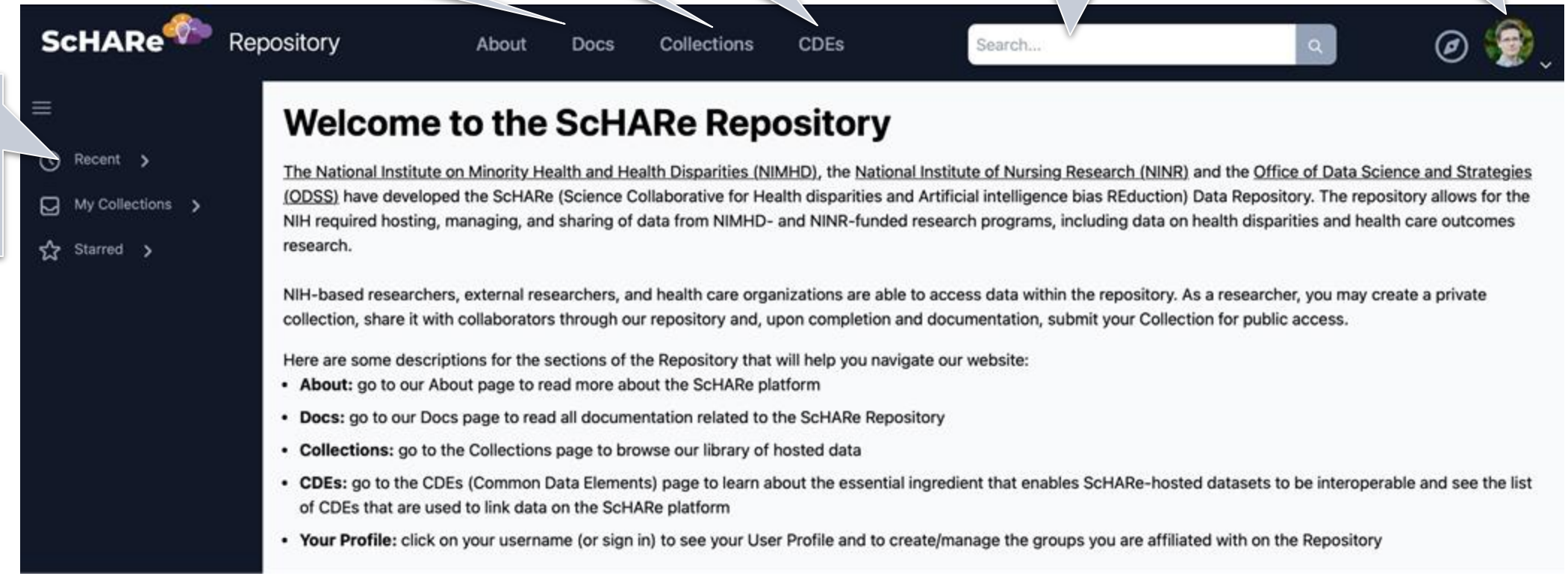
Browse for data (private and public)

View CDE definitions

Search across all data that you can see (private and public)

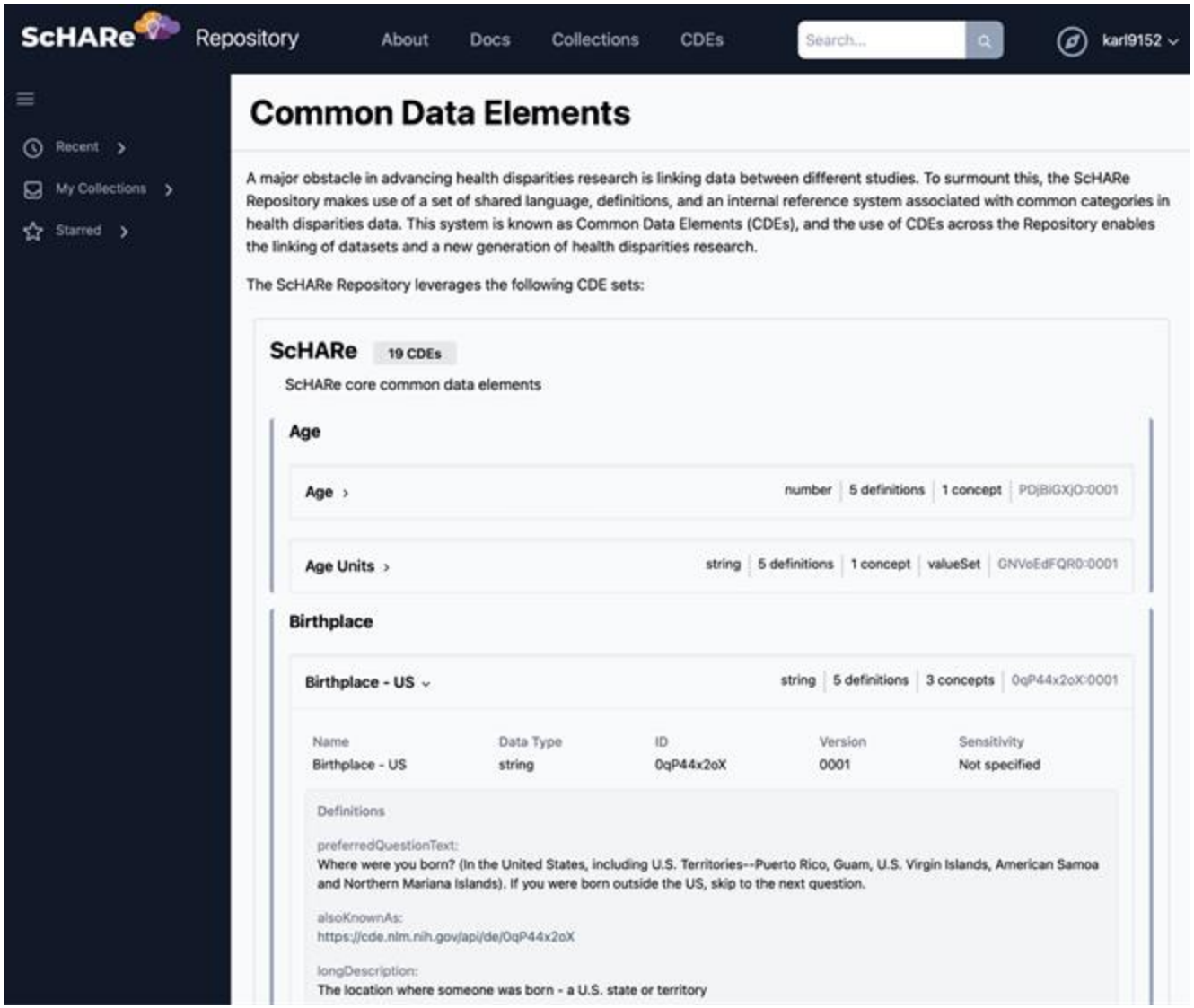
Log in, log out, and user information

Quick links to recently accessed and your own data



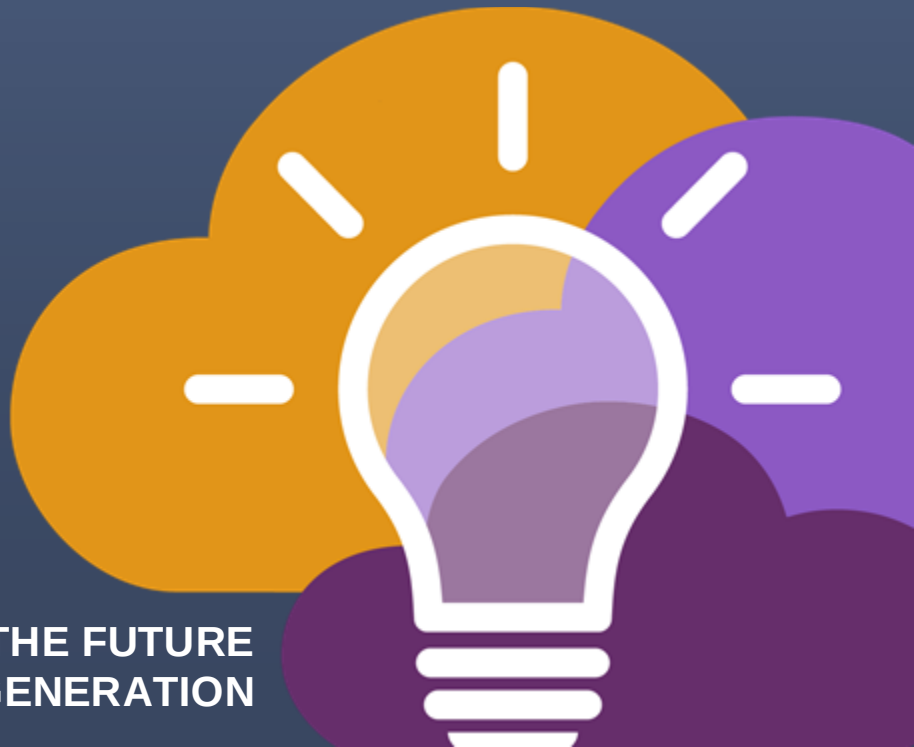
Viewing the ScHARe CDEs

- Browse through the list of ScHARe CDEs
- Expand an individual CDE's details to see definitions, permissible values, and concept codes



ScHARe

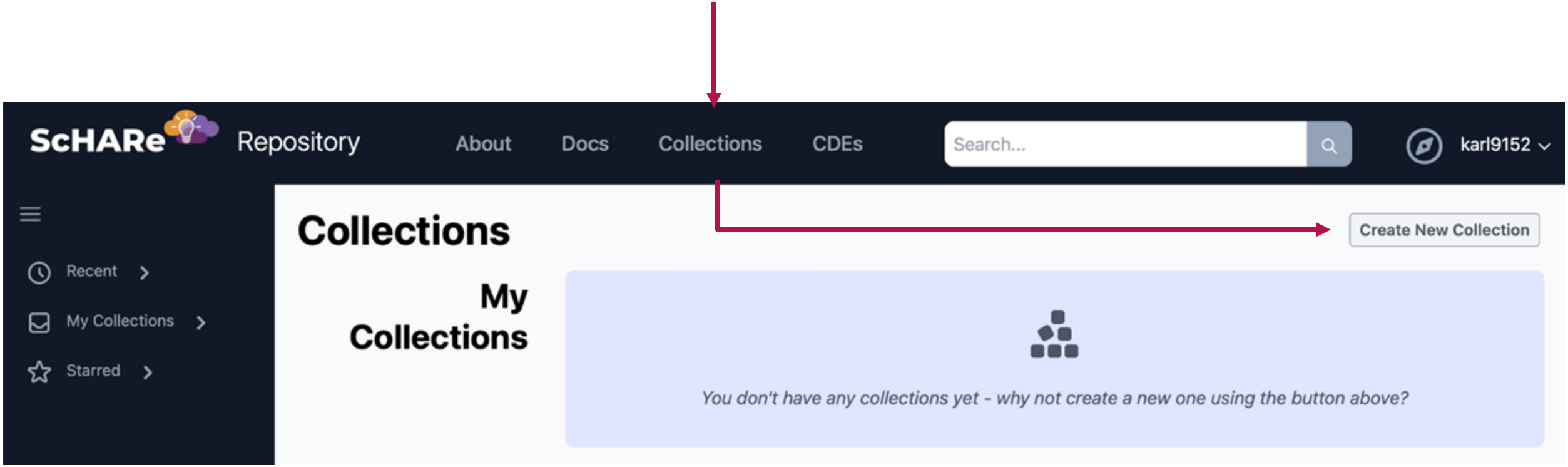
Uploading your first Data Set



BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

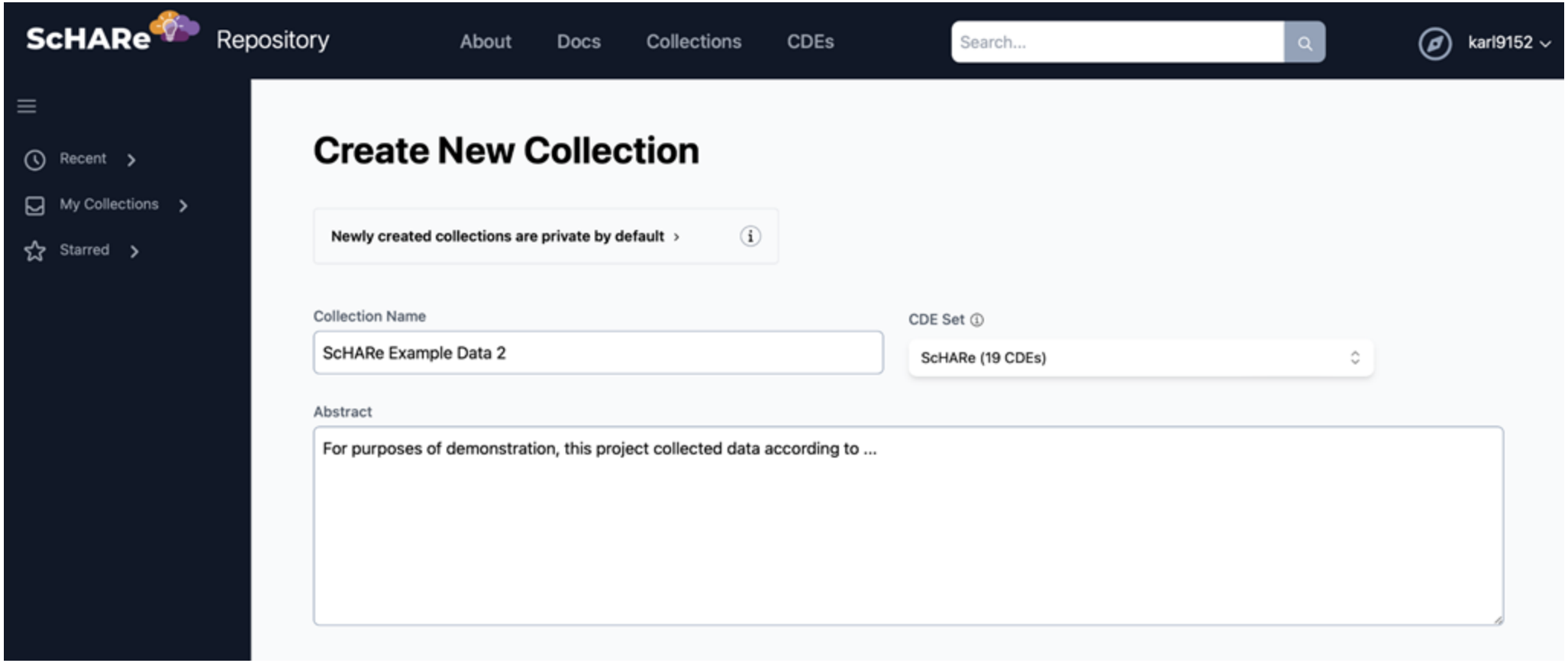
Creating a Collection

- **Collections** are a place where you can describe and store your data and any related metadata. They can be shared with colleagues and published when you're ready. You can upload any file type!
- The first step in uploading data is to create a collection



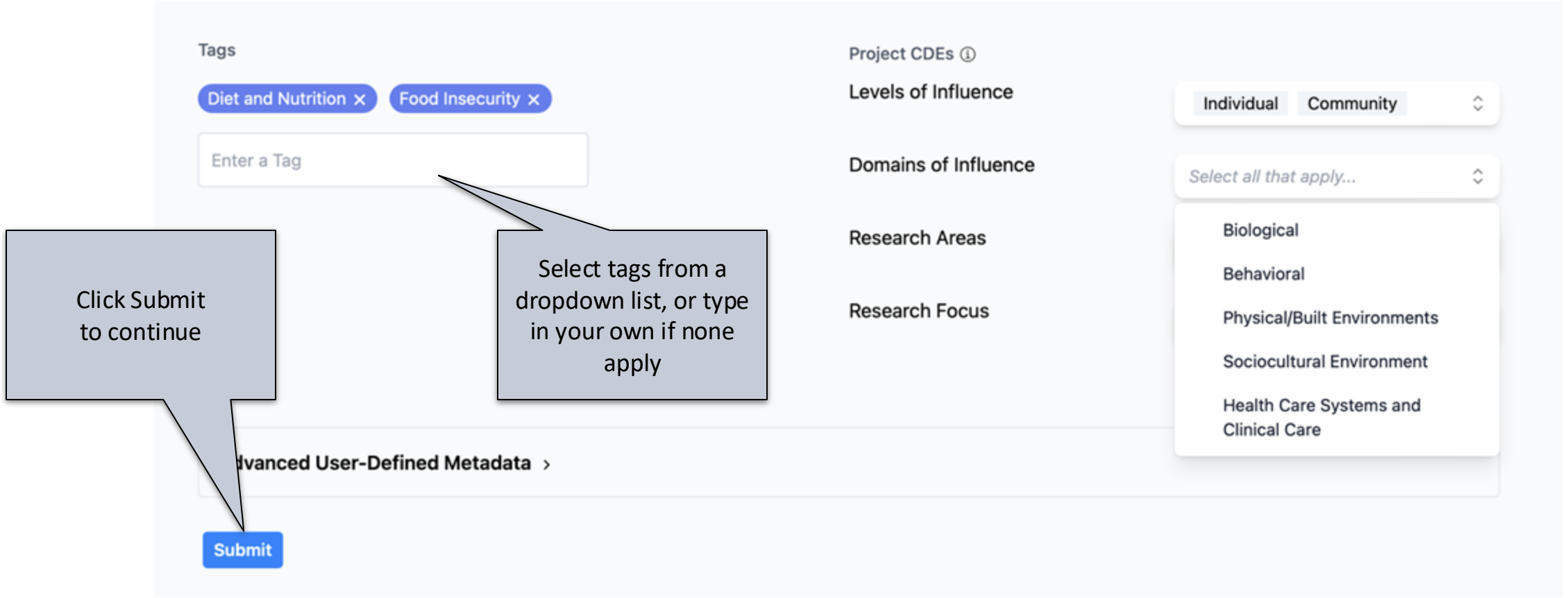
Creating a Collection

- Provide a readable name and a brief abstract (description) of the data to be uploaded
- These details can be updated later if needed



Tags and Project-Level CDEs

- Apply relevant tags and project-level CDEs to your dataset
- Refer to the NIMHD Research Framework for more information on selecting project-level CDEs:
  - https://www.nimhd.nih.gov/about/overview/research-framework/nimhd-framework.html



Project Level CCDEs – Framework

What NIMHD Research framework levels and domains of influence is your study targeting? (Select all that apply)

