National Institute of Environmental Health Sciences
Your Environment. Your Health.

Resources and Tools to Support Integration of Geospatial and Environmental Exposure Data into Epidemiological and Clinical Health Research

Joint Annual Meeting of the International Society of Exposure Science
and the International Society for Environmental Epidemiology

Atlanta, GA
August 17, 2025

CAFE
Accelerating Climate & Health Research

WORKSHOP AGENDA

WELCOME AND SETTING THE STAGE	8:00 AM – 8:15 AM
CHORDS DATA ECOSYSTEM	8:15 AM – 8:20 AM
CAFE DATA REPOSITORY	8:20 AM – 8:25 AM
SHOWCASE: CHORDS DATA CATALOG	8:25 AM – 8:50 AM
SHOWCASE: CAFE DATA RESOURCES	8:50 AM – 9:15 AM
TUTORIAL: AMADEUS SOFTWARE	9:15 AM – 10:05 AM
TUTORIAL: CAFE TOOLKIT	10:05 AM – 10:55 AM
FACILITATED DISCUSSION: DATA STANDARDS	10:55 AM – 11:15 AM
PANEL DISCUSSION/Q&A	11:15 AM – 11:30 AM
ADJOURN	11:30 AM

Learning Objectives

- Increase awareness of existing geospatial environmental exposure data, tools, and resources
- Improve understanding of evolving approaches for linking and harmonizing exposure and health data
- Considerations for data standards in the context of epidemiologic and clinical research

Workshop Goals

- Build a network of researchers working in this space
- Solicit feedback on tools and resources
- Disseminate resources and opportunities related to the CHORDS and CAFE programs to a wider audience
- Join our stakeholder group/community of practice

CHORDS Overview

Connecting Health Outcomes Research and Data Systems (CHORDS)

Goal: Strengthen data infrastructure to facilitate research connections between environmental exposures and health outcomes so researchers can 1) identify, analyze, and reduce the health effects associated with disaster-related events (e.g., wildfires) and 2) improve patient and population health outcomes

CHORDS Objectives and Deliverables

Objective 1:
Web-based data resources catalog

Objective 2:
Standardized, linked datasets

Objective 3:
Toolkit of software and tutorials

Objective 4:
Evaluation use case

Objective 5:
Addressing end user needs

CHORDS Data Platform

About CHORDS

The CHORDS project aims to build and strengthen data infrastructure for patient-centered outcomes research on environment and health.

Browse Catalog

View a catalog of over 100 CHORDS Data Resources.

Highlighted Research

Collection of case studies intended to provide examples of research articles that examine the health effects of wildfire-related exposures and highlight the key environment and health data sets used in these studies.

