PART I

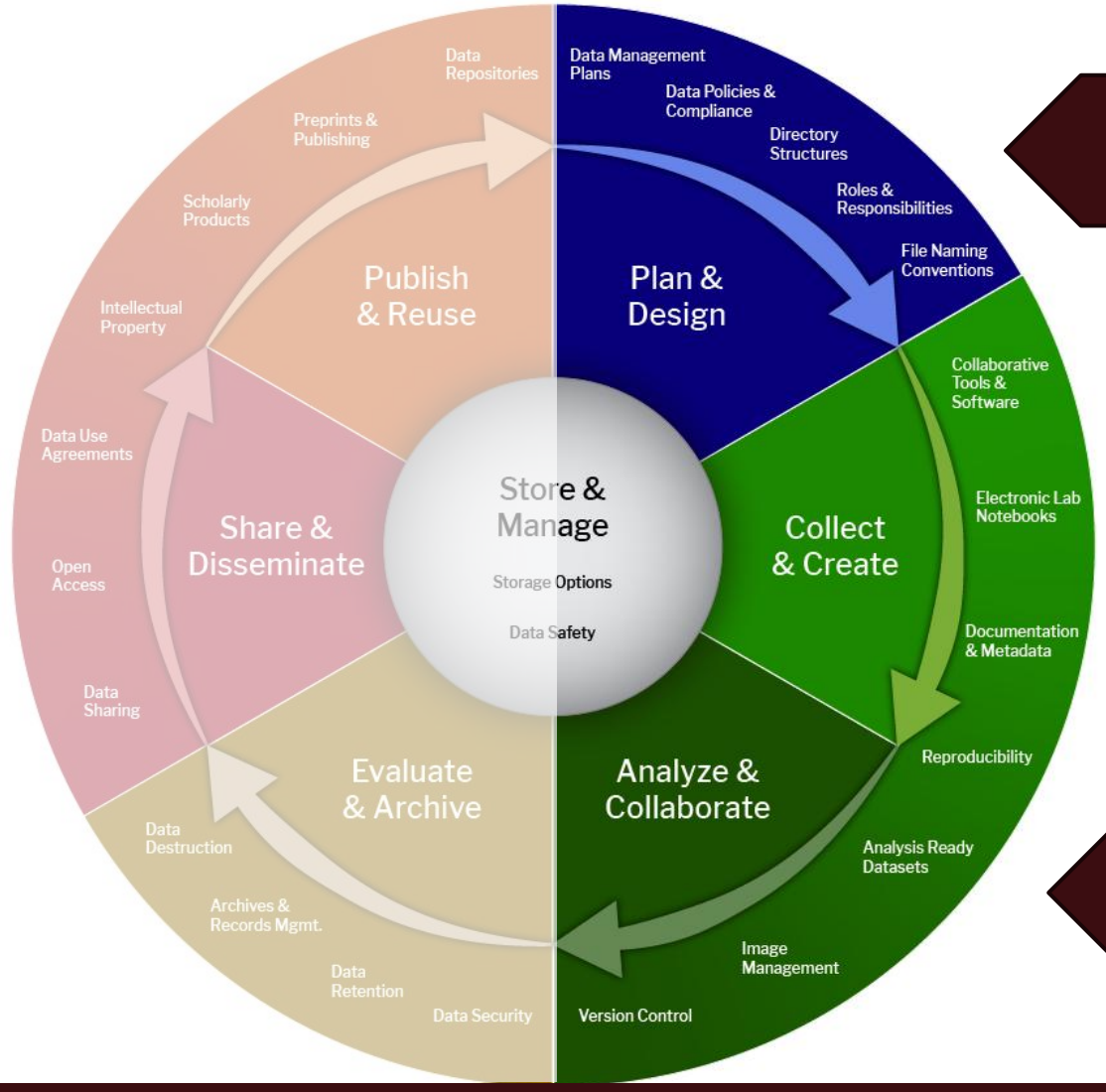
Best Practices: Data and Project Management

Calley E. Fisk
calleyfi@usc.edu

USC | UCLA Center on Biodemography and Population Health

CDR Contextual Data Resource for Aging Surveys

Part I: Overview



Data management

Data workflows

Data organization and documentation

Planning for Data Management

Roadmap your data product prior to beginning your project

Best practice to follow FAIR principles

F

Findability

A

Accessibility

I

Interoperability

R

Reuse

https://www.go-fair.org/fair-principles

F A I R

Findability

Data assigned unique and persistent identifier

Data described with rich metadata

Metadata clearly and explicitly include identifier

Data are registered in a searchable resource (e.g., ICPSR)

Accessibility

Interoperability

Reuse

https://www.go-fair.org/fair-principles

F

Findability

Data assigned unique and persistent identifier

Data described with rich metadata

Metadata clearly and explicitly include identifier

Data are registered in a searchable resource (e.g., ICPSR)

A

Accessibility

Data are retrievable from their identifier using standard communications protocol

Protocol is open, free, and universally implementable

Metadata is accessible, even when data is no longer available

I

Interoperability

R

Reuse

https://www.go-fair.org/fair-principles

F

Findability

Data assigned unique and persistent identifier

Data described with rich metadata

Metadata clearly and explicitly include identifier

Data are registered in a searchable resource (e.g., ICPSR)