Levels of Influence

- ☐ Individual
- ☐ Interpersonal
- ☐ Community
- ☐ Societal

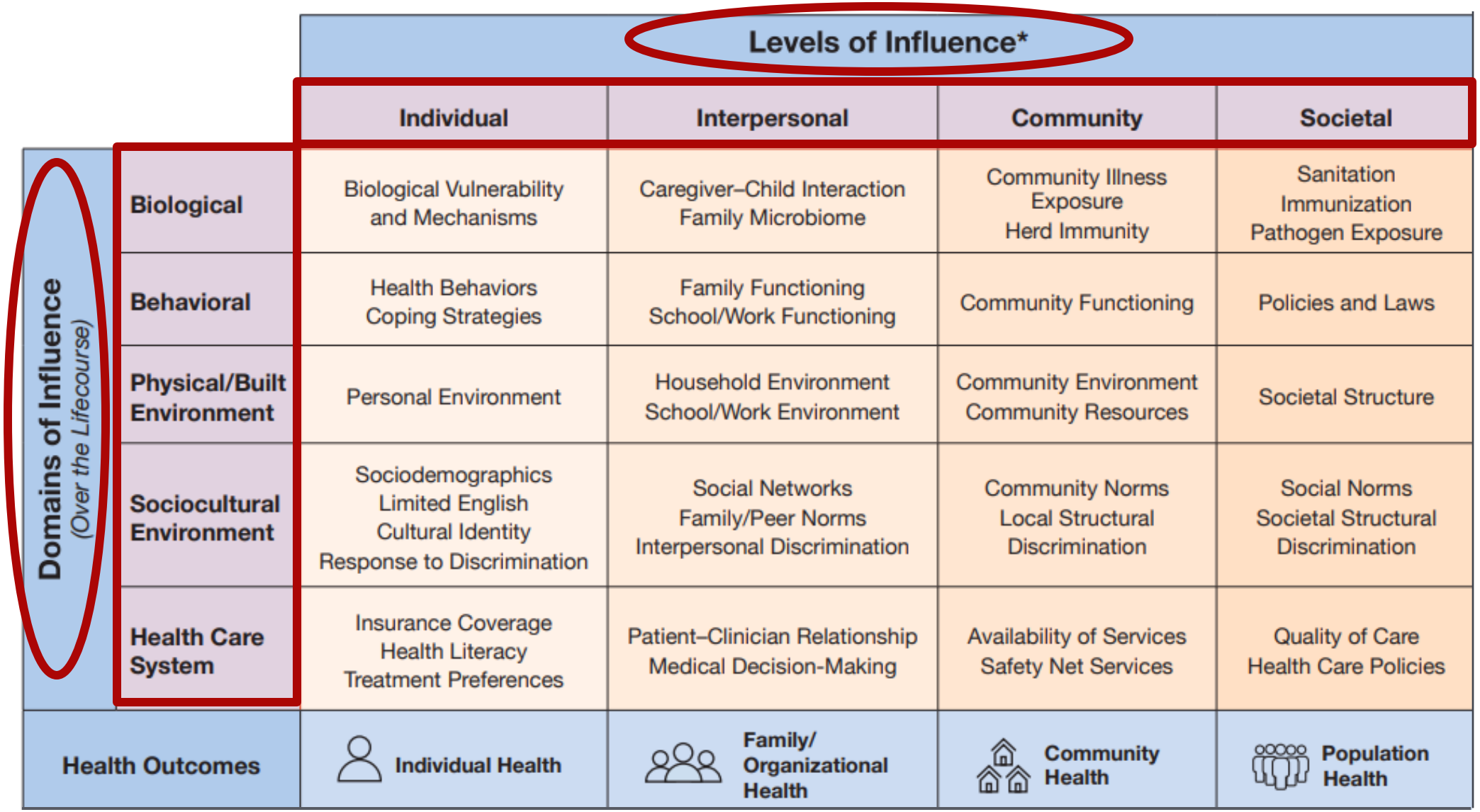
Domains of Influence

- ☐ Biological
- ☐ Behavioral
- ☐ Physical/Built Environments
- ☐ Sociocultural Environment
- ☐ Health Care Systems and Clinical Care

NIMHD Research Framework. https://www.nimhd.nih.gov/about/overview/research-framework/nimhd-framework.html

NIMHD Research Framework

bit.ly/NIMHD-RF



Domains of Influence (Over the Lifecourse)	Levels of Influence*			
	Individual	**Interpersonal**	**Community**	**Societal**
Biological	Biological Vulnerability and Mechanisms	Caregiver–Child Interaction Family Microbiome	Community Illness Exposure Herd Immunity	Sanitation Immunization Pathogen Exposure
Behavioral	Health Behaviors Coping Strategies	Family Functioning School/Work Functioning	Community Functioning	Policies and Laws
Physical/Built Environment	Personal Environment	Household Environment School/Work Environment	Community Environment Community Resources	Societal Structure
Sociocultural Environment	Sociodemographics Limited English Cultural Identity Response to Discrimination	Social Networks Family/Peer Norms Interpersonal Discrimination	Community Norms Local Structural Discrimination	Social Norms Societal Structural Discrimination
Health Care System	Insurance Coverage Health Literacy Treatment Preferences	Patient–Clinician Relationship Medical Decision-Making	Availability of Services Safety Net Services	Quality of Care Health Care Policies
Health Outcomes	**Individual Health**	**Family/ Organizational Health**	**Community Health**	**Population Health**

Project Level CCDEs – Research Area Focus

Which of the following content areas of research is this study addressing, if any? Select all that apply.

- ☐ Minority health
- ☐ Health Disparity (select the focus area)
  - ☐ Higher incidence and/or prevalence of disease, including earlier onset or more aggressive progression of disease
  - ☐ Premature or excessive mortality from specific health conditions
  - ☐ Greater global burden of disease, such as Disability Adjusted Life Years (DALY), as measured by population health metrics
  - ☐ Poorer health behaviors and clinical outcomes related to the aforementioned
  - ☐ Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions
- ☐ Other Health Outcomes / Health Delivery or care

Duran DG, Pérez-Stable EJ. Novel Approaches to Advance Minority Health and Health Disparities Research. Am J Public Health. 2019 Jan; 109(S1):S8-S10. doi: 10.2105/AJPH.2018.304931. PMID: 30699017; PMCID: PMC6356124. *ADAPTED with Other health outcomes delivery/care*

NIMHD's Mission: Improve Minority Health

Minority Health:
Distinctive health characteristics and attributes of racial and/or ethnic minority populations who are socially disadvantaged due in part to being subject to racist or discriminatory acts and are underserved in health care.

Minority Health Research

The scientific investigation of singular and combinations of attributes, characteristics, behaviors, biology, and societal and environmental factors that influence the health of minority racial and/or ethnic population(s), including within-group or ethnic sub-populations, with the goals of improving health and preventing disease.

Minority Health Populations

The OMB Directive 15 defines racial and ethnic minority populations as:

- American Indian or Alaska Native
- Asian
- Black or African American
- Hispanic or Latino American
- Native Hawaiian and Pacific Islander

NIMHD's Mission: Reduce Health Disparities

Health Disparity:
A health disparity is a health difference that adversely affects disadvantaged populations in comparison to a reference population, based on one or more health outcomes.

All populations with health disparities are socially disadvantaged due in part to being subject to racist or discriminatory acts and are underserved in health care.

Health Disparity Research

A multi-disciplinary field of study devoted to:

- Gaining greater scientific knowledge about the influence of health determinants.
- Understanding the role of mechanisms.
- Determining how this knowledge is translated into interventions to reduce or eliminate adverse health outcomes.

Populations with Health Disparities

Populations that experience health disparities include:

- Racial and ethnic minority groups
- People with lower socioeconomic status (SES)
- Underserved rural communities
- Sexual and gender minority (SGM) groups
- People with disabilities

Health Disparity Outcomes

Unfair disadvantages that people face in different aspects of life, like education, income, or opportunities **can lead to health disparities**

Some groups of people may experience poorer health outcomes than others as a result

Health Disparity Outcomes

The health outcomes are categorized as:

- Higher incidence and/or prevalence of disease, including earlier onset or more aggressive progression of disease.
- Premature or excessive mortality from specific health conditions.
- Greater global burden of disease, such as Disability Adjusted Life Years (DALY), as measured by population health metrics.
- Poorer health behaviors and clinical outcomes related to the aforementioned.
- Worse outcomes on validated self-reported measures that reflect daily functioning or symptoms from specific conditions.

The Collection Main Page

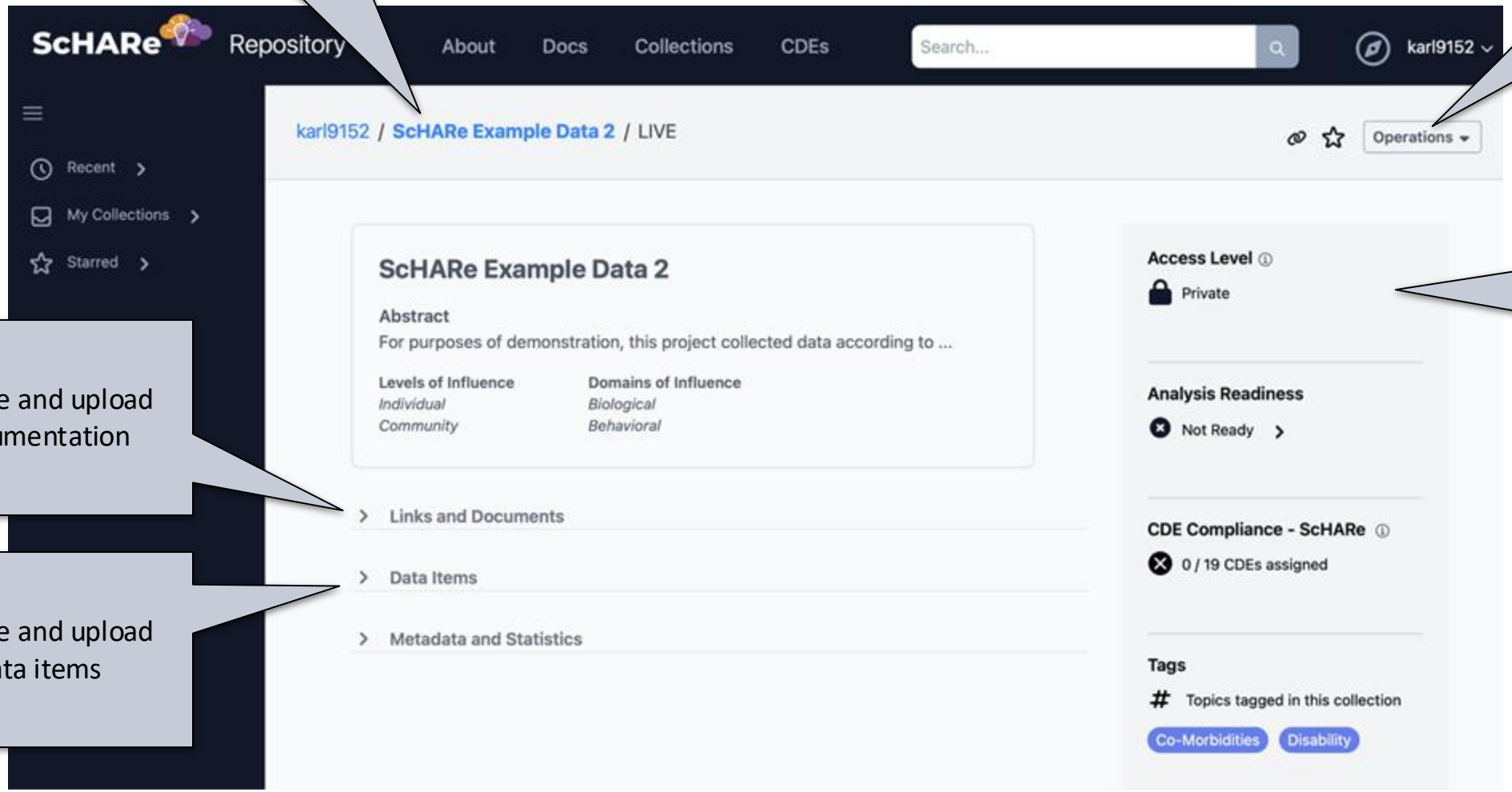
Click the collection name to return to the main page

Actions on the collection itself (edit metadata, share, delete)

Sidebar with key information

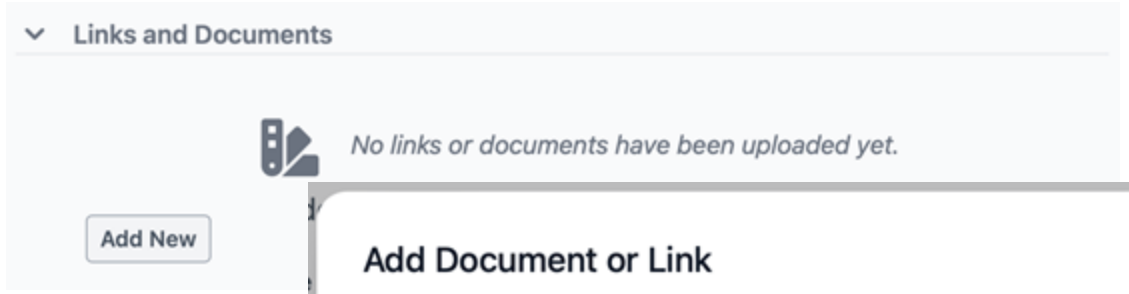
Browse and upload documentation

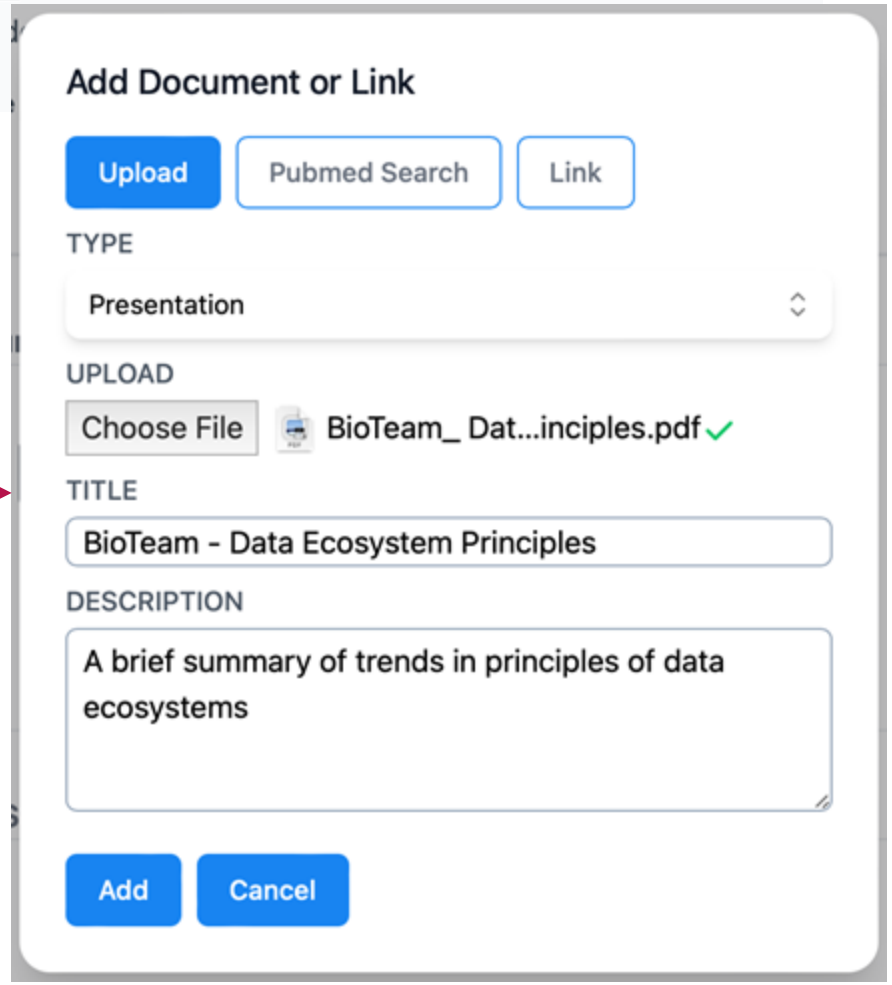
Browse and upload data items

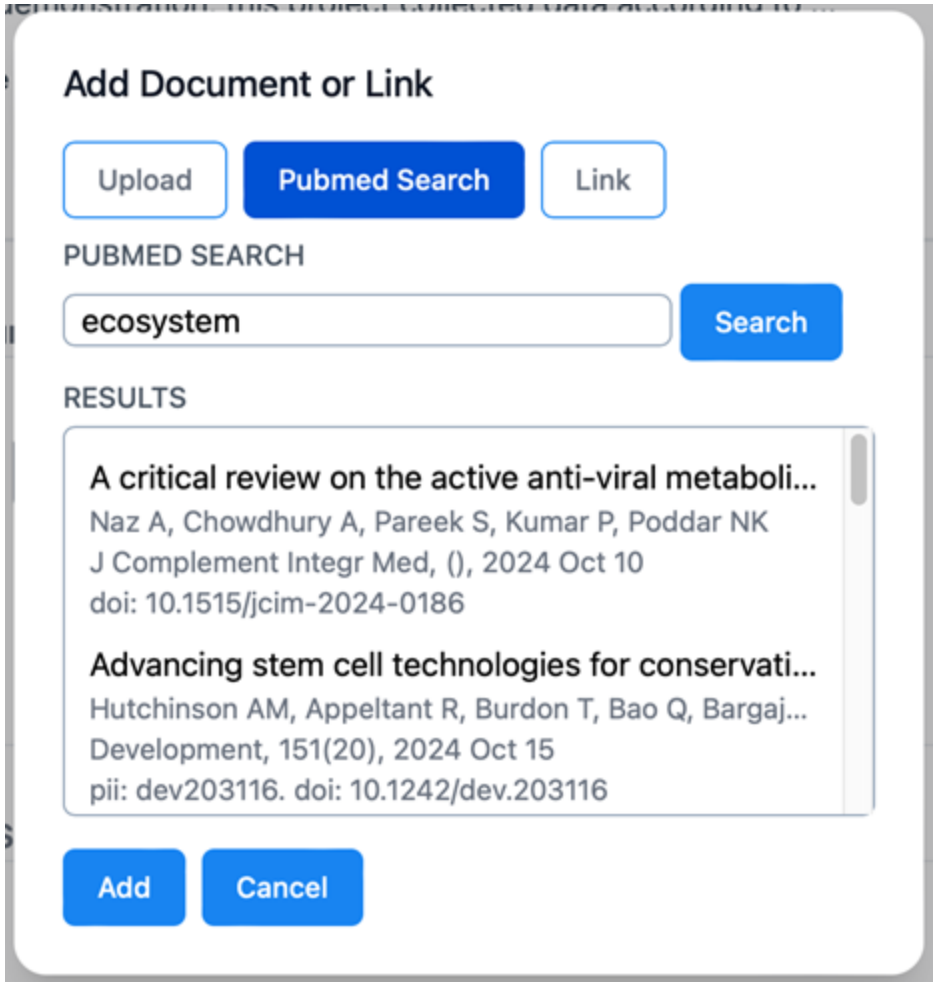


Adding Links and Documents

- Upload any document, presentation, etc.
- Link to papers through Pubmed







Uploading a Data File

You can upload to the ScHARe Repository…

Any file

- Documents (PDF, docx, etc.)
- Archives (zip, etc.)