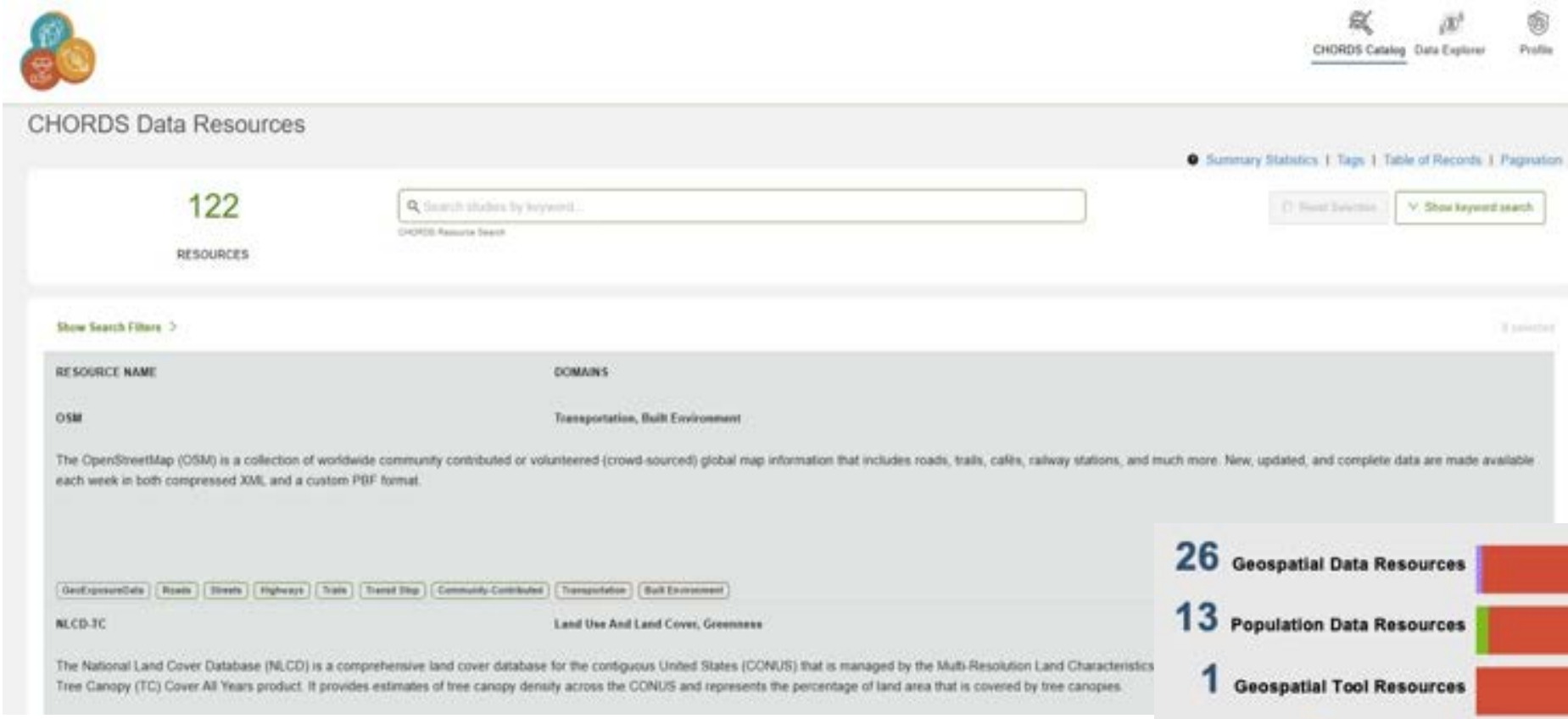
Software

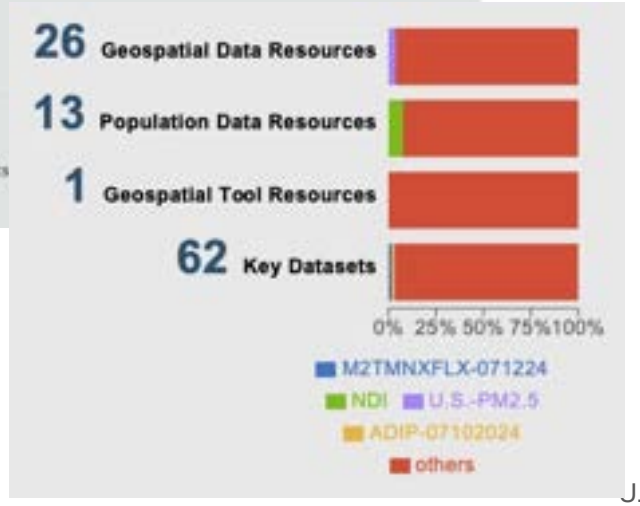
Find out more and download the CHORDS software.

Training and Use Cases

The CHORDS project seeks to connect researchers with guides, tutorials, and example code.