A

Accessibility

Data are retrievable from their identifier using standard communications protocol

Protocol is open, free, and universally implementable

Metadata is accessible, even when data is no longer available

I

Interoperability

Data use is formal, accessible, shared, and applicable language used for knowledge representation

Data uses vocabularies that follow FAIR principles

Data include qualified references to other (meta)data

R

Reuse

https://www.go-fair.org/fair-principles

F A I R

Findability

Data assigned unique and persistent identifier

Data described with rich metadata

Metadata clearly and explicitly include identifier

Data are registered in a searchable resource (e.g., ICPSR)

Accessibility

Data are retrievable from their identifier using standard communications protocol

Protocol is open, free, and universally implementable

Metadata is accessible, even when data is no longer available

Interoperability

Data use is formal, accessible, shared, and applicable language used for knowledge representation

Data uses vocabularies that follow FAIR principles

Data include qualified references to other (meta)data

Reuse

Metadata richly described with accurate and relevant attributes

Data are released with clear and accessible data usage license

Data are associated with detailed provenance

Data meet domain-relevant community standards

https://www.go-fair.org/fair-principles

Benefits of Planning & FAIR Principles

- Promotes consistent practices
- Reduces issues of recall
- Compliance with external stakeholders
- Good research ethics
- Workflow efficiency

Data Management Plans (DMP)

Documentation of the planning process

Data Management Plans (DMP)

Documentation of the planning process

Common information included:

- Description and format of data to be produced and shared
- Description of documentation
- Plans for data formatting and metadata
- Plans for data use and security
- Information on data team members

Education and Dementia - Write

dmptool.org/plans/135673/edit?phase_id=2696

DMP Tool

Dashboard | Create Plan | Public Plans | Funder Requirements | About | Logout

University of Southern California (usc.edu)

University of Southern California (usc.edu)

USC Office of Research

Education and Dementia

Project Details | Collaborators | Write Plan | Research outputs | Finalize | Download

This plan is based on the "NIH-Default DMSP" template provided by National Institutes of Health (nih.gov) - (ver: 12, pub: 2024-08-15).

expand all | collapse all

0/12

+ Element 1: Data Type (0 / 3)

+ Element 2: Related Tools, Software and/or Code (0 / 1)

+ Element 3: Standards (0 / 1)

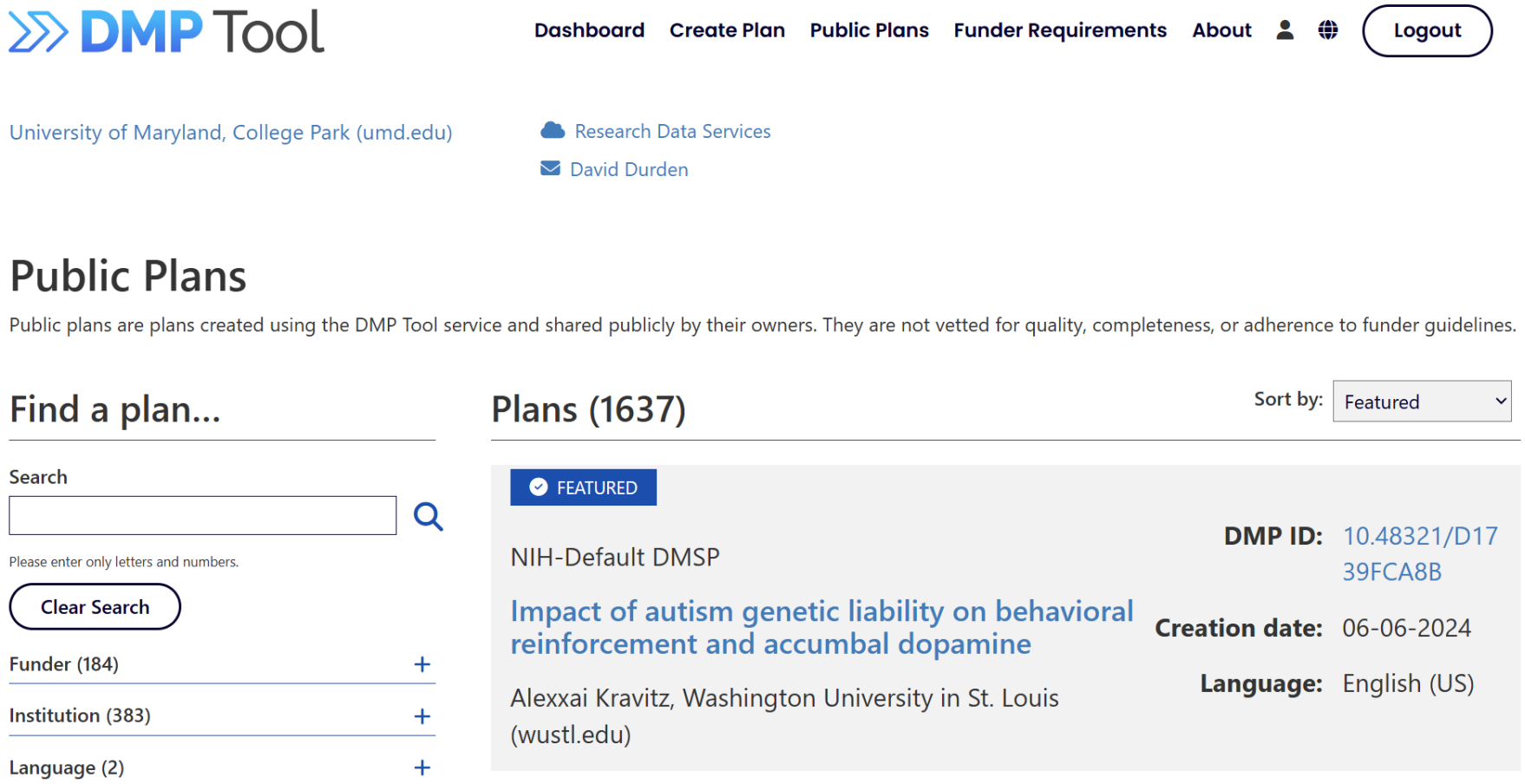
+ Element 4: Data preservation, Access, and Associated Timelines (0 / 3)