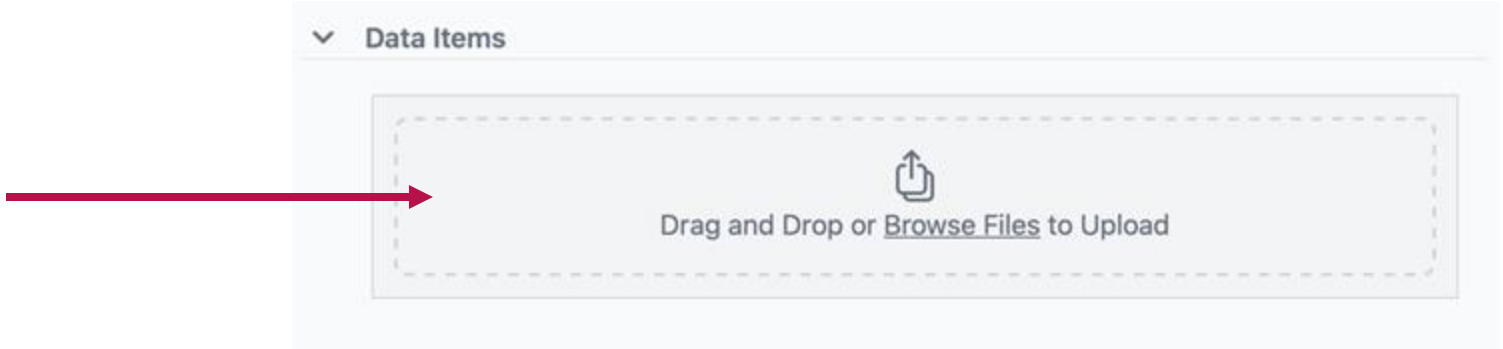
Some file formats can be previewed in the web interface

Anything uploaded can be downloaded

Tabular data

- CSV and TSV
- Excel (xlsx, xls)
- SAS (XPORT V5/6, not V8)
- SPSS
- STATA
- Parquet

Tabular data can be viewed in the web interface, plus much more…

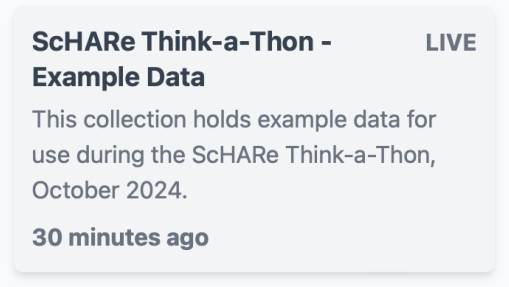


Uploading a Data File

For today's training, we'll all use the same example data file.

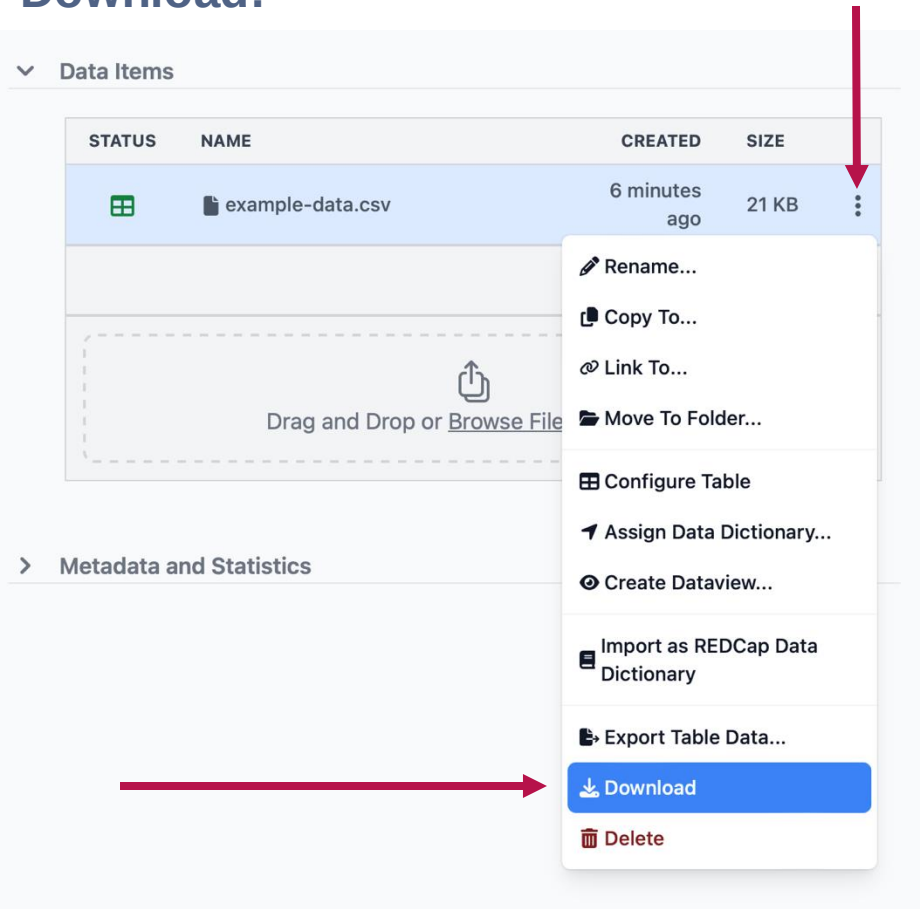
Visit:
https://test-schare.nimhd.nih.gov/c/fwf4bfnf7b

Or click "Collections" and find
"ScHARe Think-a-Thon – Example Data"

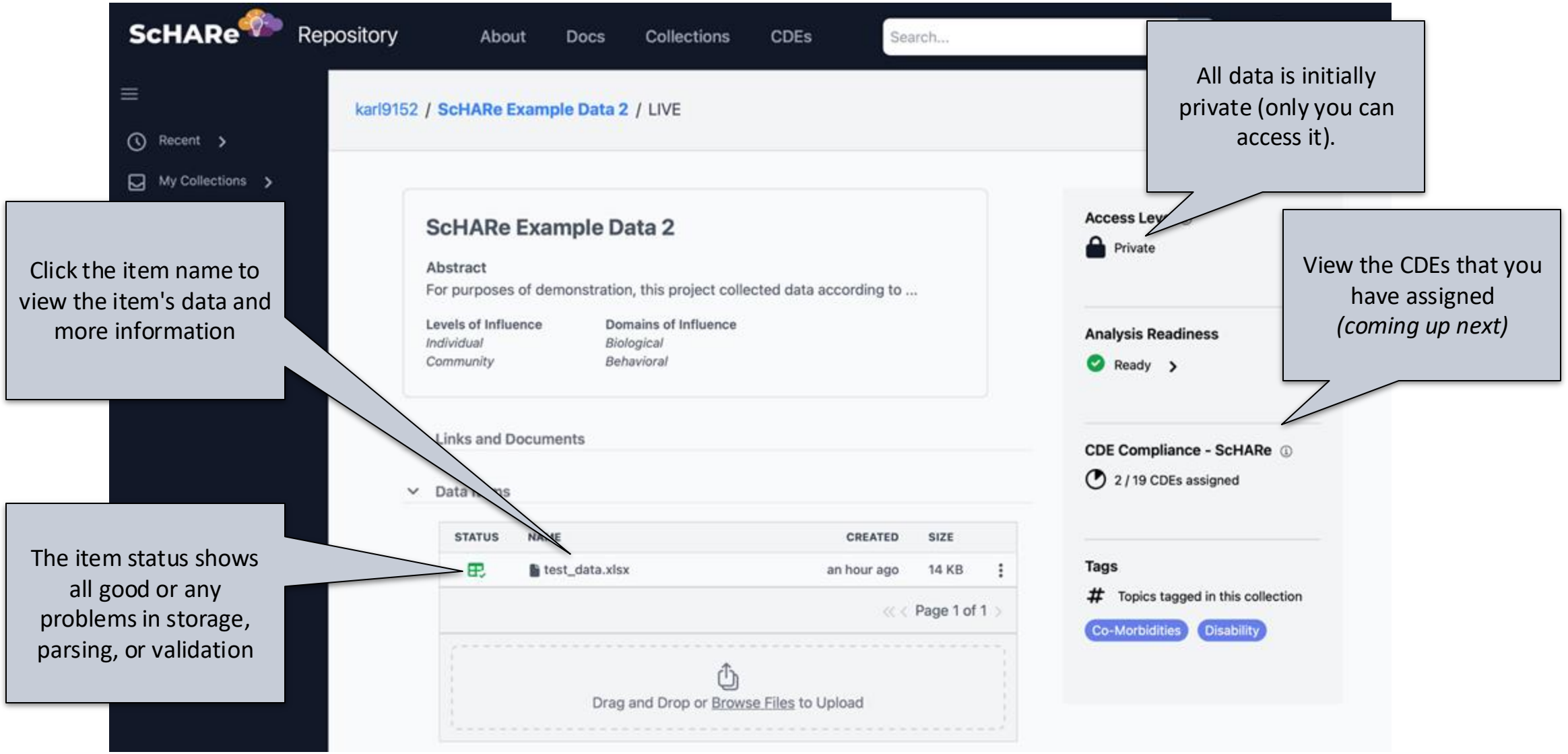


… Expand the Data Items section

Download:



Viewing Uploaded Data



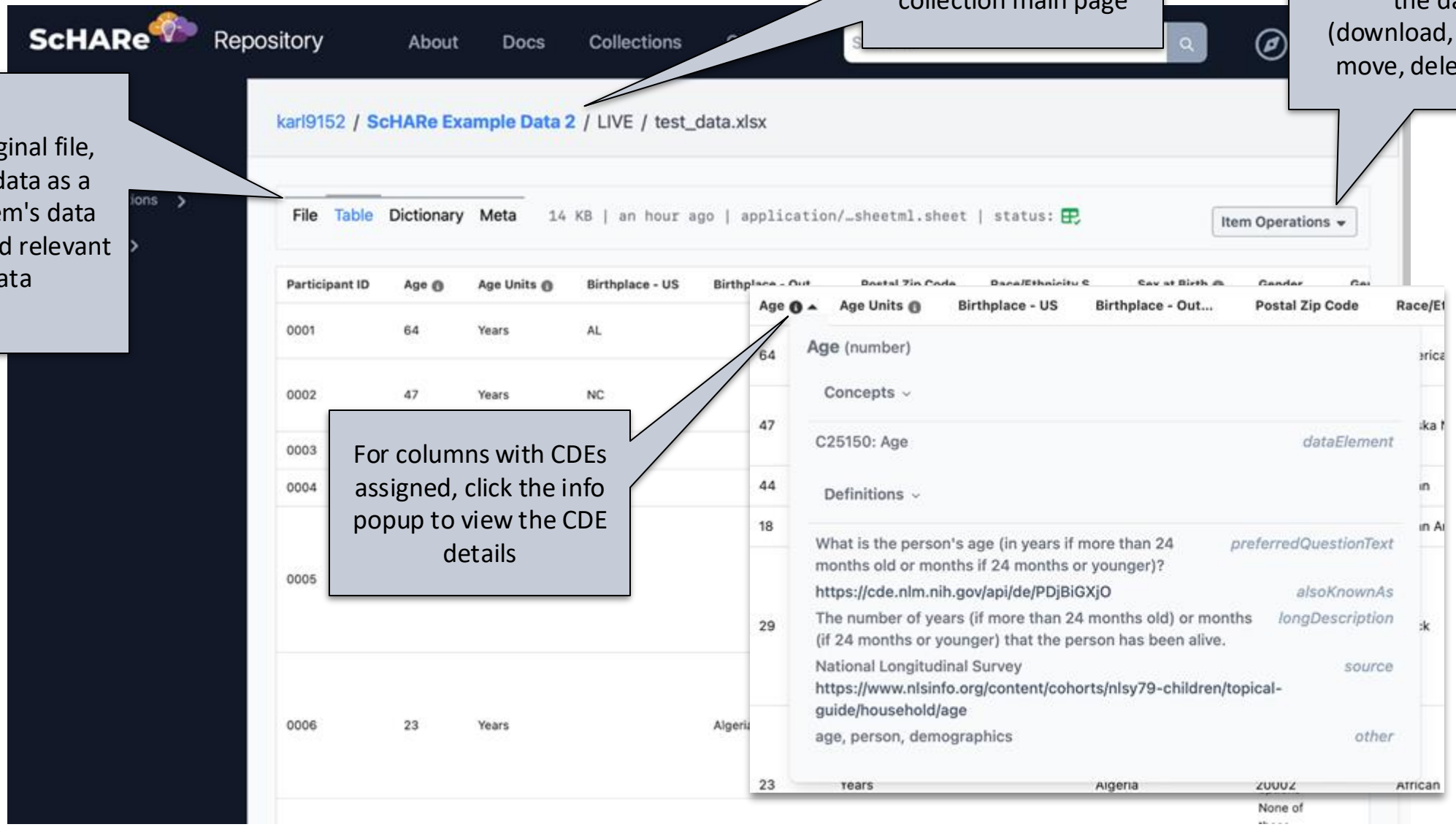
Viewing Uploaded Data

Click the collection name to return to the collection main page

Perform operations on the data (download, rename, move, delete, etc.)

View the original file, the parsed data as a table, the item's data dictionary, and relevant metadata

For columns with CDEs assigned, click the info popup to view the CDE details



ScHARe Repository
Introduction

October 16, 2024

Deborah Duran, PhD · NIMHD
Luca Calzoni, MD MS PhD Cand. · NIMHD
Elif Dede Yildirim, PhD · NIMHD

ScHARe

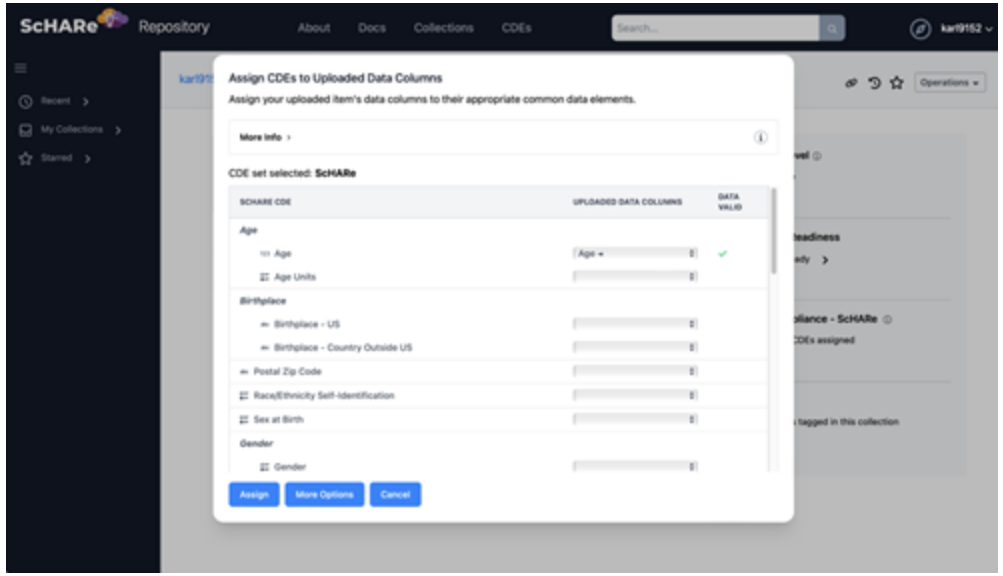
CDE Mapping

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

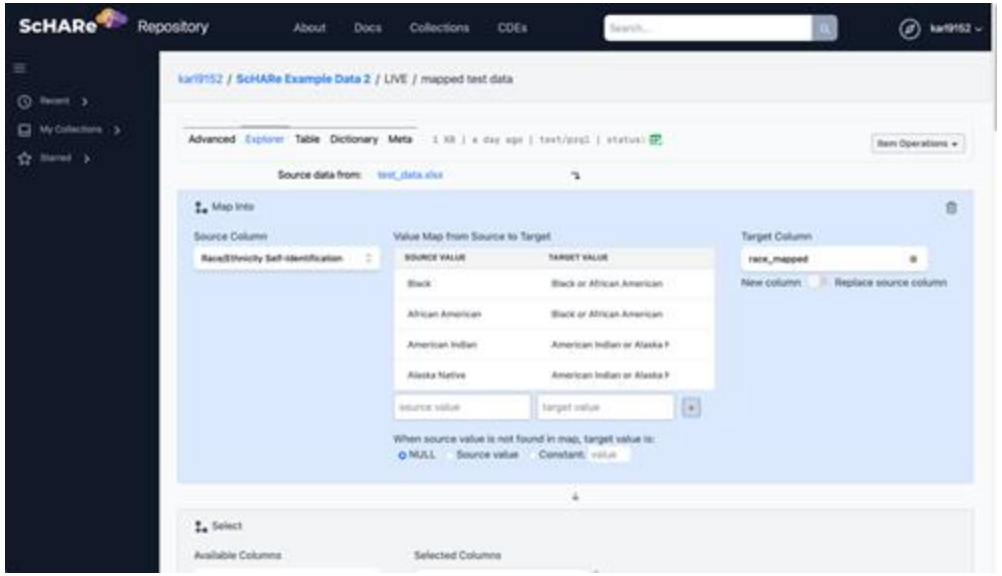
Getting Started with Mapping CDEs

- The Repository does not *enforce* CDEs, but it does *encourage* using CDEs
- Start with the data you have, and use the tools the repository offers to map as much of your data to CDEs as possible