Data Catalog: Manually Curated Searchable Geospatial and Population Resources





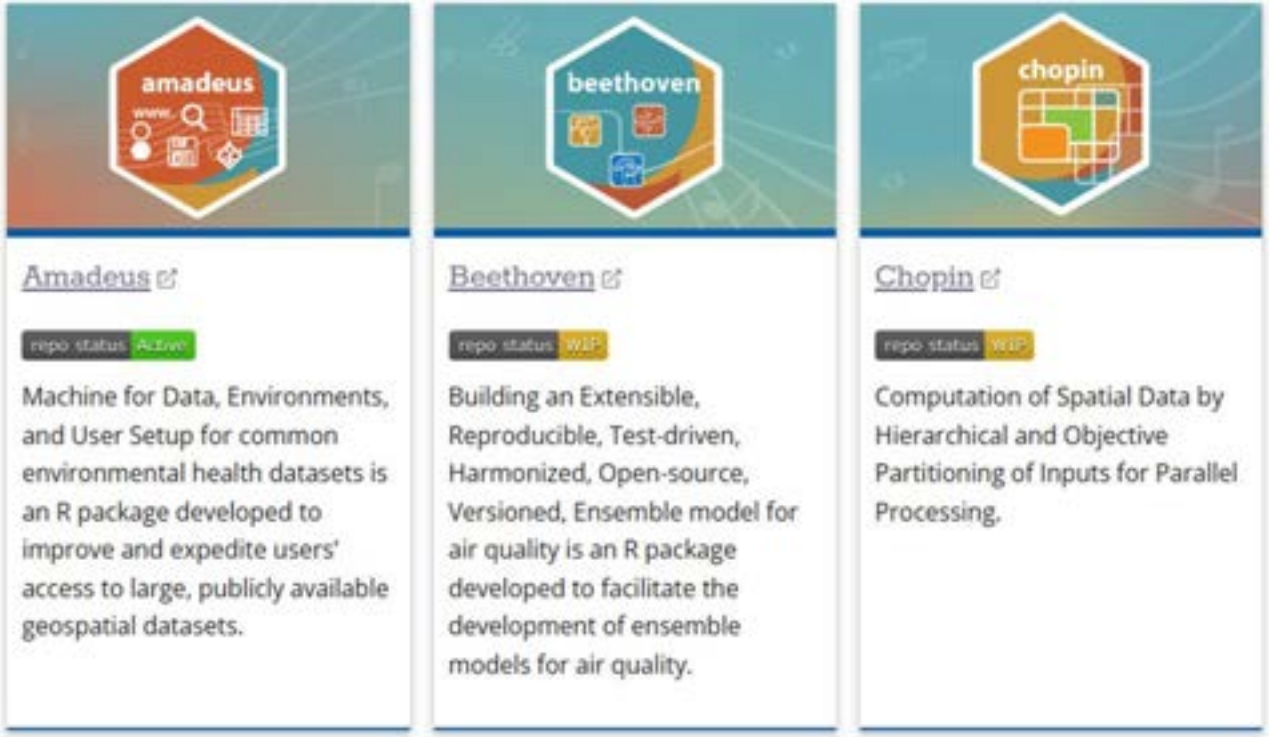
Metadata provided:

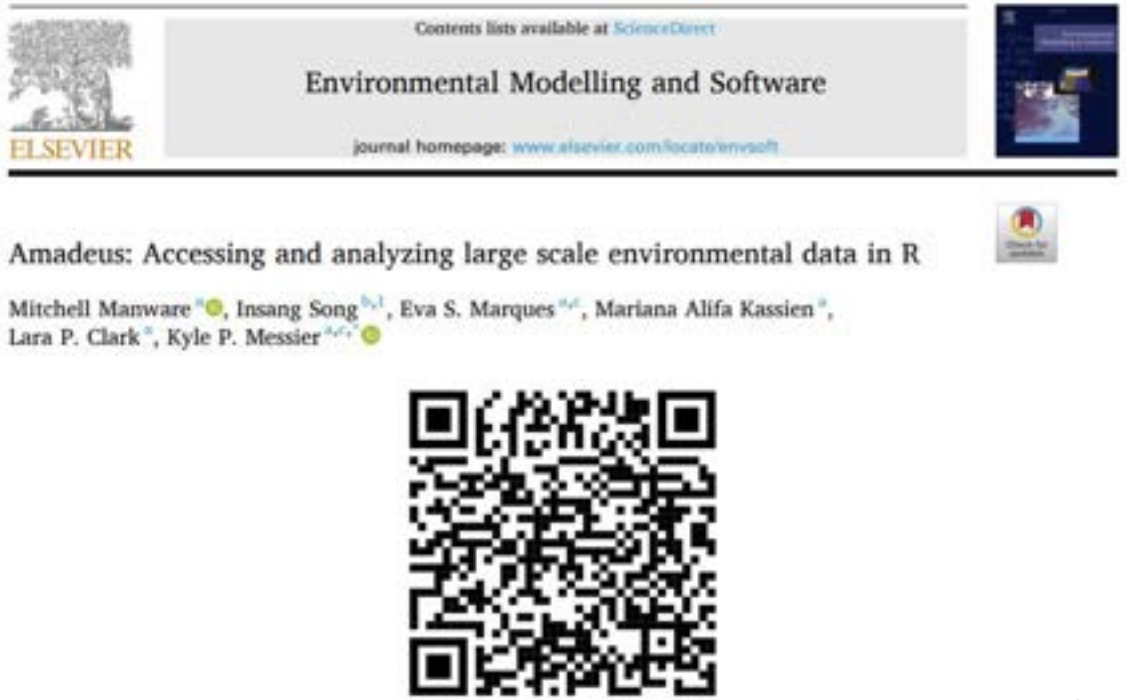
- Project information
- Measures
- Spatial characteristics
- Temporal characteristics
- Data access information
- Other:
  - Keywords
  - Citations
  - Publications