+ Element 5: Access, Distribution, or Reuse Considerations (0 / 3)

+ Element 6: Oversight of Data Management and Sharing (0 / 1)

Finding Specific Examples

Example: DMPTool Public Plans



Element 1: Data Type (0 / 3)

Briefly describe the scientific data to be managed, preserved, and shared.

1A.Types and amount of scientific data expected to be generated in the project: Summarize the types and estimated amount of scientific data expected to be generated in the project.

B I 12pt

Press Alt 0 or Option 0 for help using the rich text editor with keyboard only.

Save

Example answer

DMPTool fill-in-the-blank prompt

This project will produce _____ [*Data type, e.g., imaging, sequencing, experimental measurements*] data generated/obtained from _____ [*Data modality, e.g., instrument, method, survey, experiment, data source*]. Data will be collected from ___ [*number*] of research participants/specimens/experiments, generating ___ [*number*] datasets totaling approximately ___ [*amount of data*] in size. The following data files will be used or produced in the course of the project: _____ [*list input data files, intermediate files, and final, post-processed files*]. Raw data will be transformed by ___ [*analysis, method*], and the subsequent processed dataset used for statistical analysis. To protect research participant identities, _____ [*e.g., individual, aggregated, summarized*] data will be made available for sharing.

If working with human subjects, consider adding: Data collection will be performed at clinical sites in the ___ [*location*] area(s) with ___ [*population(s) being studied; i.e., T2 diabetes*].

Sample answer from DMPTool: Basic sciences data

In this proposed project, data will be generated via the following methods: cell culture, light microscopy, confocal microscopy, real-time quantitative polymerase chain reaction (PCR), and stereological counting techniques. This data will be collected from a minimum of 3 independent experiments, with each independent

Comments & Guidance

Guidance | Comments

NIH | DMPTool

NIH Guidance

The 2023 NIH DMS Policy (NOT-OD-21-013) applies to all research, funded or conducted in whole or in part by NIH, that results in the generation of scientific data. NIH defines scientific data as data commonly accepted in the scientific community as of sufficient quality to validate and replicate research findings, regardless of whether the data are used to support scholarly publications. Scientific data does not include laboratory notebooks, preliminary analyses, completed case report forms, drafts of scientific papers, plans for future research, peer reviews, communications with colleagues, or physical objects such as laboratory specimens.

To determine if the DMS policy applies to your research, check out the complete list of NIH activity codes subject to the DMS Policy as well as your funding opportunity.

You may be subject to additional sharing policies. To find out more, check the list of NIH Institute and Center Data Sharing Polcies and/or use the decision tool Which Policies Apply to My Research? This may result in writing a separate Resource Sharing Plan (e.g., Model Organisms or Research Tools) or adding special considerations required by ICs (e.g. validation schedule for NIMH, DMPTool has a separate NIH-NIMH template to incorporate this) or other policies (e.g. Clinical Trial Dissemination Policy) into relevant parts of a DMS plan.

In this section of the Plan, describe data in general terms that address the type, source and amount/size of scientific data expected. Descriptions may indicate data

Internal Data Workflows

1. Project Infrastructure
2. Team Infrastructure

Project Infrastructure

- Tools
- Folders
- Organization

Project Infrastructure

Tools

What do you need to complete your project?

- Options: Python/SQL, Stata, R, Microsoft Office, etc.
- Considerations: Team expertise and training, external formatting, file types

Folders

Organization

Project Infrastructure

Tools

What do you need to complete your project?

- Options: Python/SQL, Stata, R, Microsoft Office, etc.
- Considerations: Team expertise and training, external formatting, file types

Folders

Where does your project live?

- Options: Dropbox, Google Drive, etc.
- Considerations: storage size, institutional agreements, team access, file types, version control

Organization

Project Infrastructure

Tools

What do you need to complete your project?