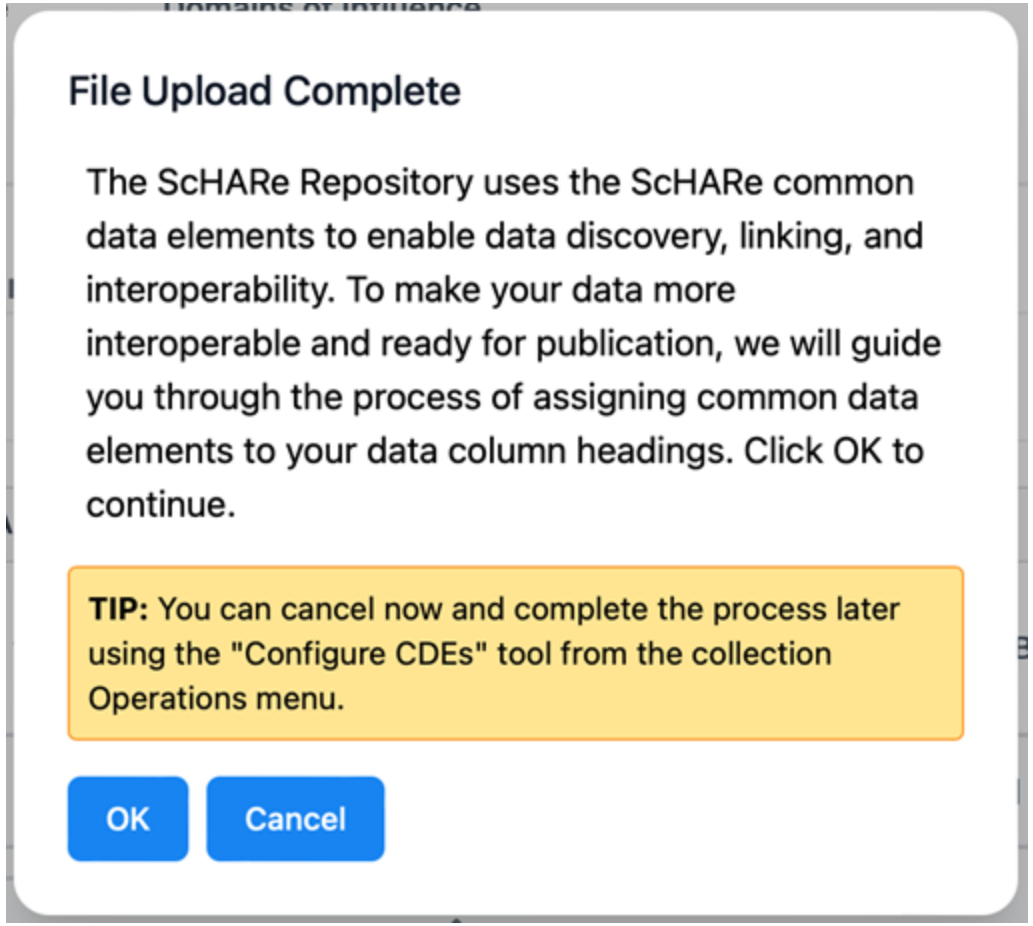
Assign CDEs at Upload



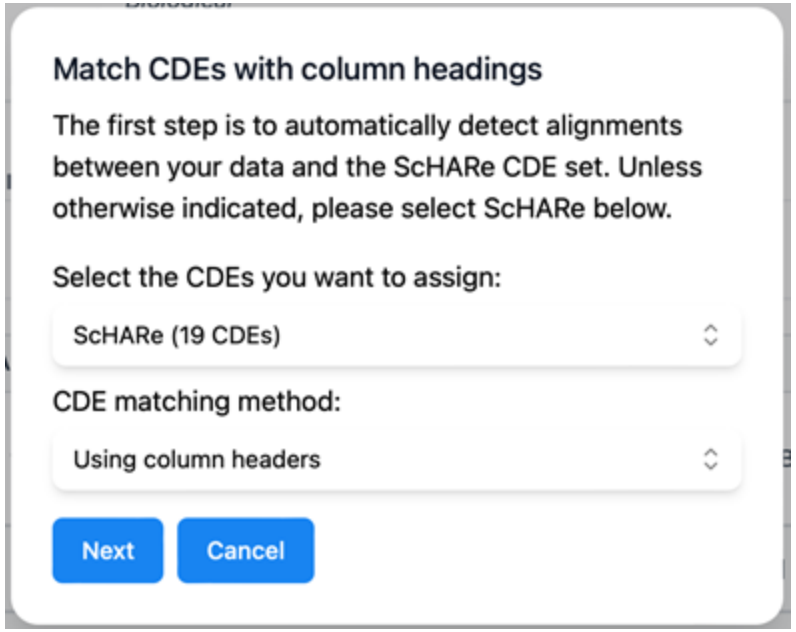
Map CDEs using Dataviews



Assigning CDEs at Upload



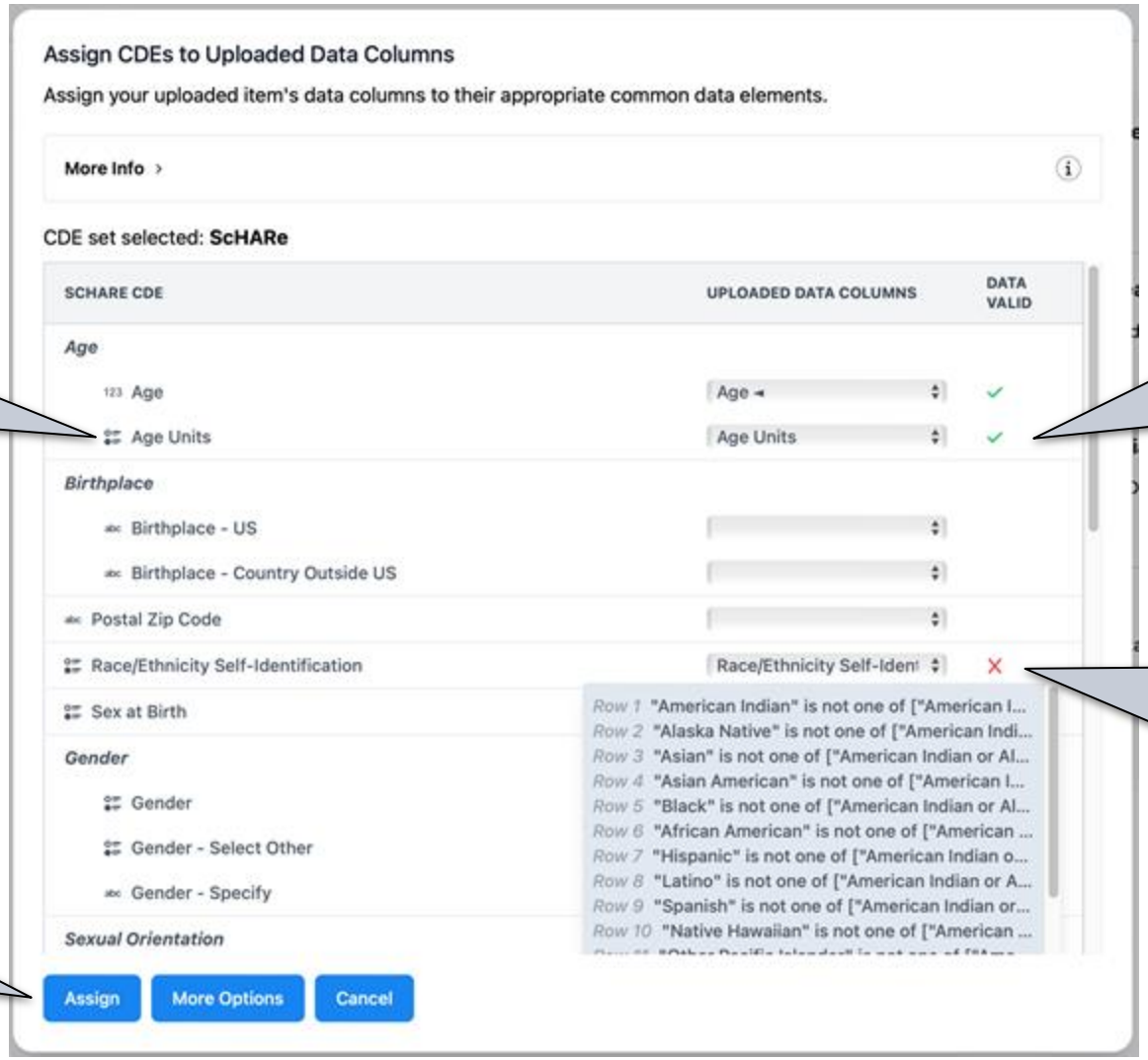
- After you upload a tabular data file, the system will recommend that you begin the process of assigning CDEs to your data column headings.



Assigning CDEs at Upload

Some of the assignments will be correct -- double-check the validation and the column name matches what you expect

Complete or cancel the process, or use "More Options" to go to a detailed tool for CDE assignment



Some of the assignments will be wrong or missing -- add anything that should be added, and double-check the validation

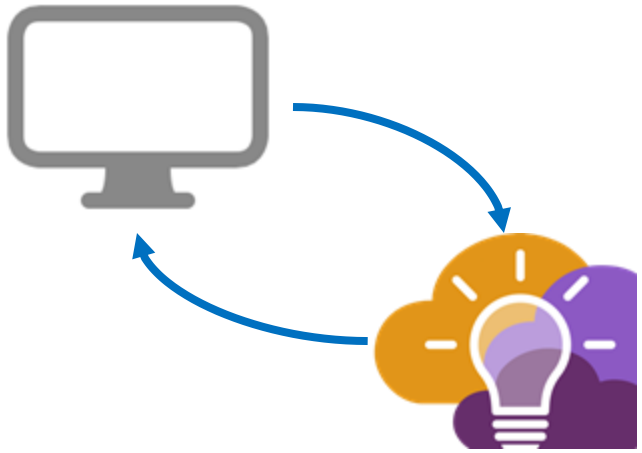
When an assigned CDE fails validation, hover over the red X to get details on why

NOTE: CDE validation errors can be corrected using dataviews (coming up later)

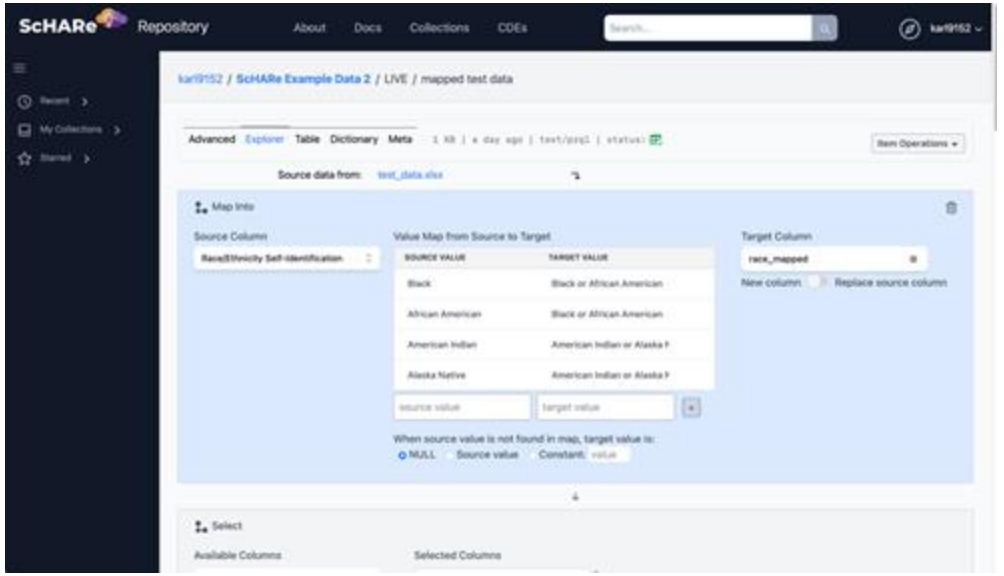
Using Dataviews to Map CDEs

If your data doesn't conform to the CDEs as it was uploaded, you have two options:

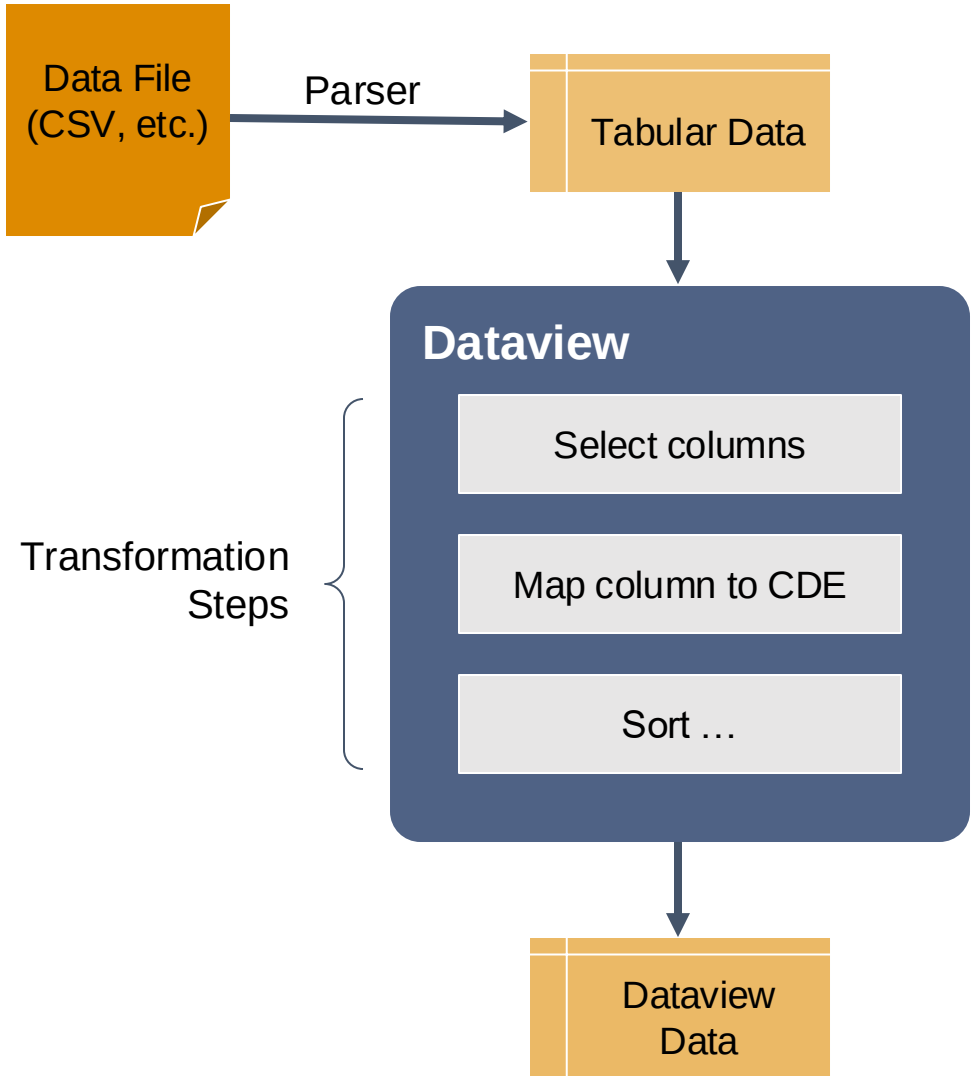
Use your own tools to adjust the data and re-upload



Use a Dataview to Map to CDEs within the Repository



What is a Dataview?

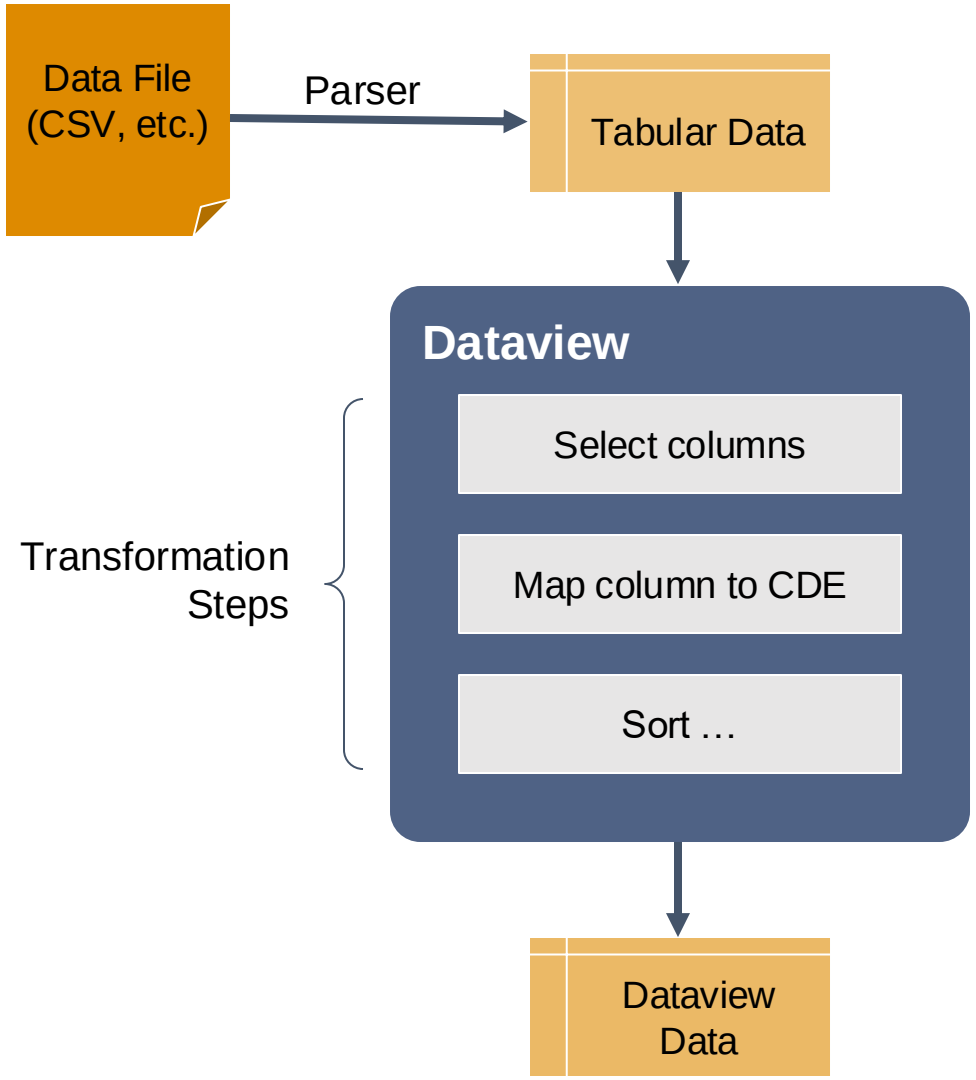


Dataviews take data from one or more sources, apply a series of transformation steps to that data *(filtering, sorting, mapping, etc.)* resulting in a new table of data as output.

Uses of Dataviews:

- Creating subsets of data
- Hiding PHI/PII for publishing
- Summarizing individual-level data into subsets and estimates
- Joining multiple datasets together
- **Mapping to CDEs**
- … many others!