Open-Source Software and Tools for Simplifying Environmental Health Data Analysis

CHORDS-specific tools

- ***amadeus:*** Downloads large-scale environmental and weather data and gets it ready for analysis
- ***beethoven:*** A reproducible and extensible pipeline for air pollution exposure designed for updated and timely releases
- ***chopin:*** Simplifies parallelization, running many calculations or processes simultaneously, of big environmental exposure data

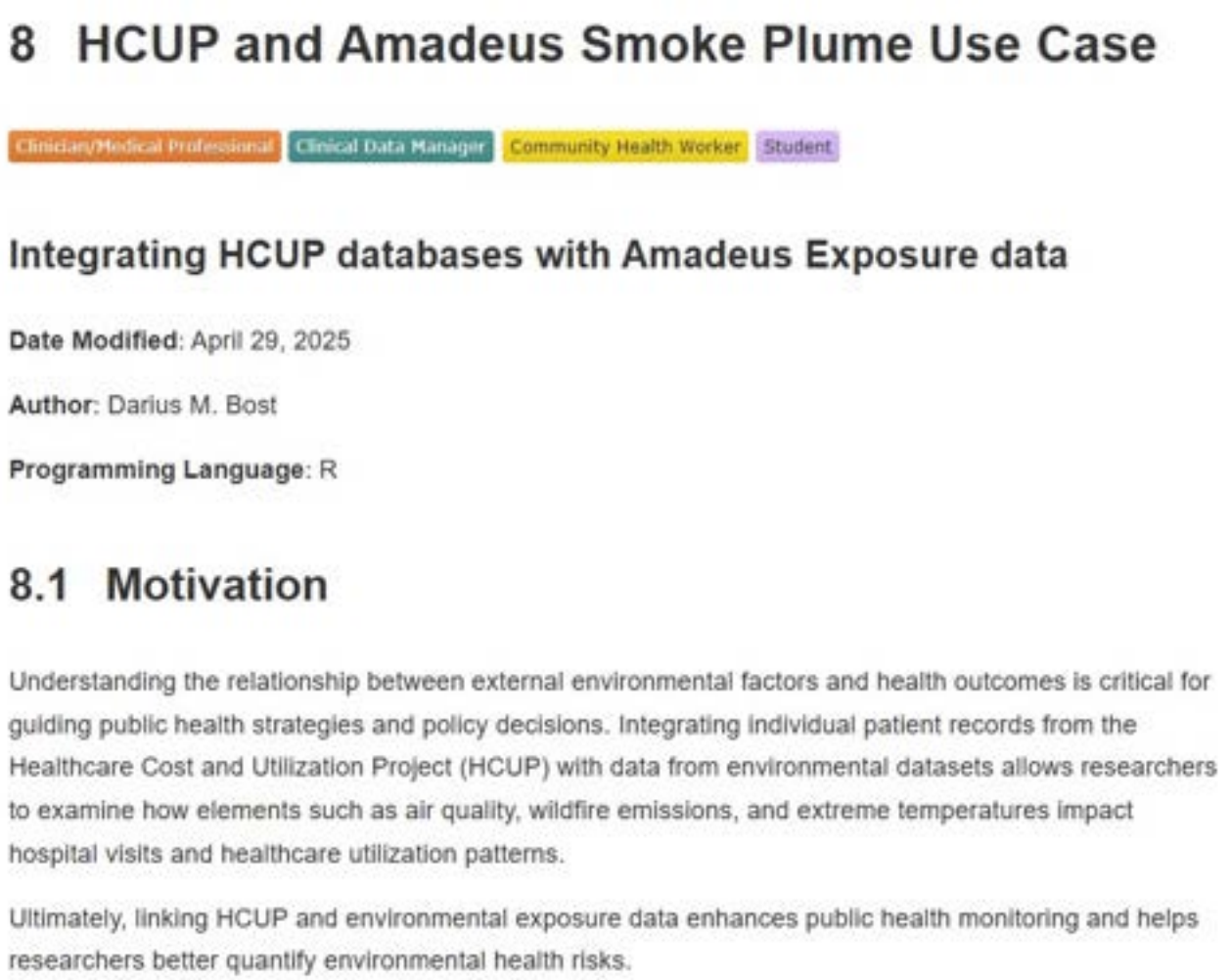




Toolkit of Code, Data, and Educational Materials

Open-source resources: Tutorials and educational materials to support different types of users (e.g., students, epidemiologists, clinicians, data managers) in accessing, processing, and integrating geospatial exposure data into health research

- **Geospatial Data Foundations**
  - Working with point, polygon, and raster data
- **Health Data Integration**
  - Linkage to Census Units
  - FHIR PIT Tutorial
- **Use Case Studies**
  - AHRQ HCUP Analysis



8 HCUP and Amadeus Smoke Plume Use Case

Clinician/Medical Professional | Clinical Data Manager | Community Health Worker | Student