- Options: Python/SQL, Stata, R, Microsoft Office, etc.
- Considerations: Team expertise and training, external formatting, file types

Folders

Where does your project live?

- Options: Dropbox, Google Drive, etc.
- Considerations: storage size, institutional agreements, team access, file types, version control

Organization

How do you organize your project folder?

- Options: Create a project skeleton with agreed-upon folder organization
- Considerations: “deep” vs. “shallow” structures, consistency, archive practices

How do you organize your project folder?

Example: CDR Crime Data Project Folder README

Uniform Crime Reports Files
USC-CDR Project-Level README
Authors: Calley Fisk (calleyfi@usc.edu)
Last Updated: 7/8/2025 9:49 PM

Project Overview

The Uniform Crime Reports (UCR) data includes annual counts of offenses collected by the FBI Uniform Crime Reporting Program. The data includes counts of murder, robbery, aggravated assault, burglary, larceny, auto theft, and arson. Data are available at the county level and for years 1994 to 2014 and 2016.

Repository Overview

This project directory is organized using the following structure:

HRS Team Folder/Crime/

|-- 01_overview
    |-- 00_source_reference **// original data source materials and web resources**
    |-- 01_Crime_Instruction Manual_1994_2016.docx **// data manual**
    |-- 02_Crime_HRS-CDR_Variables.xlsx **// list of USC-CDR variable names**
    |-- 03_City-Level Data Information_09242019.docx **// notes on city level UCR data**
    |-- 04_Fipscode change check_1994_2016.xlsx **// notes on county changes from version 1 and 2 data**
|--02_processing
    |-- 01_data
        |-- 00_original **// raw downloaded data, folders by year**
        |-- 01_working **// folders for working files in stata format**
        |-- 02_final **// final USC-CDR data files**
        |-- 00_README.txt **// README on file name changes from original to working data**
    |-- 02_scripts
        |-- archive
        |-- 00_Working Data Creation_1994_2016.do **// script for creating working data files**
        |-- 01_Master_crime_data transformation_1994_2016 **// script for processing working data**
        |-- 02_rename_geoids_HRS_to_UAS_Crime.do **// renaming FIPS codes**
        |-- 03_rename_geoids_UAS_to_HRS_Crime.do **// renaming FIPS codes**
|-- 03_data_final
    |-- 01_usc_cdr
        |-- Crime_CDR_Documentation_2020-11-07.docx **// draft of documentation**
        |-- Crime_CDR_Documentation_2020-11-07.pdf
        |-- Crime_County_1994_2014.dta **// version 1 data**
        |-- Crime _County_1994_2016.dta **// version 2 data**
    |-- 02_hrs_cdr
        |-- Crime_HRS-CDR_Documentation_2020-11-07.docx **// draft of documentation**
        |-- Crime_HRS-CDR_Documentation_2020-11-07.pdf
        |-- Crime_County_1994_2014.sas7bdat
        |-- Crime _County_1994_2016.sas7bdat
    |-- 01_uas_cdr
        |-- Crime_UAS-CDR_Documentation_2020-12-02.docx **// draft of documentation**
        |-- Crime_UAS-CDR_Documentation_2020-12-02.pdf
        |-- Crime_County_1994_2014.dta
        |-- Crime _County_1994_2016.dta
|-- archive
|--00_README **// you are here.**

Team Infrastructure

Who does what tasks?

How do they do tasks?

Team Infrastructure

Who does what tasks?

Importance of building teams with appropriate expertise

- Skills: technical, interpersonal, organizational
- Expertise: undergraduate, graduate, post-doc, research scientist

How do they do tasks?

Team Infrastructure

Who does what tasks?

Importance of building teams with appropriate expertise

- Skills: technical, interpersonal, organizational
- Expertise: undergraduate, graduate, post-doc, research scientist

How do they do tasks?

Importance of clear and consistent communication

- Software: Slack, Asana, Trello, etc.
- Balance between day-to-day tasks and major project updates

Team Infrastructure

Who does what tasks?

Importance of building teams with appropriate expertise

- Skills: technical, interpersonal, organizational
- Expertise: undergraduate, graduate, post-doc, research scientist

How do they do tasks?

Importance of clear and consistent communication

- Software: Slack, Asana, Trello, etc.
- Balance between day-to-day tasks and major project updates

Considerations: project timeline, succession planning, consistent scheduling, manager assignments

How do you organize a project workflow?