Mapping CDEs via Dataviews

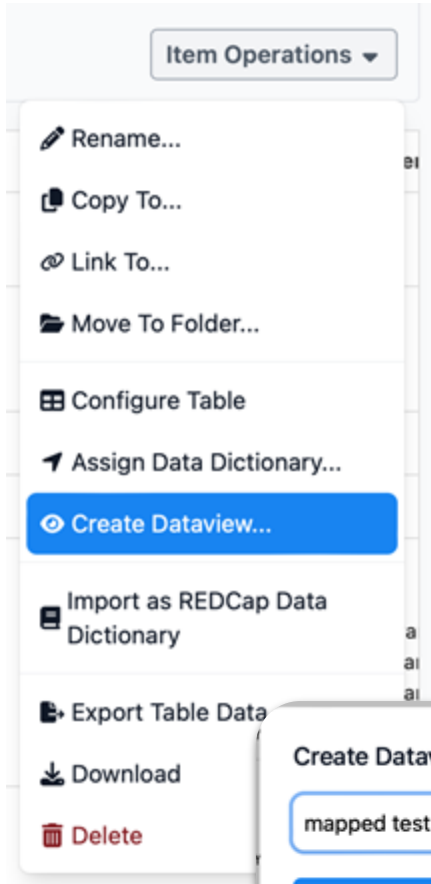


Next, we'll walk through the process together:

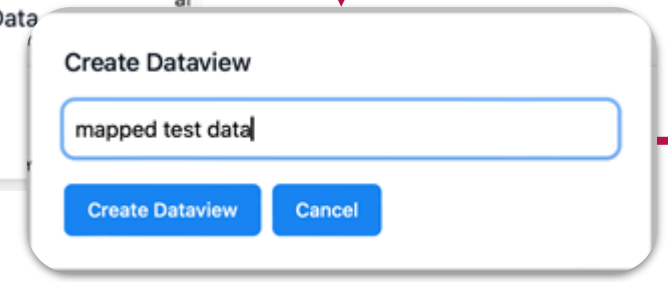
1. Create a new dataview item
2. Add mapping steps to the dataview that map your data values to align with the CDEs
3. Label the newly mapped data columns with the appropriate CDEs and check that it passes validation

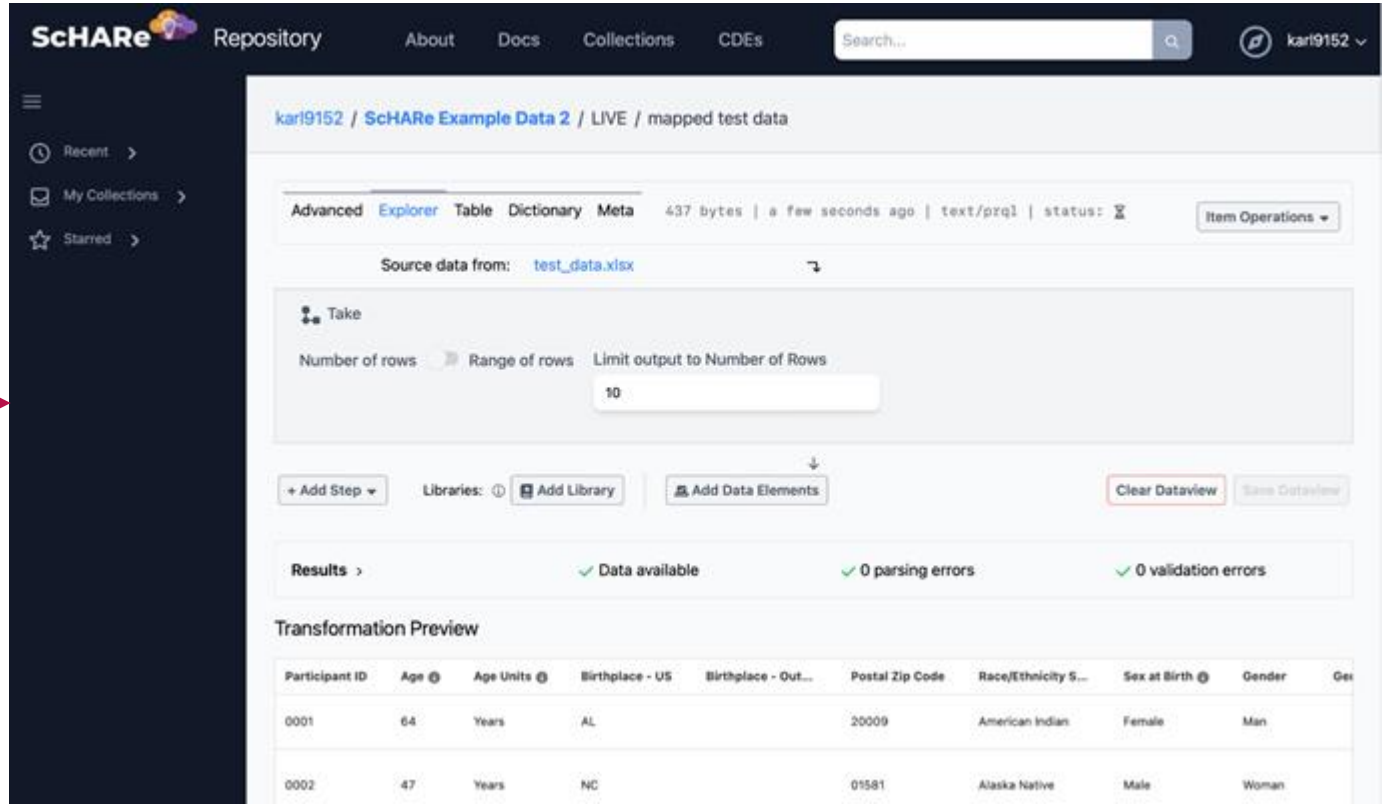
Mapping CDEs via Dataviews

Step 1: Create a new Dataview



1. Select **Create Dataview…** from the item's menu
2. Give it a name
3. It will open the Explorer, where you can begin building





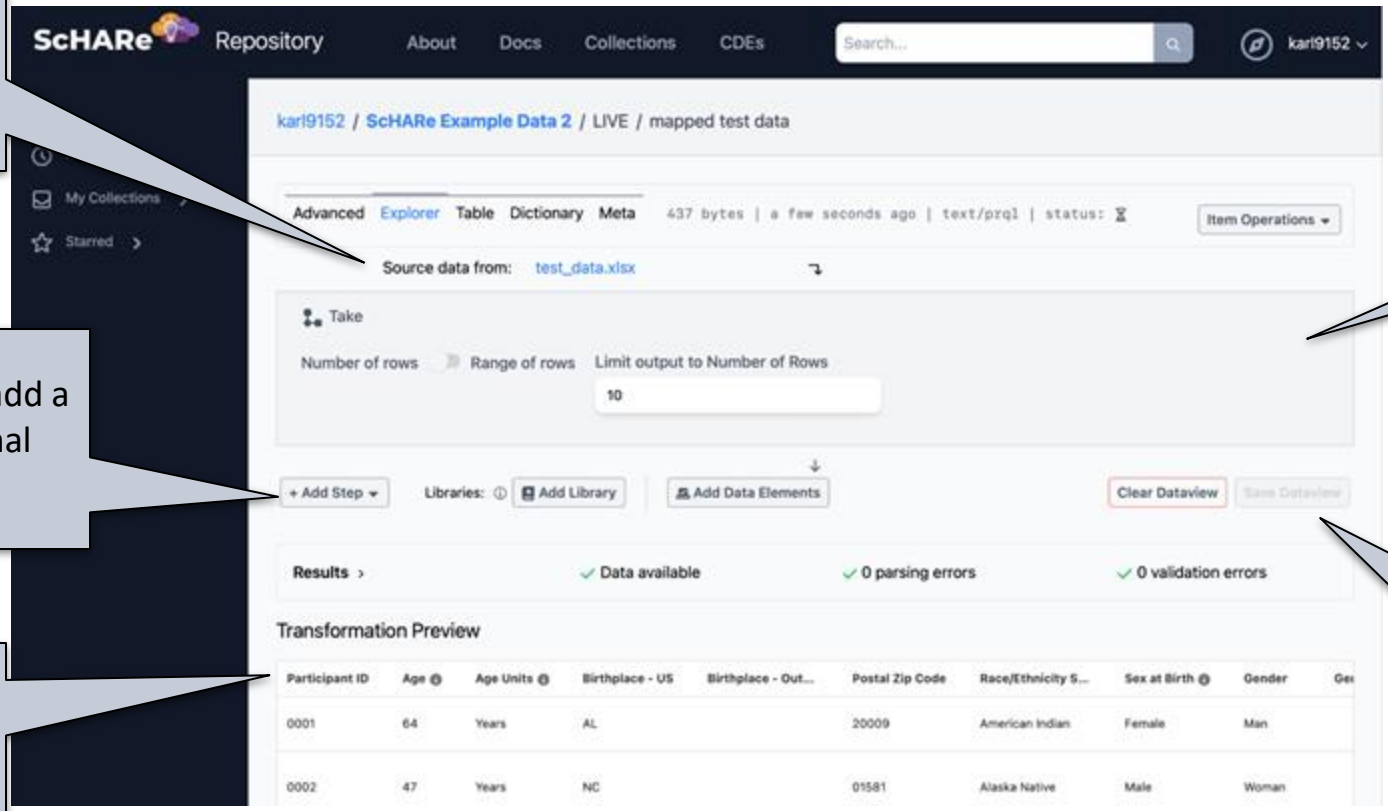
Mapping CDEs via Dataviews

A Quick Tour of the Dataview Explorer screen

Source data comes in at the top

Add more steps or add a library of additional steps

The result is previewed at the bottom



Data flows vertically down the pipeline of transformation steps

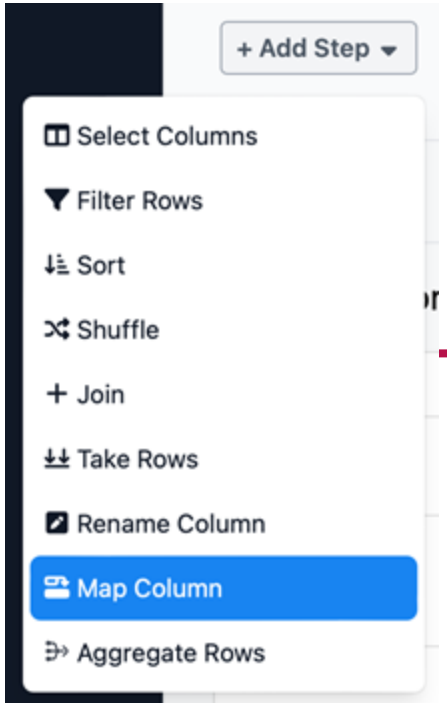
Save your changes using the Save Dataview button

Once you're done, use the **Table** tab to view the full results

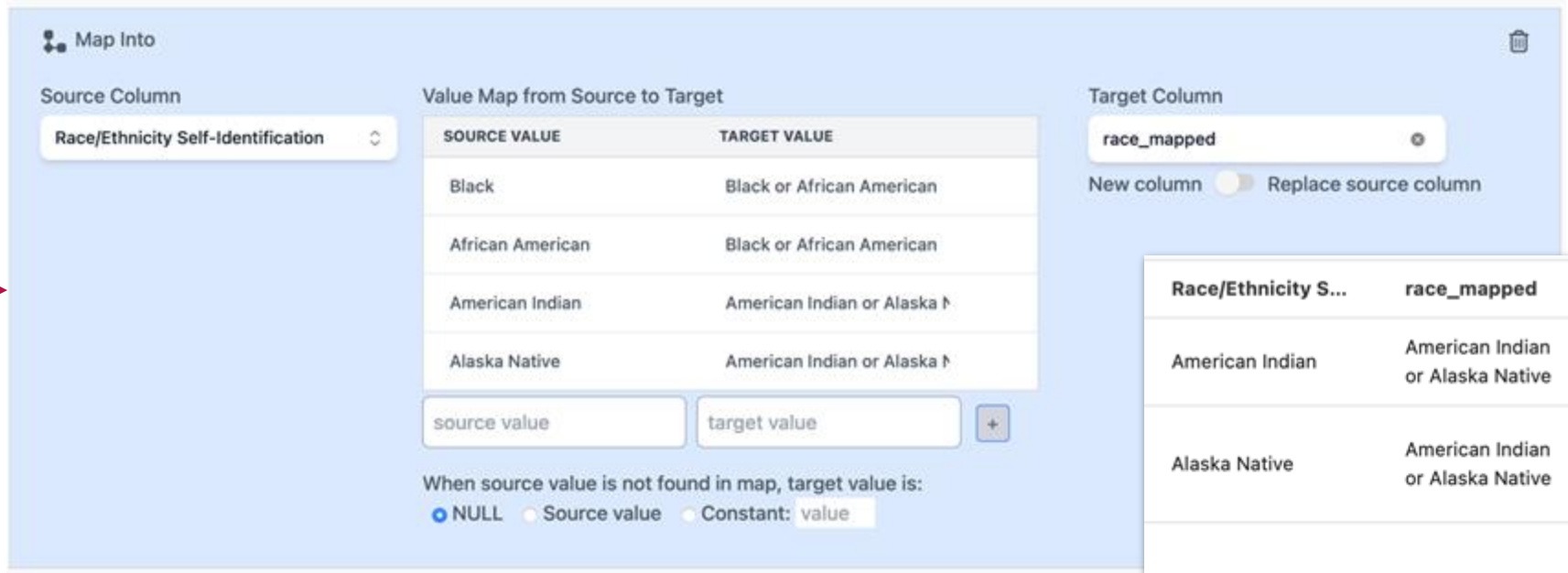
Mapping CDEs via Dataviews

Step 2: Add Mapping Steps to your Dataview

1. Using the **Add Step** menu, add a **Map Column** step

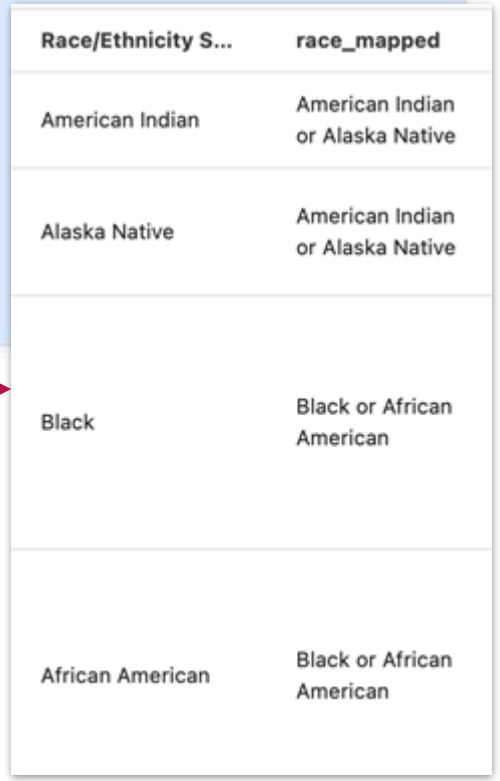


2. Fill in the step fields (source column, target column, value map)



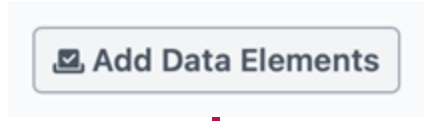
3. Review the results, adjust as needed

4. Repeat for all columns that need to be mapped

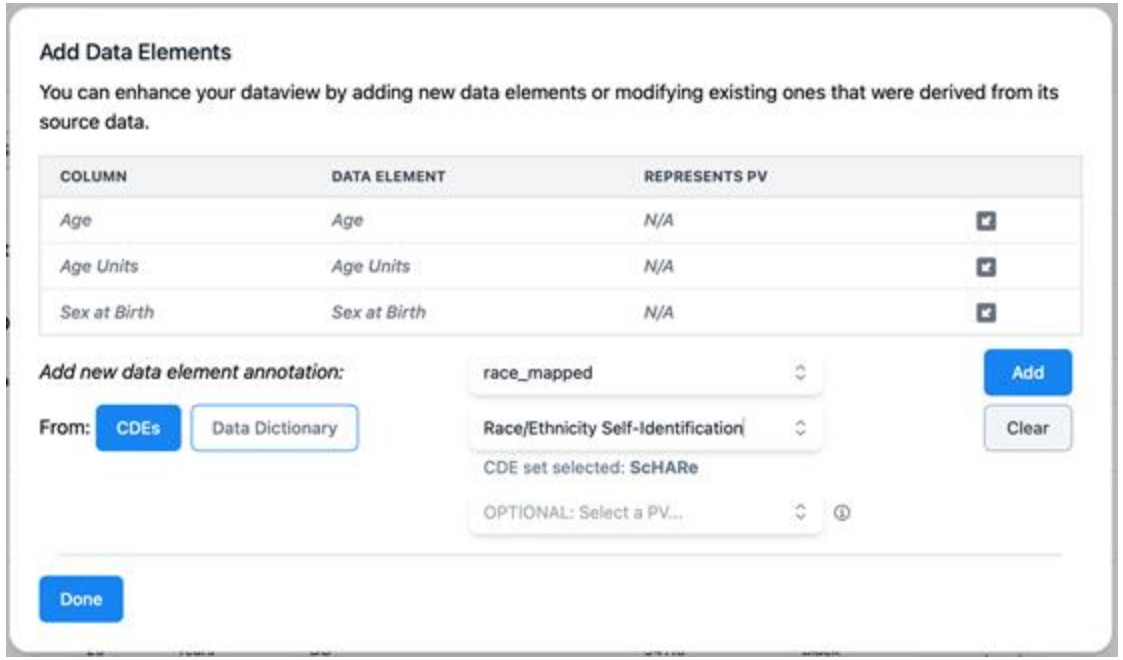


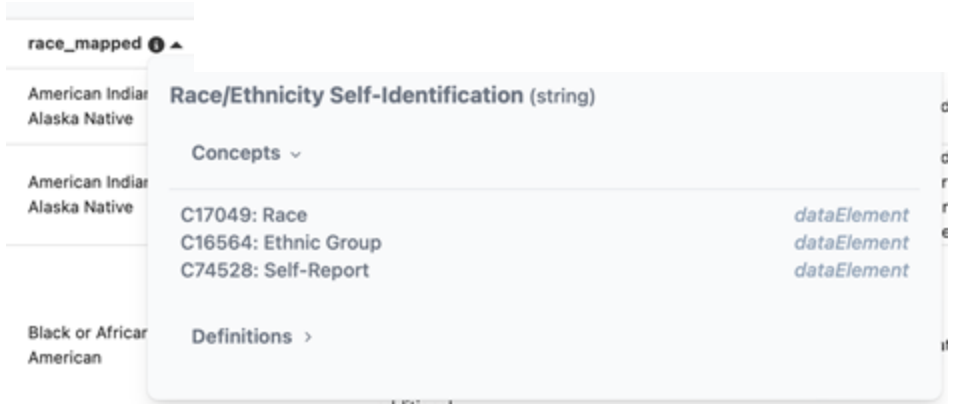
Mapping CDEs via Dataviews

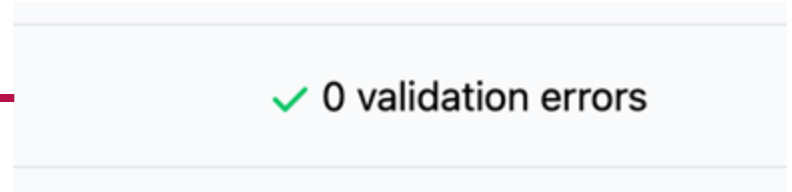
Step 3: Assign CDEs to Data Columns



1. Click **Add Data Elements**
2. Select the column and data element, click Add
3. Verify the data element is on the column and the data validates successfully