Integrating HCUP databases with Amadeus Exposure data

Date Modified: April 29, 2025

Author: Darius M. Bost

Programming Language: R

8.1 Motivation

Understanding the relationship between external environmental factors and health outcomes is critical for guiding public health strategies and policy decisions. Integrating individual patient records from the Healthcare Cost and Utilization Project (HCUP) with data from environmental datasets allows researchers to examine how elements such as air quality, wildfire emissions, and extreme temperatures impact hospital visits and healthcare utilization patterns.

Ultimately, linking HCUP and environmental exposure data enhances public health monitoring and helps researchers better quantify environmental health risks.

Use Case: Healthcare Cost and Utilization Project (HCUP) Analyses

Spatial analysis: bivariate map of asthma prevalence and heavy smoke exposure across ZIP codes in Oregon

- Map shows areas with overlapping smoke and health risks

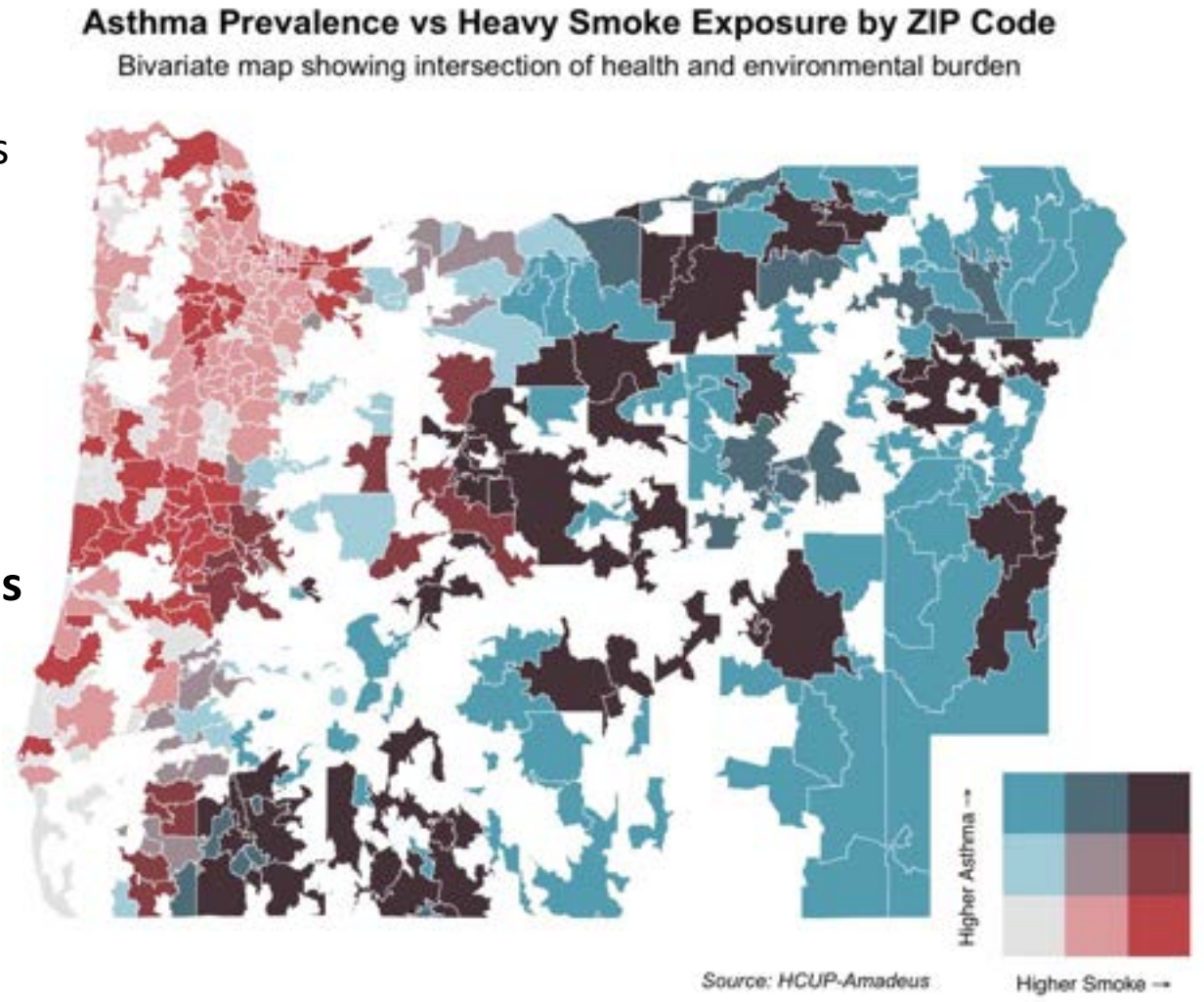
Logistic regression model: examine relationship between asthma diagnoses and exposure to different levels of smoke density