Example: CDR Shared Project Workflow Diagram

		Who		Data Production			Documentation			Updates
1	README									
2										
3		Who		Data Production			Documentation			Updates
4	**Data Product**	Manager	Team	Downloaded	Processing	Finalized (e.g., variable decisions)	Drafted	Reviewed	Specified to Linkable Survey (e.g., HRS and UAS versions)	Update Needed
5	Decennial Census			Yes	Completed	Completed	Completed	Completed	Completed	No
6	American Community Survey			Yes	Completed	Completed	Completed	Completed	Completed	No
7	USDA Food Environment and Access			Yes	Completed	Completed	Completed	Completed	Completed	Update in Progress
8	FBI Uniform Crime Reports			Yes	Completed	Completed	Completed	Completed	Completed	Update in Progress
9	EPA FAQSD PM2.5 and O3			Yes	Completed	Completed	Completed	Completed	Completed	Update in Progress
10	ATMOS			Yes	Completed	Completed	Completed	In Progress		Yes
11	Nitrogen Dioxide Land Use Estimates			Yes	Completed	Completed	Completed	Completed	Completed	No
12	Dartmouth Atlas of Health Care			Yes	Completed	Completed	Completed	Completed	Completed	No
13	Area Health Resources File			Yes	Completed	Completed	Completed	In Progress	Completed	Update in Progress
14	StreetMap Walkability			Yes	Completed	Completed	Completed	Completed	Completed	No
15	gridMET Weather			Yes	Completed	Completed	Completed	Completed	In Progress	Update in Progress
16	EPA Environmental Quality			Yes	Completed	Completed	Completed	Completed	In Progress	No

Build project infrastructure that emphasizes routine and automation

Tools and organization system only useful if team is trained and invested

Consider scope of project and experience needed

Data Organization

Data organization decisions depend on:

- What does your data look like and what do you want it to look like
- How will you use your data
- Will your data change overtime

Data Organization

Data organization decisions depend on:

- What does your data look like and what do you want it to look like
- How will you use your data
- Will your data change overtime

Variable Naming Conventions

Version Control

Data Formatting and/or Release

Internal and External Documentation

Variable Naming Conventions

Create unique identifier consistent with plans for data analysis

If possible, decide on variable naming rules prior to data processing

BE CONSISTENT!

For value coding, also consider missing codes, reverse coding, value labeling, etc.

Variable Naming Conventions

Create unique identifier consistent with plans for data analysis

If possible, decide on variable naming rules prior to data processing

BE CONSISTENT!

For value coding, also consider missing codes, reverse coding, value labeling, etc.

Example: CDR gridMET Data Naming Conventions

	A	B	C	D	E
1	**NAME**	**DEFINITION**	**GEOGRAPHY**	**YEARS**	
2		**Geographic Identifiers**			
3	LINKCEN2010	geographic identifier	varies	all	
4		**Heat Index Monthly Average Count by Level**			
5	m100	Number of level 0 heat index (no hazard) days in month	tract	annual 1979-2023	
6	m101	Number of level 1 heat index (caution) days in month	tract	annual 1979-2023	
7	m102	Number of level 2 heat index (extreme caution) days in month	tract	annual 1979-2023	
8	m103	Number of level 3 heat index (danger) days in month	tract	annual 1979-2023	
9	m104	Number of level 4 heat index (extreme danger) days in month	tract	annual 1979-2023	
10	m110	Percent of level 0 heat index (no hazard) days in month	tract	annual 1979-2023	
11	m111	Percent of level 1 heat index (caution) days in month	tract	annual 1979-2023	
12	m112	Percent of level 2 heat index (extreme caution) days in month	tract	annual 1979-2023	
13	m113	Percent of level 3 heat index (danger) days in month	tract	annual 1979-2023	
14	m114	Percent of level 4 heat index (extreme danger) days in month	tract	annual 1979-2023	
15					
16					

Variable Naming Conventions

Create unique identifier consistent with plans for data analysis

If possible, decide on variable naming rules prior to data processing

BE CONSISTENT!

For value coding, also consider missing codes, reverse coding, value labeling, etc.

Example: CDR gridMET Data Naming Conventions

	A	B	C	D	E
1	**NAME**	**DEFINITION**	**GEOGRAPHY**	**YEARS**	
2		**Geographic Identifiers**			
3	LINKCEN2010	geographic identifier	varies	all	
4		**Heat Index Monthly Average Count by Level**			
5	m100	Number of level 0 heat index (no hazard) days in month	tract	annual 1979-2023	
6	m101	Number of level 1 heat index (caution) days in month	tract	annual 1979-2023	
7	m102	Number of level 2 heat index (extreme caution) days in month	tract	annual 1979-2023	
8	m103	Number of level 3 heat index (danger) days in month	tract	annual 1979-2023	
9	m104	Number of level 4 heat index (extreme danger) days in month	tract	annual 1979-2023	
10	m110	Percent of level 0 heat index (no hazard) days in month	tract	annual 1979-2023	
11	m111	Percent of level 1 heat index (caution) days in month	tract	annual 1979-2023	
12	m112	Percent of level 2 heat index (extreme caution) days in month	tract	annual 1979-2023	
13	m113	Percent of level 3 heat index (danger) days in month	tract	annual 1979-2023	
14	m114	Percent of level 4 heat index (extreme danger) days in month	tract	annual 1979-2023	
15					
16					

m100tr197901…m100tr202301

Variable Naming Conventions

Create unique identifier consistent with plans for data analysis

If possible, decide on variable naming rules prior to data processing

BE CONSISTENT!

For value coding, also consider missing codes, reverse coding, value labeling, etc.