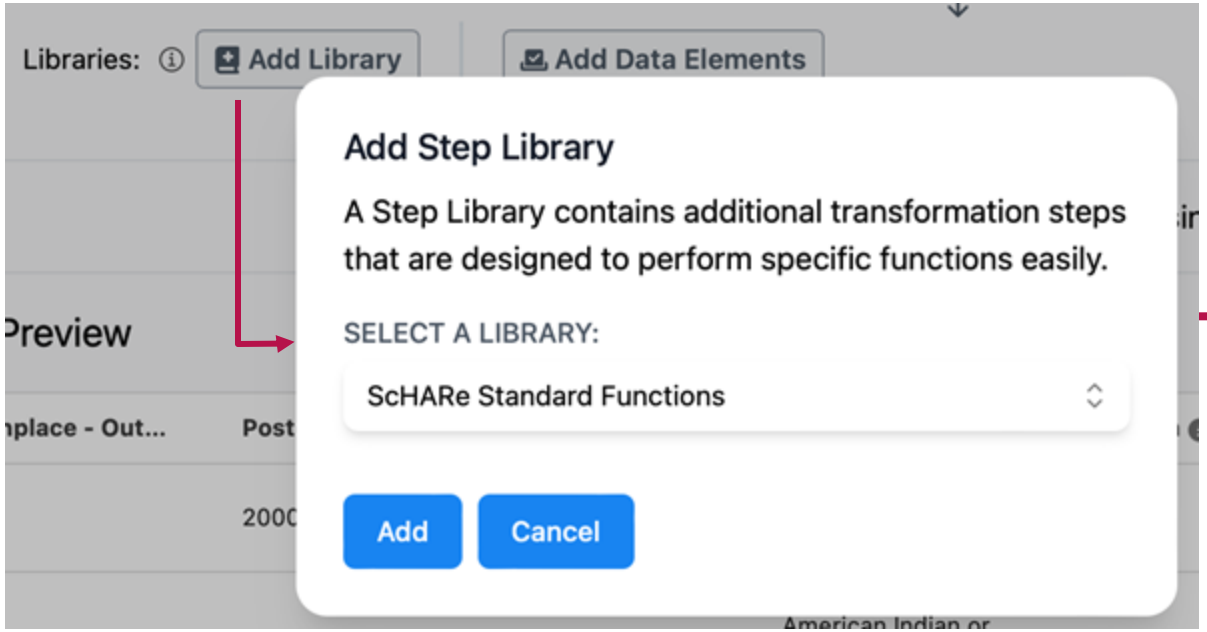


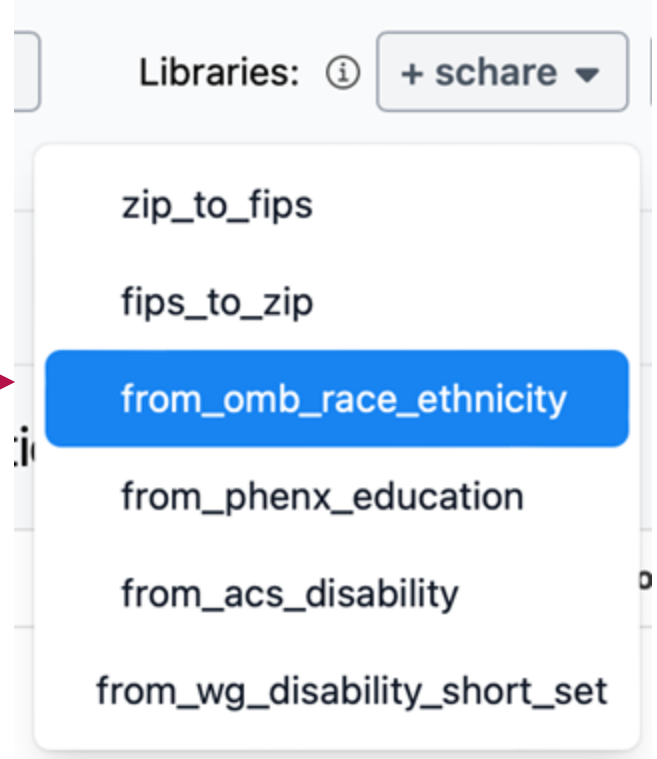


Mapping CDEs via Dataviews

The system also has a number of specific mapping functions available, for cases when the mapping is standardized or when the mapping is more complex than can be handled by the generic mapping functions.

Please contact our support team if you need any help with custom mapping functions!





You're done uploading your first data set!

What we've accomplished:

- Created a private Collection to hold all of your project data in one place
- Uploaded a data dictionary, documentation, linked to papers, etc.
- Uploaded your tabular data and viewed it
- Assigned and mapped your data to CDEs

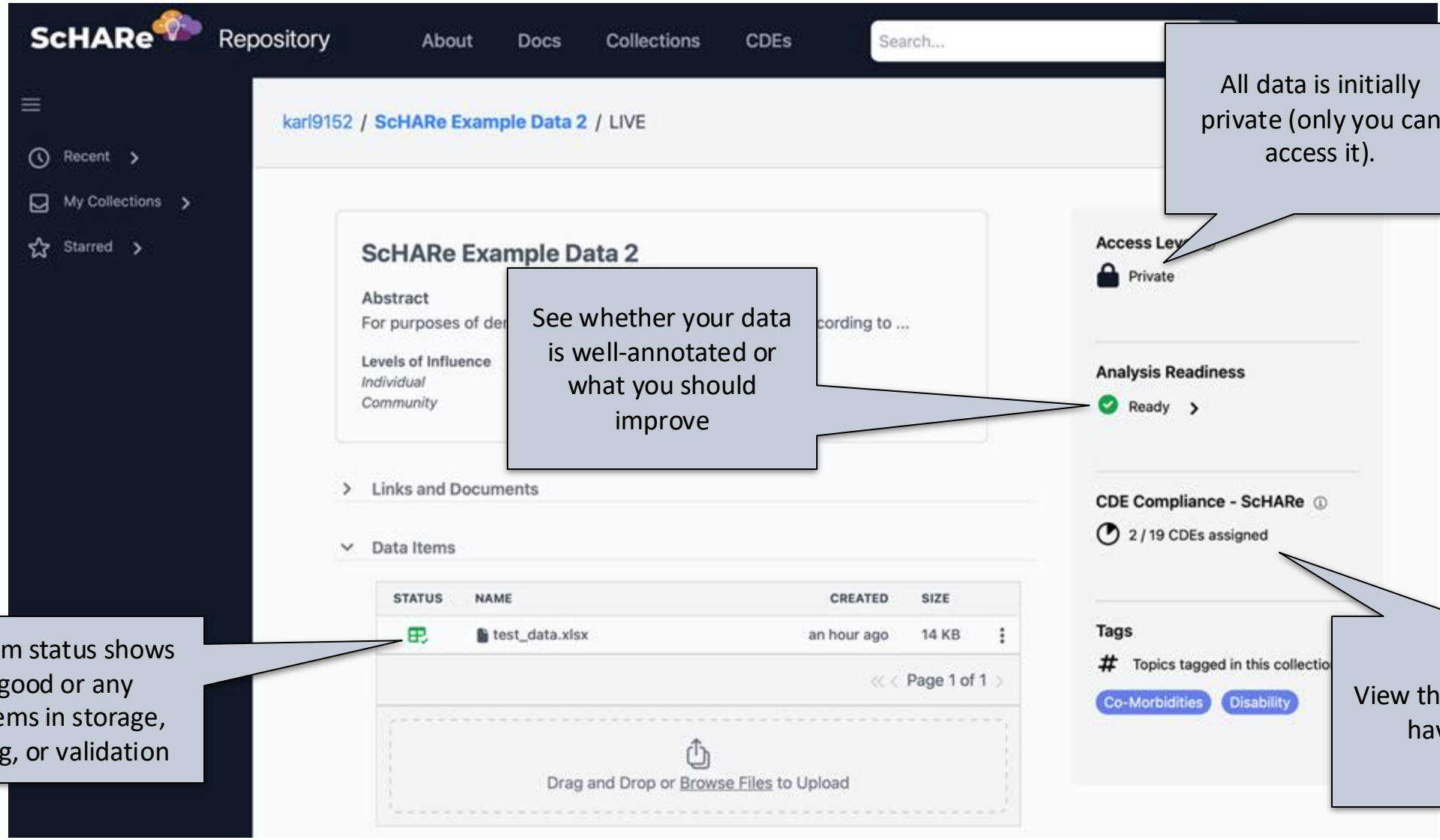
Questions so far?

ScHARe

Sharing Data

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

Viewing your Collection



All data is initially private (only you can access it).

See whether your data is well-annotated or what you should improve

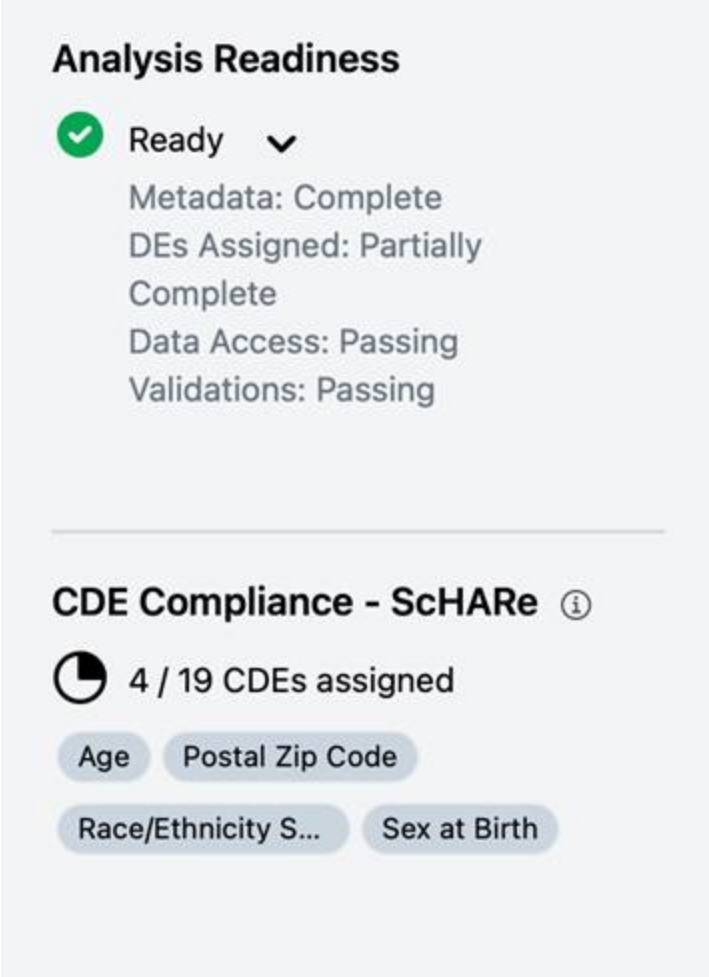
The item status shows all good or any problems in storage, parsing, or validation

View the CDEs that you have assigned

CDE Compliance and Analysis Readiness

Analysis Readiness - a simple metric on whether your data is ready for downstream use

1. Did you assign **metadata** to your collection (tags, project-level CDEs)?
2. Have you partially or fully **assigned CDEs** to your data?
3. Is the data **accessible** to the system (no broken links)?
4. Does the data pass **validation** according to the assigned CDEs?

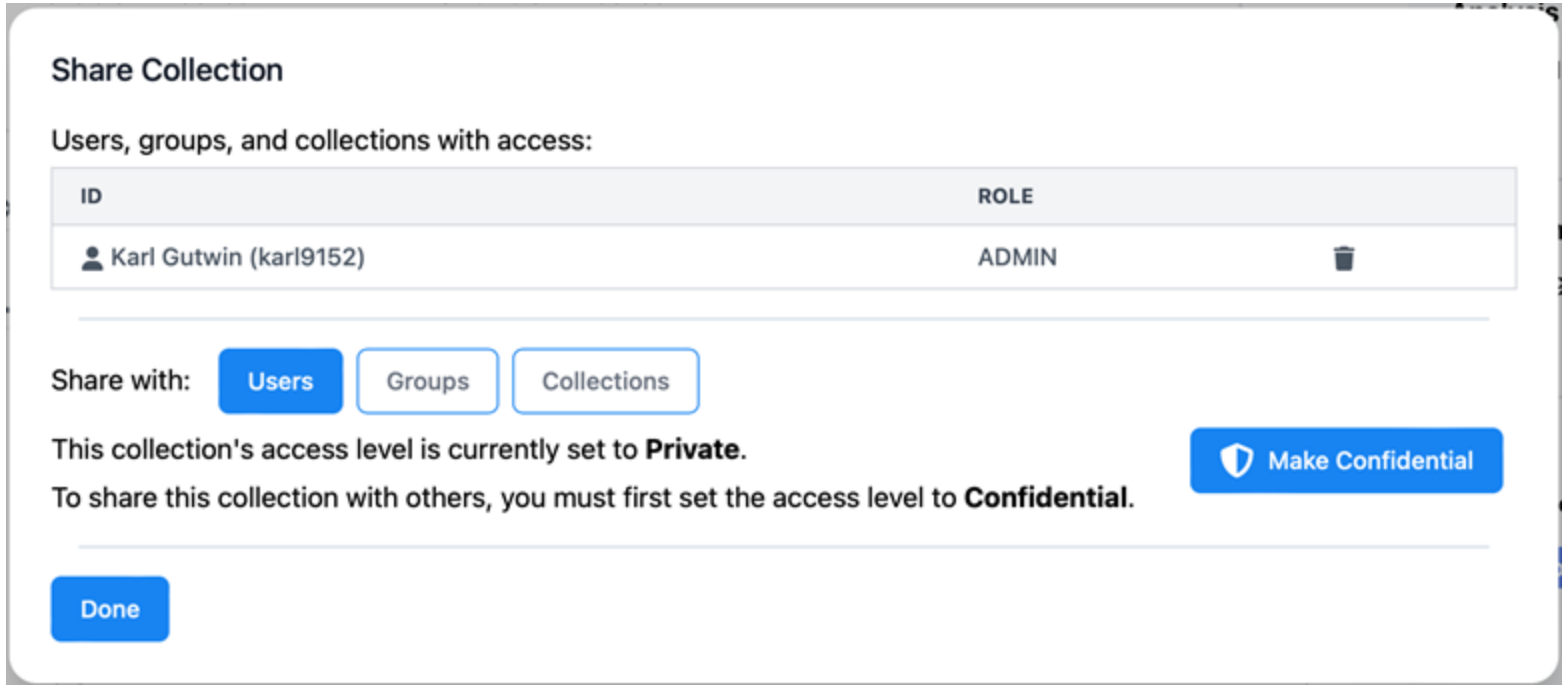


CDE Compliance - Showing how many CDEs have been assigned across the data in the collection

- Click on the metric to pop up the list of CDEs assigned
- Click on an individual CDE to find more information about that CDE

Access Levels and Sharing Data

You have control over how your data is shared on the ScHARe Repository. By default, all collections start out as **Private**.



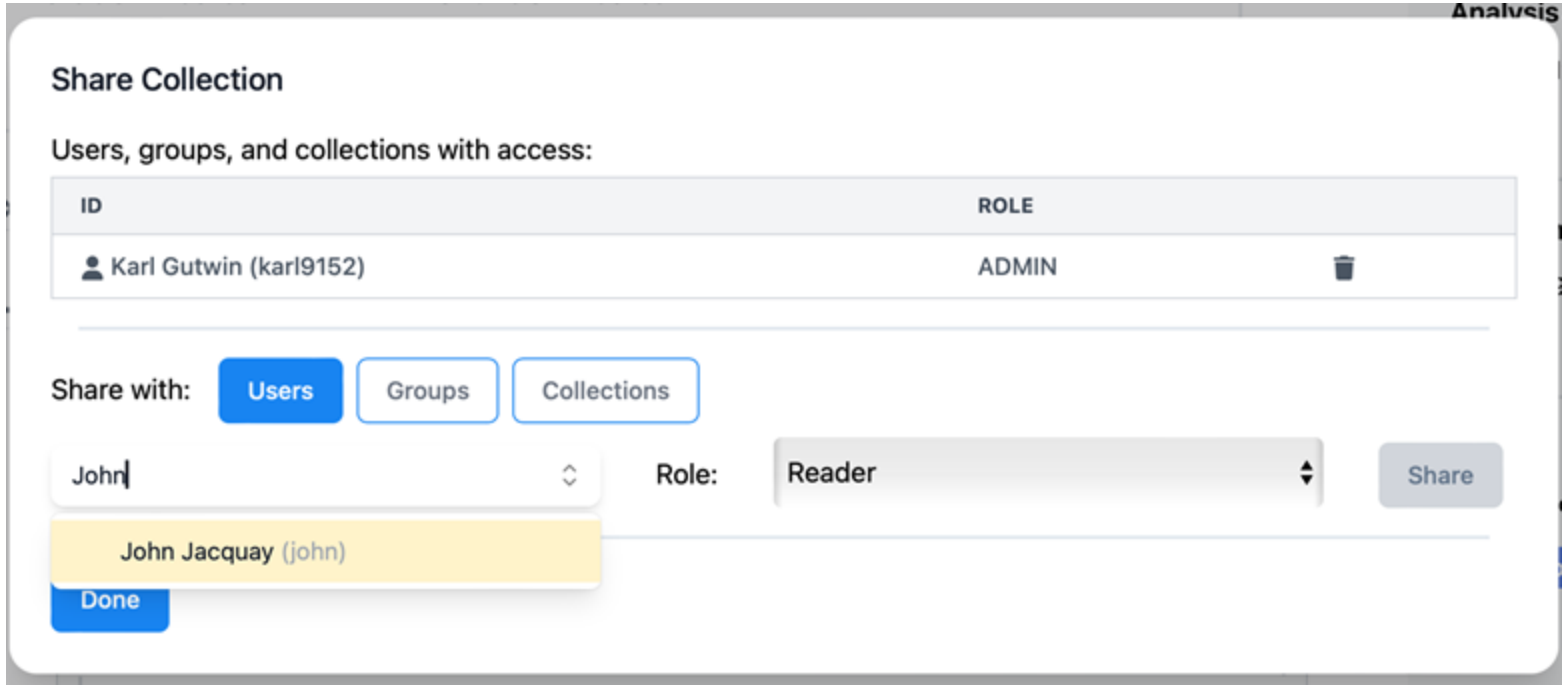
Access Levels

The access level of a collection defines the maximum permissions that can be used to share it with others. The following access levels are supported:

- **Private**: Only the collection's owner can access
- **Confidential**: The collection can be shared with named users
- **Controlled**: The collection can be shared with members of a controlled access group, as well as named users
- **Public**: The collection can be read by any user, including those not logged in; it can also be shared with named users

Access Levels and Sharing Data

After changing your collection's access level to Confidential, you can share it with other users, groups, and collections.