- Exposure to medium and heavy smoke is associated with significantly increased odds of asthma

Findings can support targeted public health interventions

Legend:
- X-axis (**red**): higher smoke exposure
- Y-axis (**blue**): higher asthma

Color interpretations:
- **Dark red**: high smoke, low asthma
- **Dark blue**: high asthma, low smoke
- **Dark purple**: high smoke, high asthma
- **Light gray**: low smoke, low asthma



CHORDS Team and Contributors

Core Team

- Adam Burkholder
- Alison Motsinger-Reif
- Ann Liu
- Aubrey Miller
- Charles Schmitt
- David Fargo
- David Reif
- Kyle Messier
- Trisha Castranio

NIH

- Asif Rizwan (NCI)
- Erin Iturriaga (NHLBI)
- Regina Bures (NHLBI)
- Richard Kwok (NIA)
- Umit Tokac (NHLBI)
- Cristina Justice (AofU)

Other Agencies

- Pamela Owens (AHRQ)
- Patricia Keenan (AHRQ)
- NASA representatives

NIEHS

- Darius Bost
- Deep Patel
- Jennifer Fostel
- Lara Clark
- Marcus Jackson
- Maria Shatz
- Mariana Kassien
- Mike Conway
- Mitchell Manware
- Rupali Gupta
- Skylar Marvel
- Sue Nolte