Example: CDR gridMET Data Variable Labeling Script

```
**# Step 3. Label Variables
foreach file in `files' {
    use `file'

    * extracts the year from the filename
    local year = substr(`"`file'"', 3, 4)

    * heat index files
    if strpos("`file'", "hi") > 0 {

        foreach var of varlist m1* {

            * number of days vs. percent of days
            if strpos("`var'", "m10") > 0 {
                local lab1 = "Number of "
                local lab3 = "days in "
            }
            if strpos("`var'", "m11") > 0 {
                local lab1 = "Percent of days at "
                local lab3 = "in "
            }

            * heat index levels
            if strpos("`var'", "0tr") > 0 {
                local lab2 = "level 0 heat index (no hazard) "
            }
            if strpos("`var'", "1tr") > 0 {
                local lab2 = "level 1 heat index (caution) "
            }
            if strpos("`var'", "2tr") > 0 {
                local lab2 = "level 2 heat index (extreme caution) "
            }
            if strpos("`var'", "3tr") > 0 {
                local lab2 = "level 3 heat index (danger) "
            }
            if strpos("`var'", "4tr") > 0 {
                local lab2 = "level 4 heat index (extreme danger) "
            }
            * month and year
            if strpos("`var'", "`year'01") > 0 {
                local lab4 = "January `year'"
            }
            if strpos("`var'", "`year'02") > 0 {
                local lab4 = "February `year'"
            }
            if strpos("`var'", "`year'03") > 0 {
                local lab4 = "March `year'"
            }
            if strpos("`var'", "`year'04") > 0 {
                local lab4 = "April `year'"
            }
            if strpos("`var'", "`year'05") > 0 {
                local lab4 = "May `year'"
            }
            if strpos("`var'", "`year'06") > 0 {
                local lab4 = "June `year'"
            }
            if strpos("`var'", "`year'07") > 0 {
                local lab4 = "July `year'"
            }
            if strpos("`var'", "`year'08") > 0 {
                local lab4 = "August `year'"
            }
            if strpos("`var'", "`year'09") > 0 {
                local lab4 = "September `year'"
            }
            if strpos("`var'", "`year'10") > 0 {
                local lab4 = "October `year'"
            }
            if strpos("`var'", "`year'11") > 0 {
                local lab4 = "November `year'"
            }
            if strpos("`var'", "`year'12") > 0 {
                local lab4 = "December `year'"
            }

            label variable `var' "`lab1'`lab2'`lab3'`lab4'"
            macro drop lab1 lab2 lab3 lab4
        }
    }
```

Version Control

Identify what a "version" means for your project

Add version indicators and description of change

Keep a change log to track data lineage

Example: CDR Census Data Version Change Log

Version	Version Note	Years	Geographies
1		Census: 1990, 2000, 2010 ACS 5-yr: 2008-12 and 2012-16	State, county, tract
2	Updates ACS 5-year estimates for 2005-09, 2006-10, 2007-11, 2009-13, 2010-14, 2011-15, 2013-17 and 2014-18.	Census: 1990, 2000, 2010 ACS 5-yr: 2005-2009 to 2014-2018	State, county, tract
3	Updates ACS 5-year estimates to 2021, adds ACS 1-year estimates to 2021, reports ACS data at the CBSA level - Inflation-adjusted to 2021 dollars - County change adjusted to 2020 - Corrections made to the variables % males 65+ living alone and % females 65+ living alone	Census: 1990, 2000, 2010 ACS 5-yr: 2005-2009 to 2017-2021	State, CBSA, county, census tract
		ACS 1-yr: 2006 to 2021	State, CBSA, county

Data Formatting and/or Release

Differentiate filenames for internal and external releases

Coordinate with external stakeholders about expectations for data release

Choose data format that coincides with needs for accessibility and metadata

Track data releases

Data Formatting and/or Release

Differentiate filenames for internal and external releases

Coordinate with external stakeholders about expectations for data release

Choose data format that coincides with needs for accessibility and metadata

Track data releases

Example: CDR Census Data Dissemination Overview

Study	Upload Date	Version Number	Who	Upload Location	Delivery Note
MiCDA	March 2023	2		SFTP	Error estimate files are not included for dissemination
	June 2024	3		Dropzone	Error estimate files are included for dissemination
HRS	Dec 2018	1			
	Nov 2020	2		SFTP	Error estimate files are not included for dissemination
	Feb 2025	3		Dropbox	Error estimate files are included for dissemination
UAS	Aug 2020	2		Google drive	Error estimate files are not included for dissemination
	May 2024	3		Google drive	Error estimate files are included for dissemination

Internal and External Documentation

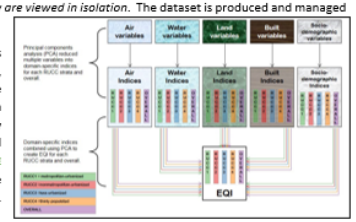
Internal documentation can cover multiple levels of project

Internal documentation should be revisited frequently

Example: CDR Environmental Quality Index Data Acquisition and Management Documentation

Overview

The Environmental Protection Agency Environmental Quality Index (referred to as EPA-EQI in this document) *presents data in five domains: air, water, land, built, and sociodemographic environments to provide a county-by-county snapshot of overall environmental quality across the entire U.S. The EQI helps researchers better understand how health outcomes relate to cumulative environmental exposures that typically are viewed in isolation.* The dataset is produced and managed by the United States Environmental Protection Agency.