Access Levels

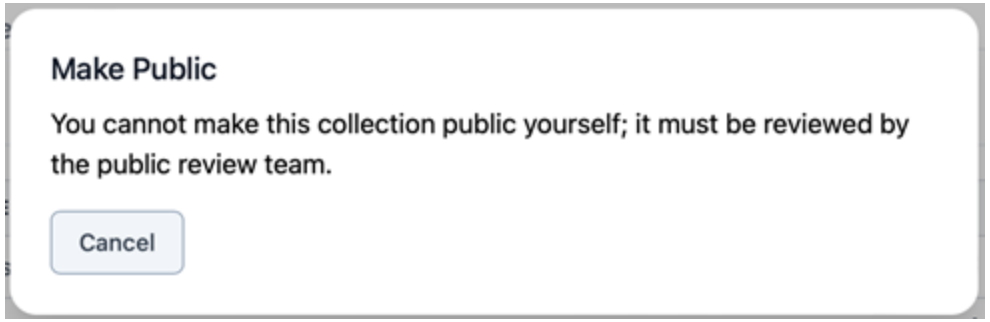
The access level of a collection defines the maximum permissions that can be used to share it with others. The following access levels are supported:

- **Private**: Only the collection's owner can access
- **Confidential**: The collection can be shared with named users
- **Controlled**: The collection can be shared with members of a controlled access group, as well as named users
- **Public**: The collection can be read by any user, including those not logged in; it can also be shared with named users

Publishing Data

We are still working on the final version of the process; however, it will essentially be:

1. Prepare your data according to the published guidelines
2. Add the Public Review group as Admin on your collection
    a. This is necessary for them to be able to alter your collection's Access Level on your behalf
3. Notify the Public Review group by email
    a. They will review your data for compliance with the data sharing guidelines
    b. If any changes are necessary, you can work with the reviewer
    c. Once complete, they will create a static version of your collection and set its Access Level to Public



ScHARe

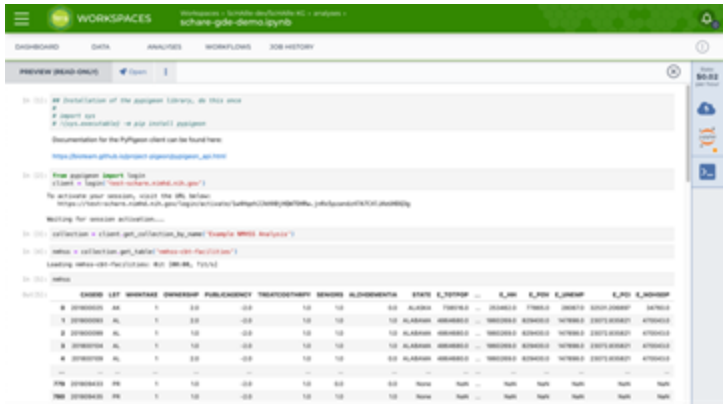
Data Aggregation and Analysis

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

How do I analyze or aggregate data from the ScHARe Repository?

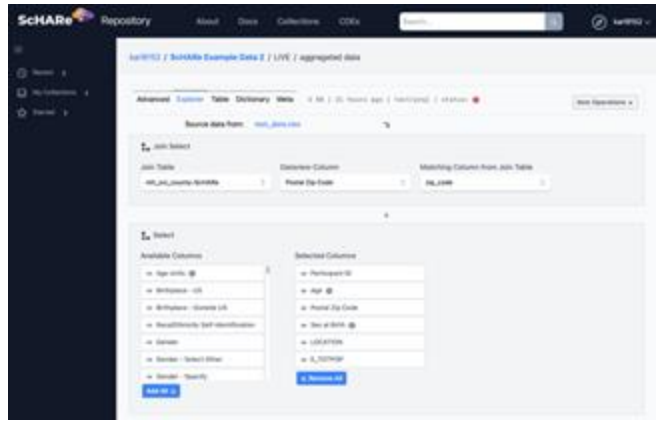
Three quick options:

Connect to a Terra Jupyter notebook



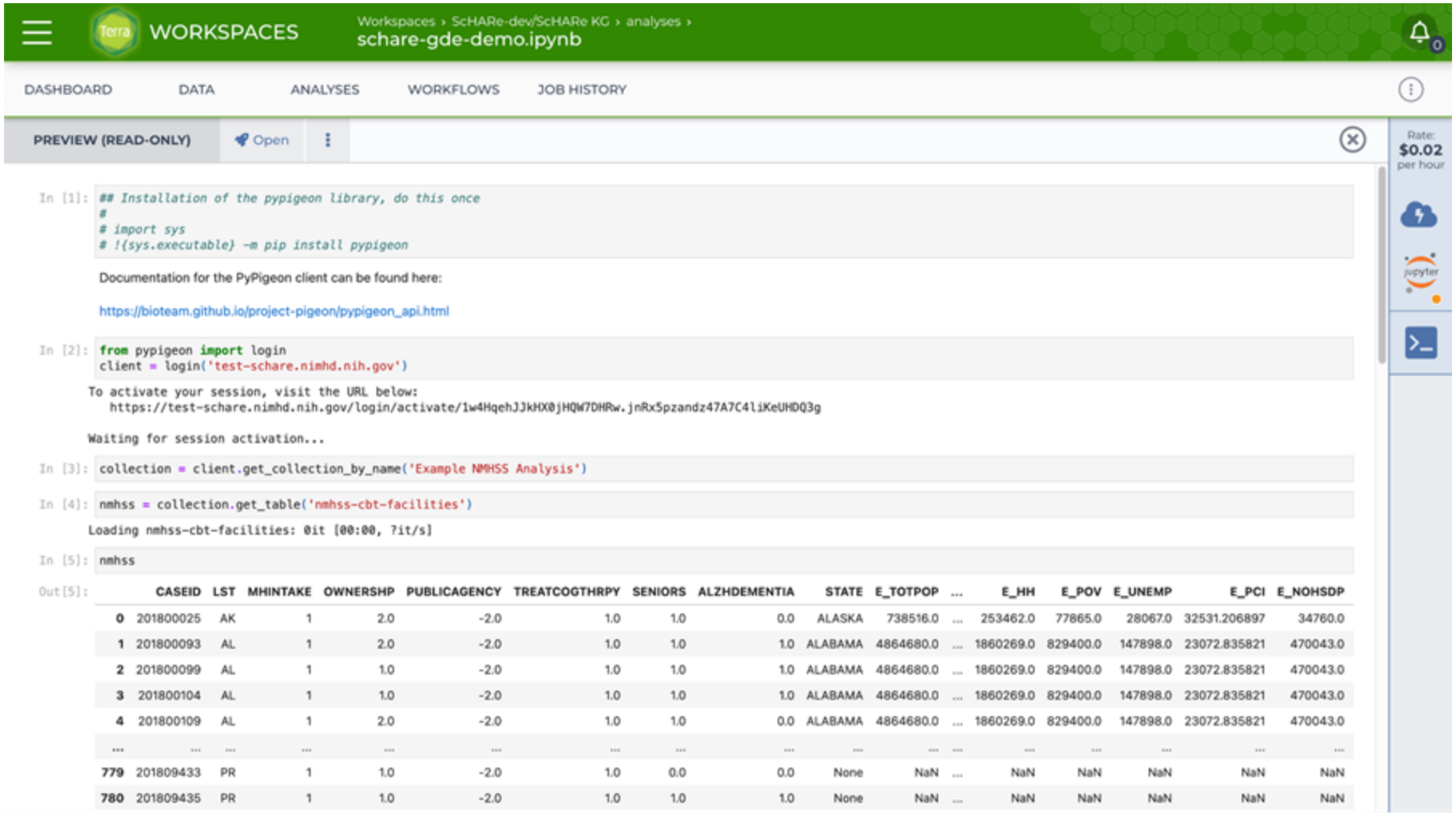
Download to your own computer

Use Dataviews on the Repository



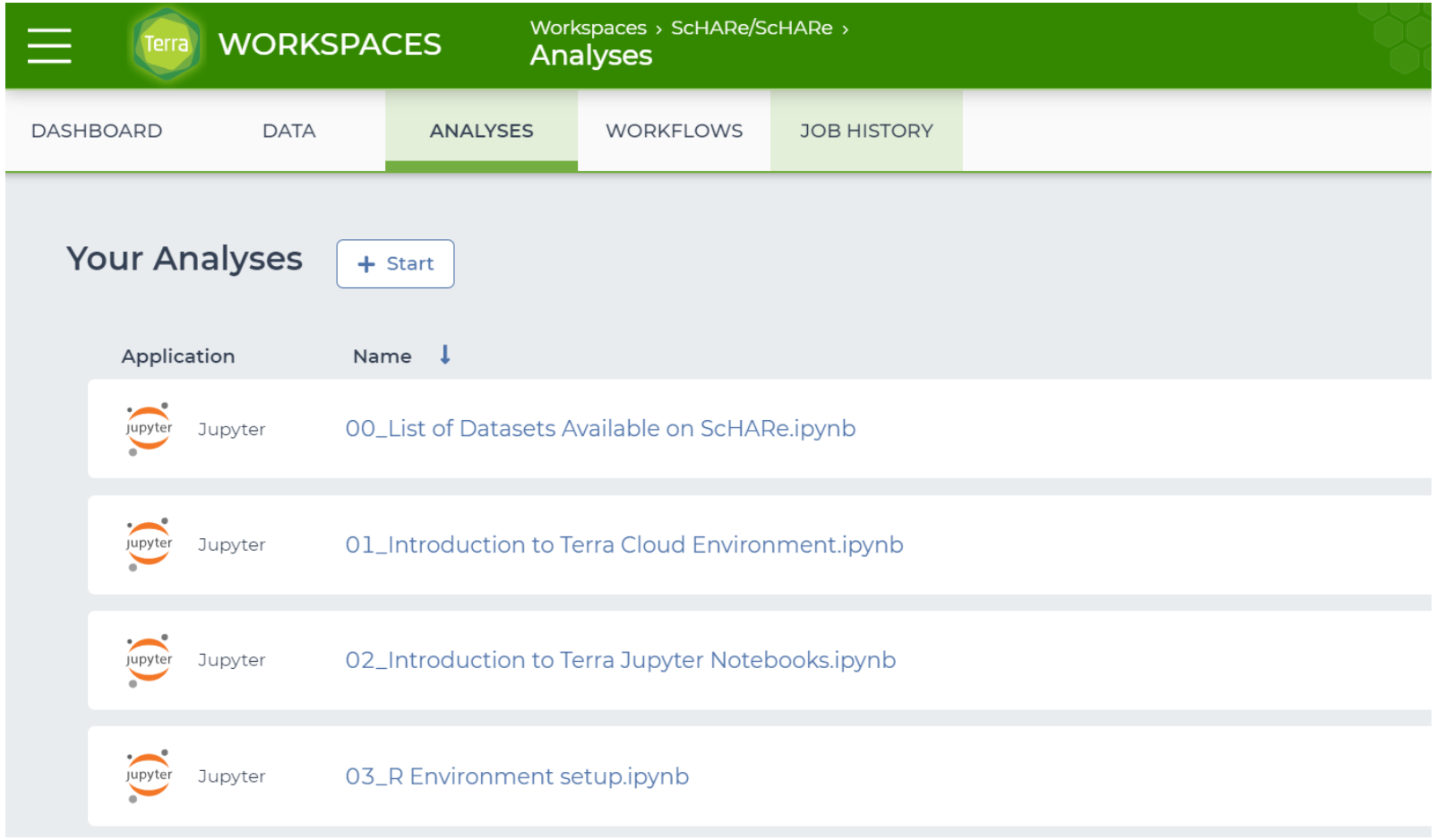
Look for a future Think-a-Thon where we will share more about how to link your data to Terra and how to use the ScHARe Repository to create aggregated data sets!

Analyzing Data on Terra (Jupyter Notebooks)

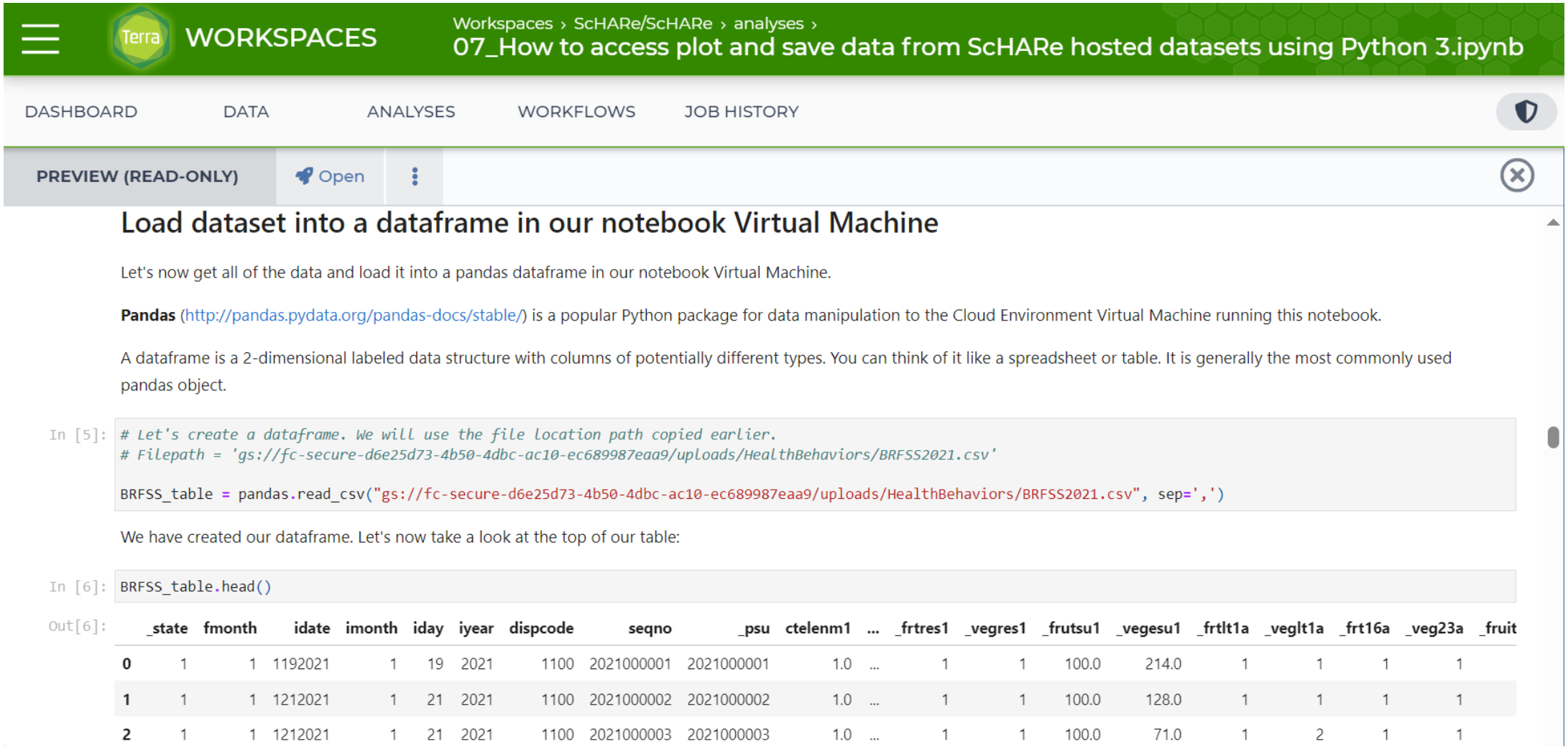


	CASEID	LST	MHINTAKE	OWNERSHP	PUBLICAGENCY	TREATCOGTHRPY	SENIORS	ALZHDEMENTIA	STATE	E_TOTPOP	...	E_HH	E_POV	E_UNEMP	E_PCI	E_NOHSDP
0	201800025	AK	1	2.0	-2.0	1.0	1.0	0.0	ALASKA	738516.0	...	253462.0	77865.0	28067.0	32531.206897	34760.0
1	201800093	AL	1	2.0	-2.0	1.0	1.0	1.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
2	201800099	AL	1	1.0	-2.0	1.0	1.0	1.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
3	201800104	AL	1	1.0	-2.0	1.0	1.0	1.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
4	201800109	AL	1	2.0	-2.0	1.0	1.0	0.0	ALABAMA	4864680.0	...	1860269.0	829400.0	147898.0	23072.835821	470043.0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
779	201809433	PR	1	1.0	-2.0	1.0	0.0	0.0	None	NaN	...	NaN	NaN	NaN	NaN	NaN
780	201809435	PR	1	1.0	-2.0	1.0	1.0	1.0	None	NaN	...	NaN	NaN	NaN	NaN	NaN

Analyzing Data on Terra (Jupyter Notebooks)



Analyzing Data on Terra (Jupyter Notebooks)