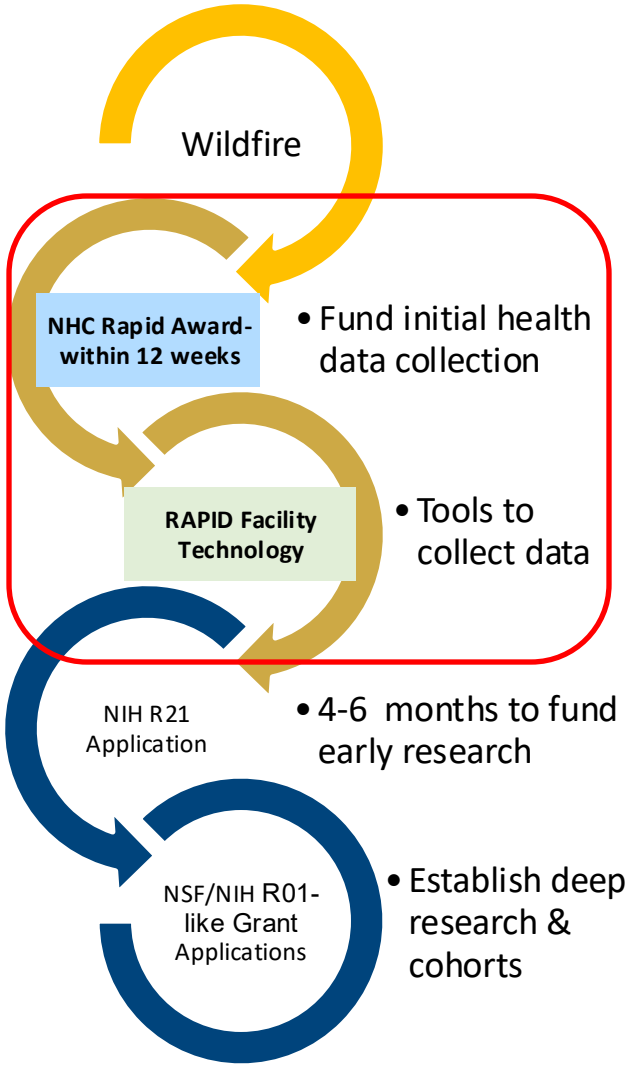
Technical Expert Panel (TEP)

- Angela Werner (CDC)
- Caleb Dresser (Harvard)
- Cavin Ward-Caviness (EPA)
- Cole Brokamp (Univ. Cincinnati)
- Genee Smith (Johns Hopkins)
- Kevin Lane (BU/CAFÉ)
- Patrick Wall (CDC)
- Philip Awadalla (Canada Data Integ. Ctr)
- Rima Habre (USC)
- Yang Liu (Emory)

NIH-NSF Collaboration Centers: Promoting Time-Critical Extreme Weather and Natural Disaster Health Data Collection and Research

- **Natural Hazards Center (NHC) (University of Colorado-Boulder)**
  - Fund awards to collect perishable health data focused on high-risk groups and time-sensitive situations (e.g., pregnancy, comorbidities, workers, at-risk populations)
  - Quick response awards between $10-50K within 12 weeks of hurricanes, wildfires, floods, tornadoes, etc.

- **RAPID Facility Program (University of Washington)**
  - Acquisition and access to health-focused sensors, instrumentation, support, and training for collection of post-natural disaster exposure and health data
  - Field testing resources during 2025 Los Angeles wildfires



CHORDS DATA CATALOG

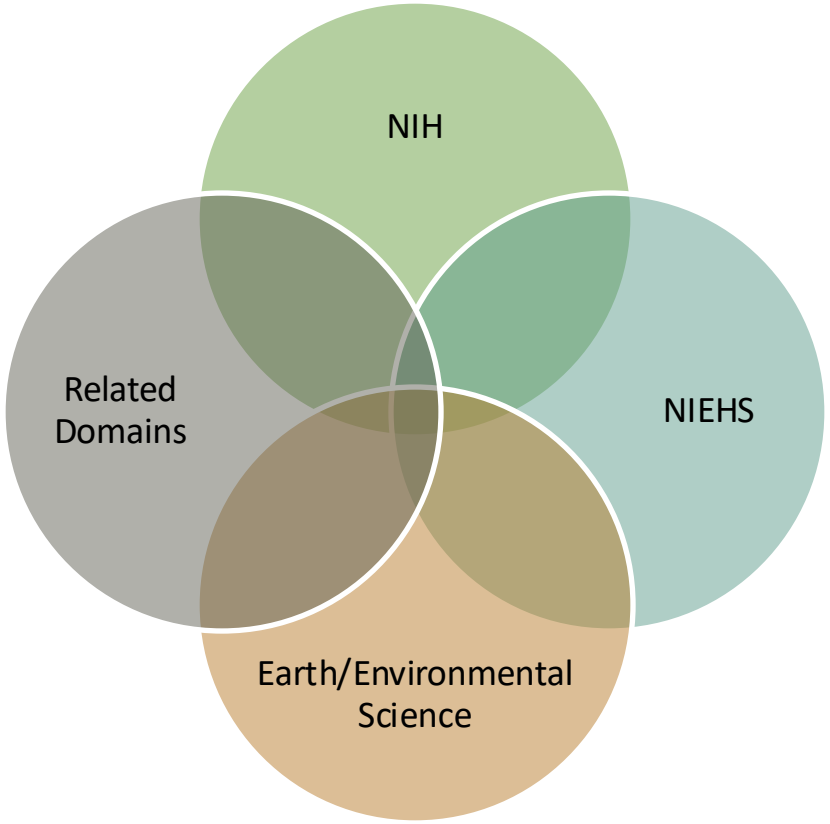
Joint Annual Meeting of the International Society of Exposure Science
and the International Society for Environmental Epidemiology