The original EPA-EQI dataset provides separate EQI values for each of the five domains (air, water, land, sociodemographic, and built environment). EQI values are generated using principal component analysis (PCA) on a set of selected data from multiple sources. For an overview on the data sources used to create domain-specific EQI values, refer to the EPA-EQI PCA Variable Source Reference file in the \Additional Reference folder. The overall EQI is created using the five domain-specific indices.

Original Data Format

Original data is provided in tabular format as CSV (.csv) files. Each data file provides estimates for the given variable for the given year range for each county in the United States.

New data files must be manually downloaded and may be retrieved here. Refer to the Downloading the Original Data section in this document for instructions for manual download.

As of November 2023, data is available for 2000-2005 and 2006-2010. Data appears to be released every five years; the most-recent data release (2005-2010 data) occurred in September 2020. Based on the previous release schedule, the next data release is expected to occur in September 2025.

Dataset	Release Date
2000-2005	September 2014
2006-2010	September 2020
2011-2015 (expected)	September 2025

Variables (Final Data)

For information on variables derived from the EPA-EQI dataset, refer to the EPA-EQI Variable Reference document.

Variable Series	Description
xxx	2010 county FIPS code
eq series	overall EQI
ai series	air EQI
wa series	water EQI
la series	land EQI
sd series	sociodemographic EQI
be series	built environment EQI

In addition to the sequence identifier, parameter-based variables include extensions to specify the geography and year range. Therefore, parameter-based variables are named based on the following convention:

[variable][geography code][8-digit year range]

For example, data on overall EQI from 2000-2005 would be given the variable name *eqco2000*.

Downloading the Original Data

New EPA files must be manually downloaded from the EPA-EQI Website. Data from the EPA-EQI website is stored as a series of files for each year range for the data. The user should download the .csv file associated with the final EQI data and store that file in the /Original base folder. The user should NOT create folders within the /Original base folder. Based on previous data releases, there does not appear to be a standard naming convention for final EQI dataset .csv files; filenames for previous final EPA-EQI .csv files are listed below for reference.

- **2000-2005 EPA-EQI File:** Eqi_results_2013JULY22.csv
- **2006-2010 EPA-EQI File:** 2006_2010_EQI_2Jan2018_VC.csv

It is recommended that the user also download all files associated with a newly released data year range and store all files associated with a new year range of data in a dedicated folder in the /Additional Reference folder.

Based on the previous release schedule, the next data release is expected to occur in September 2025.

Using the Data Cleaning Script

The spatial aggregation R script EPAEQI_CLEAN, located in the /Working folder, re-organizes and renames data from the original EPA-EQI dataset. Running the script will output a set of two files with identical content but using different storage formats (.csv and .dta) for each year range to the /Final folder.

The data cleaning R script EPAEQI_CLEAN is organized into three sections: Introduction and Overview, User Inputs, and Data Management. The user SHOULD NOT EDIT any code in the data cleaning script outside of specifying user inputs in the User Inputs section.

The user should proceed through the code in the following steps:

1. In the User Inputs section:
   - Set the **working directory**. The working directory should point to the location of the /EPA-EQI folder on the user's machine.
2. Run the code in the Data Management section.

Script Runtime

Running the EPAEQI_CLEAN script takes a few seconds.

R and Package Versions

As of November 2023, the EPAEQI_CLEAN script runs successfully using **R version 4.3.1 (Beagle Scouts)** and using the following package versions.