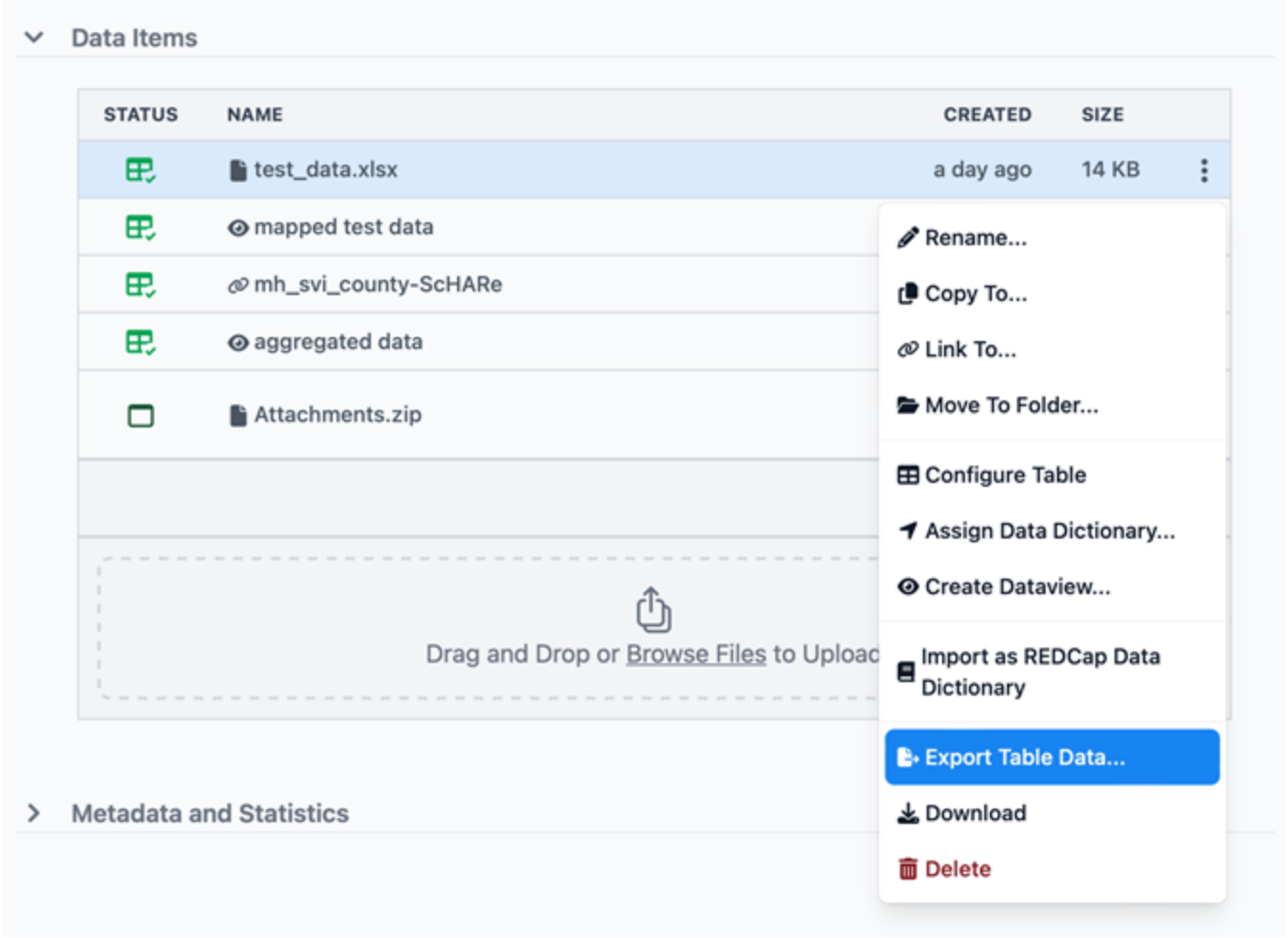
Downloading data to your own computer

Use the item's menu to download the original file

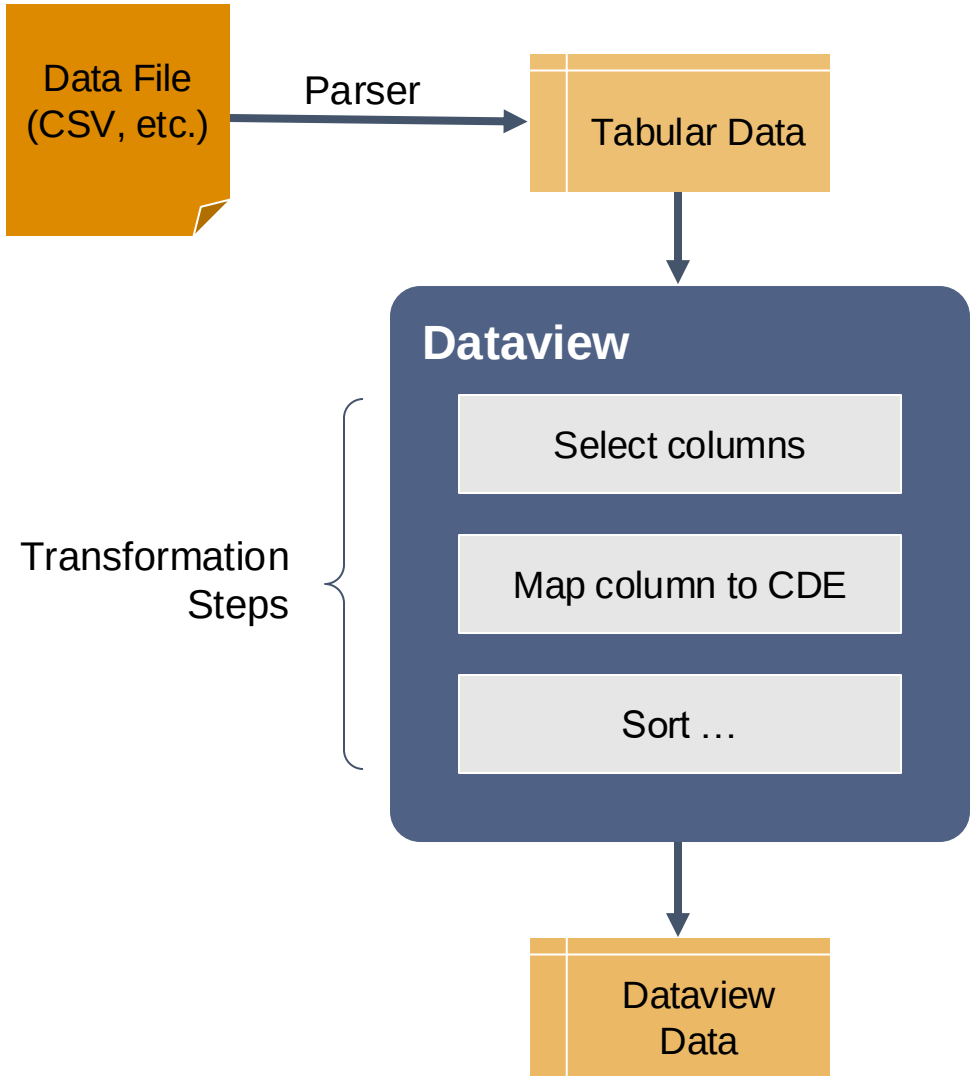
- Best for working directly with raw data

You can also export and download the tabular data in CSV, TSV, or Parquet format

- Best for working with mapped data, summary data, and other processed data



Using Dataviews to Aggregate Data

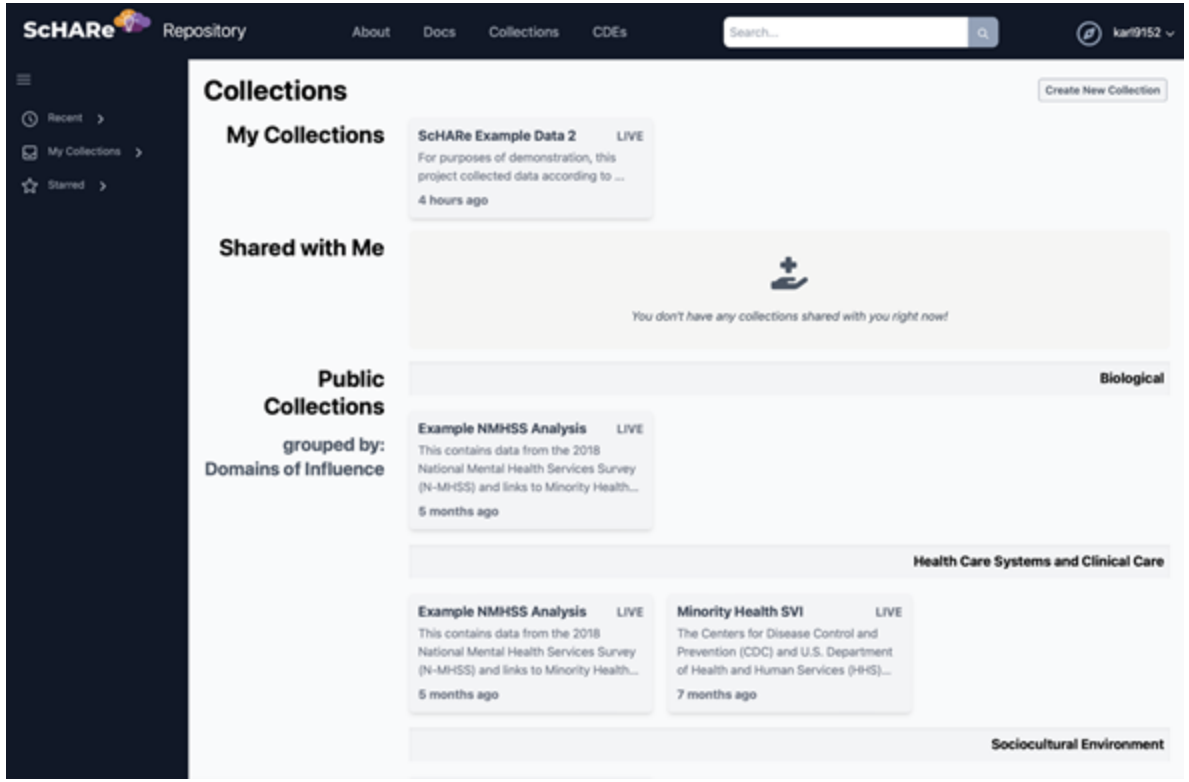


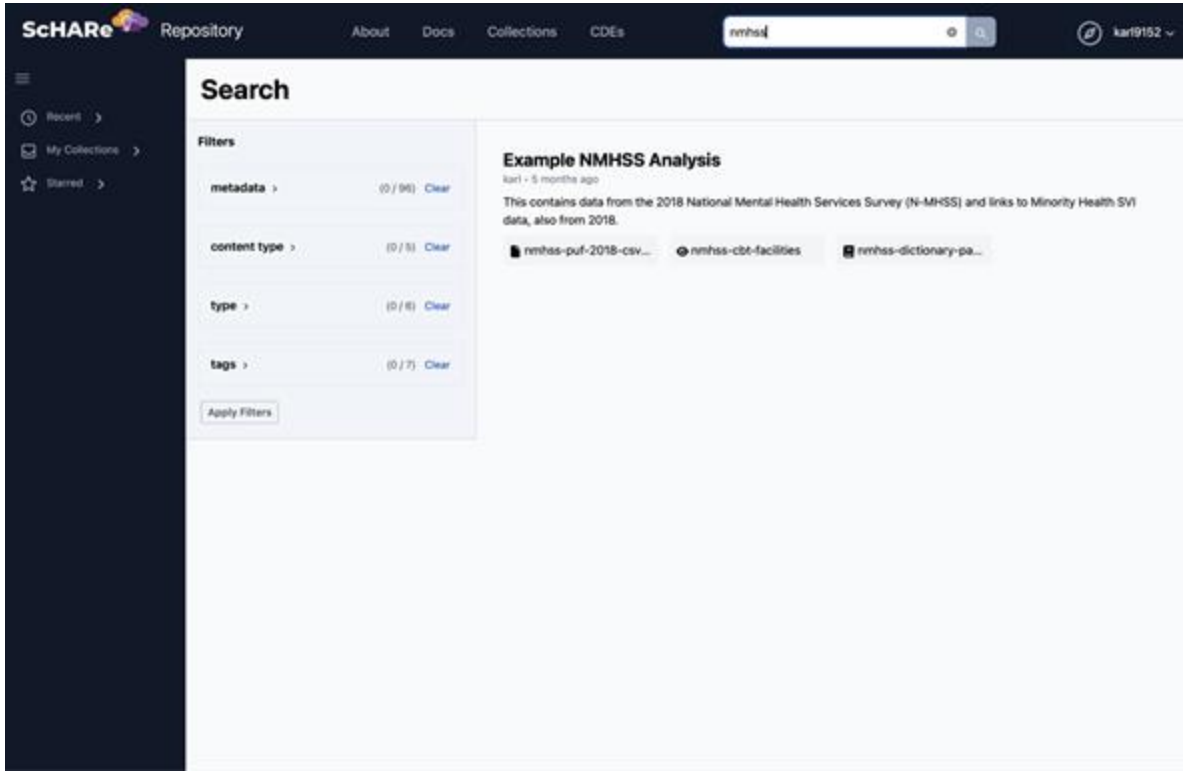
Dataviews take data from one or more sources, apply a series of transformation steps to that data *(filtering, sorting, mapping, etc.)* resulting in a new table of data as output.

Uses of Dataviews:

- Creating subsets of data
- Hiding PHI/PII for publishing
- Summarizing individual-level data into subsets and estimates
- **Joining multiple datasets together**
- Mapping to CDEs
- … many others!

Browsing and Searching for Data

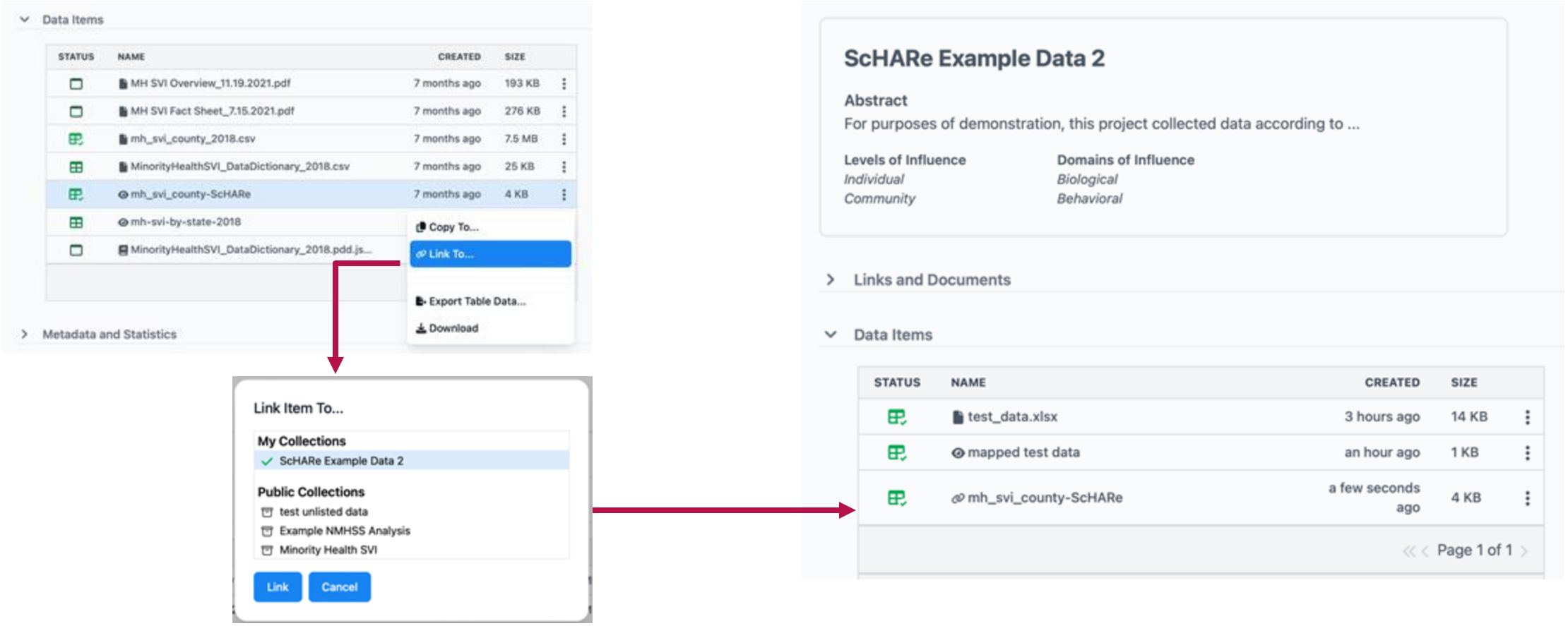




Look for improvements to the data organization and search experience in 2025!

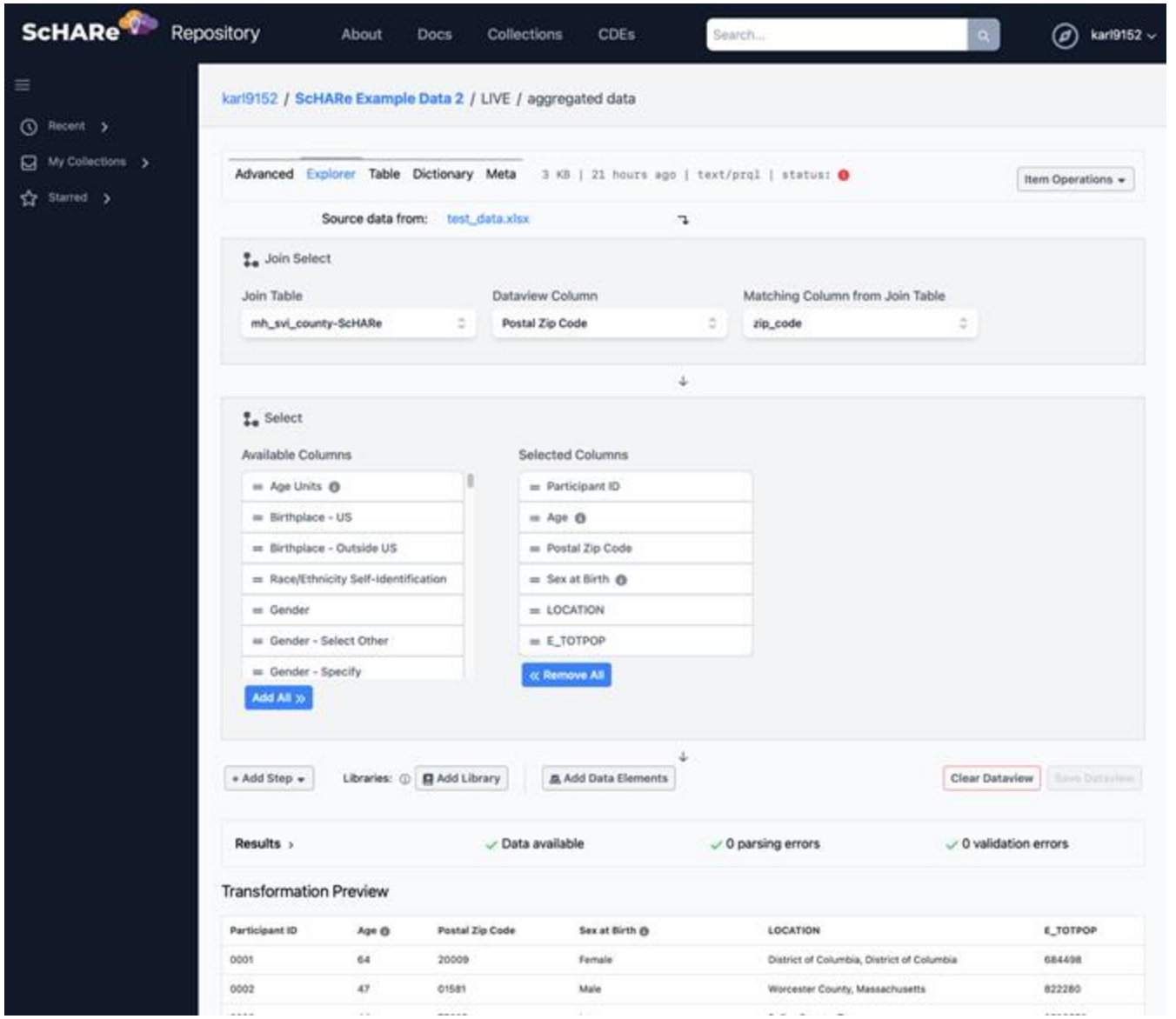
Linking and Aggregating Data

If you find public data that you would like to combine with your project data, you can link it into your own collection and use it just like a file that you had uploaded yourself.



Linking and Aggregating Data

Once linked into your collection, you can create a dataview that combines the public data with your own project data.



ScHARe

Conclusion

BE A PART OF THE FUTURE
OF KNOWLEDGE GENERATION

ScHARe Repository

In summary:

- When getting started, first create a collection, provide metadata, and upload documents.
- Upload your data and use the system to help you map to the ScHARe CDEs.
- View your data, see CDE compliance and analysis readiness.
- Data can be shared with your colleagues, and can be made publicly available after review.
- You can use dataviews to create subsets of your data, join datasets together and more.
- Data can also be analyzed on the ScHARe Terra workspace.

We are here to support you!

We want to hear your questions, issues and comments about the ScHARe Repository!

For any questions regarding how to use the Repository, please reach out to:
✉ schare@mail.nih.gov
Response time: within 24 hours

For Technical Support inquiries, contact:
✉ schare-repository-support@bioteam.net
🕐 Office Hours: Wednesdays, 4-5 PM US Eastern

Questions?

Think-a-Thon poll

1. **Rate how useful this session was:**

- ☐ Very useful
- ☐ Useful
- ☐ Somewhat useful
- ☐ Not at all useful

Think-a-Thon poll

2. Rate the pace of the instruction for yourself:

- ☐ Too fast
- ☐ Adequate for me
- ☐ Too slow

Think-a-Thon poll

3. How likely will you participate in the next Think-a-Thon?

- ☐ Very interested, will definitely attend
- ☐ Interested, likely will attend
- ☐ Interested, but not available
- ☐ Not interested in attending any others

ScHARe

Thank you

ScHARe

Next Think-a-Thons:

bit.ly/think-a-thons

Register for ScHARe:

bit.ly/join-schare

schare@mail.nih.gov