Package	Version	Use in Script
foreign	0.8.85	write .dta files

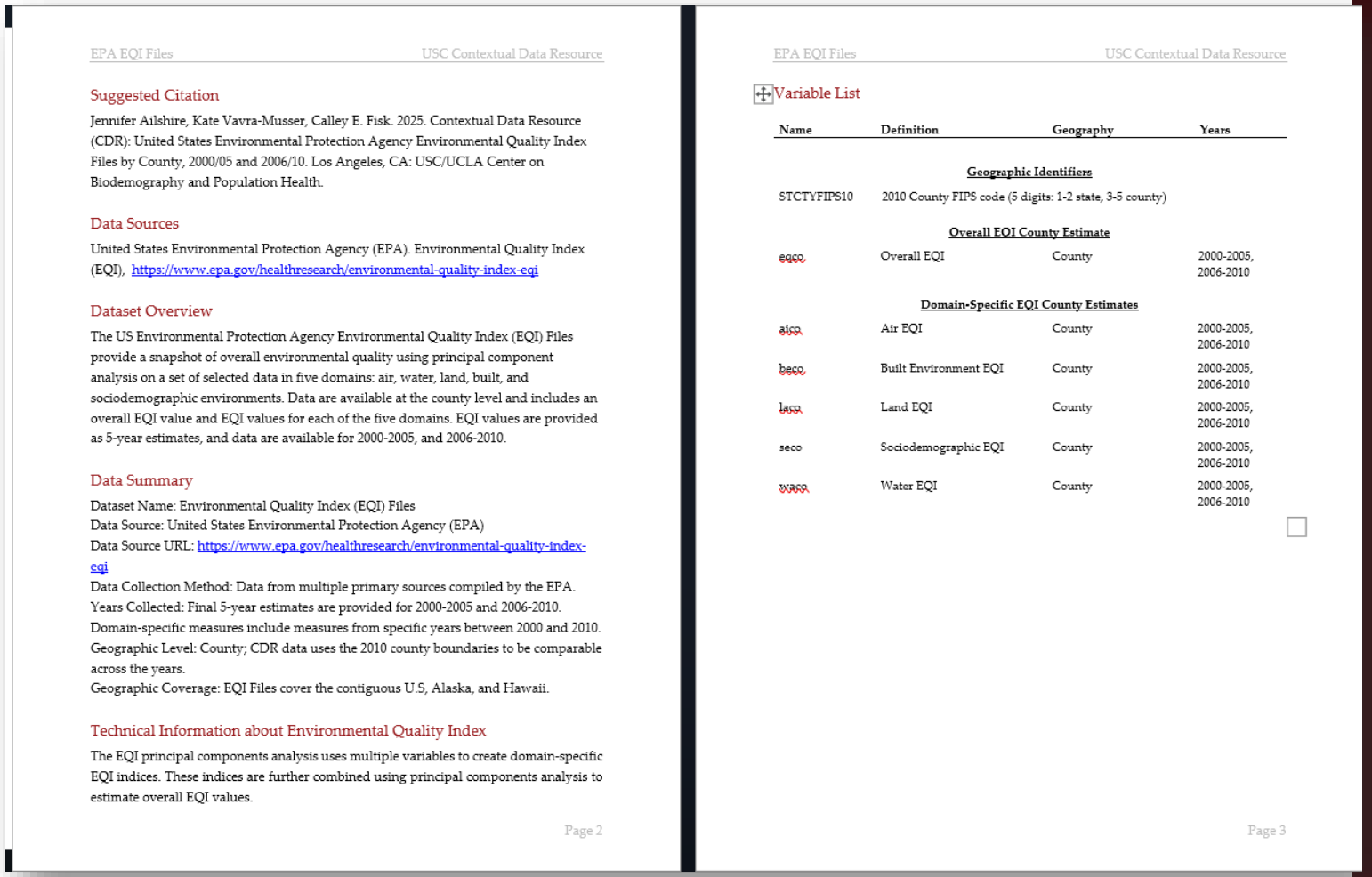
Internal and External Documentation

For external documentation, amount and quality matters

External documentation often includes

1. Project information
2. Data definitions and codebook
3. Metadata
4. Suggested citations
5. Code for data management
6. Version information

Example: CDR Environmental Quality Index Data Documentation



Suggested Citation

Jennifer Ailshire, Kate Vavra-Musser, Calley E. Fisk. 2025. Contextual Data Resource (CDR): United States Environmental Protection Agency Environmental Quality Index Files by County, 2000/05 and 2006/10. Los Angeles, CA: USC/UCLA Center on Biodemography and Population Health.

Data Sources

United States Environmental Protection Agency (EPA). Environmental Quality Index (EQI), https://www.epa.gov/healthresearch/environmental-quality-index-eqi

Dataset Overview

The US Environmental Protection Agency Environmental Quality Index (EQI) Files provide a snapshot of overall environmental quality using principal component analysis on a set of selected data in five domains: air, water, land, built, and sociodemographic environments. Data are available at the county level and includes an overall EQI value and EQI values for each of the five domains. EQI values are provided as 5-year estimates, and data are available for 2000-2005, and 2006-2010.

Data Summary

Dataset Name: Environmental Quality Index (EQI) Files
Data Source: United States Environmental Protection Agency (EPA)
Data Source URL: https://www.epa.gov/healthresearch/environmental-quality-index-eqi
Data Collection Method: Data from multiple primary sources compiled by the EPA.
Years Collected: Final 5-year estimates are provided for 2000-2005 and 2006-2010.
Domain-specific measures include measures from specific years between 2000 and 2010.
Geographic Level: County; CDR data uses the 2010 county boundaries to be comparable across the years.
Geographic Coverage: EQI Files cover the contiguous U.S, Alaska, and Hawaii.

Technical Information about Environmental Quality Index

The EQI principal components analysis uses multiple variables to create domain-specific EQI indices. These indices are further combined using principal components analysis to estimate overall EQI values.

Variable List

Name	Definition	Geography	Years
	Geographic Identifiers		
STCTYFIPS10	2010 County FIPS code (5 digits: 1-2 state, 3-5 county)		
	Overall EQI County Estimate		
eqco	Overall EQI	County	2000-2005, 2006-2010
	Domain-Specific EQI County Estimates		
aico	Air EQI	County	2000-2005, 2006-2010
beco	Built Environment EQI	County	2000-2005, 2006-2010
laco	Land EQI	County	2000-2005, 2006-2010
seco	Sociodemographic EQI	County	2000-2005, 2006-2010
waco	Water EQI	County	2000-2005, 2006-2010

PART I

Q & A

Thank you!

Calley E. Fisk
calleyfi@usc.edu

CBPH USC | UCLA Center on Biodemography and Population Health

CDR Contextual Data Resource for Aging Surveys