Atlanta, GA
August 17, 2025

Topics

- Design Principles
- CHORDS Portal Phase 1
- Curation Pipeline
- CHORDS **Phase 2**
  - Automated curation
  - Multivalent catalog
  - Formalized workflows
- Lessons Learned

CHORDS in the Context of the Exposome

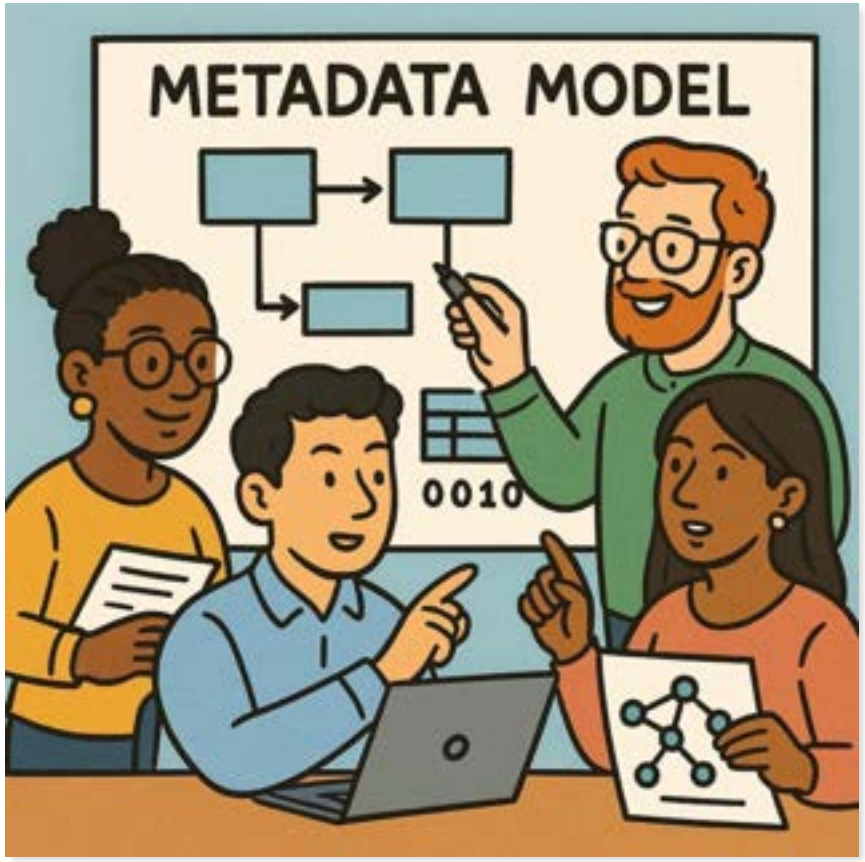


While focused on specific CHORDS use cases, this is an opportunity to think about the broader ecosystem studying the Exposome.

- Look ‘inward’ at other parts of NIH
- Ease federation with other platforms common in other domains
- Be flexible in terms of source and target schema while maintaining as much structure and validation as possible
- Think about the provenance and verification of data over time
- Prepare for increased use of AI for both metadata extraction and catalog automation as well as the use of AI in data discovery

Designing the Catalog

- This is really about the metadata model, first and foremost
- With such a broad set of domains, we would have to build out the model as we iterated
  - Tooling to allow curators to design and communicate design as well as curate the data
  - Workflows that allowed constant migration
- Community standards, ontologies for crossing environment and health not readily available
  - ECTO - https://pmc.ncbi.nlm.nih.gov/articles/PMC9951428/
  - EXO - https://pmc.ncbi.nlm.nih.gov/articles/PMC3314380/



Designing the System

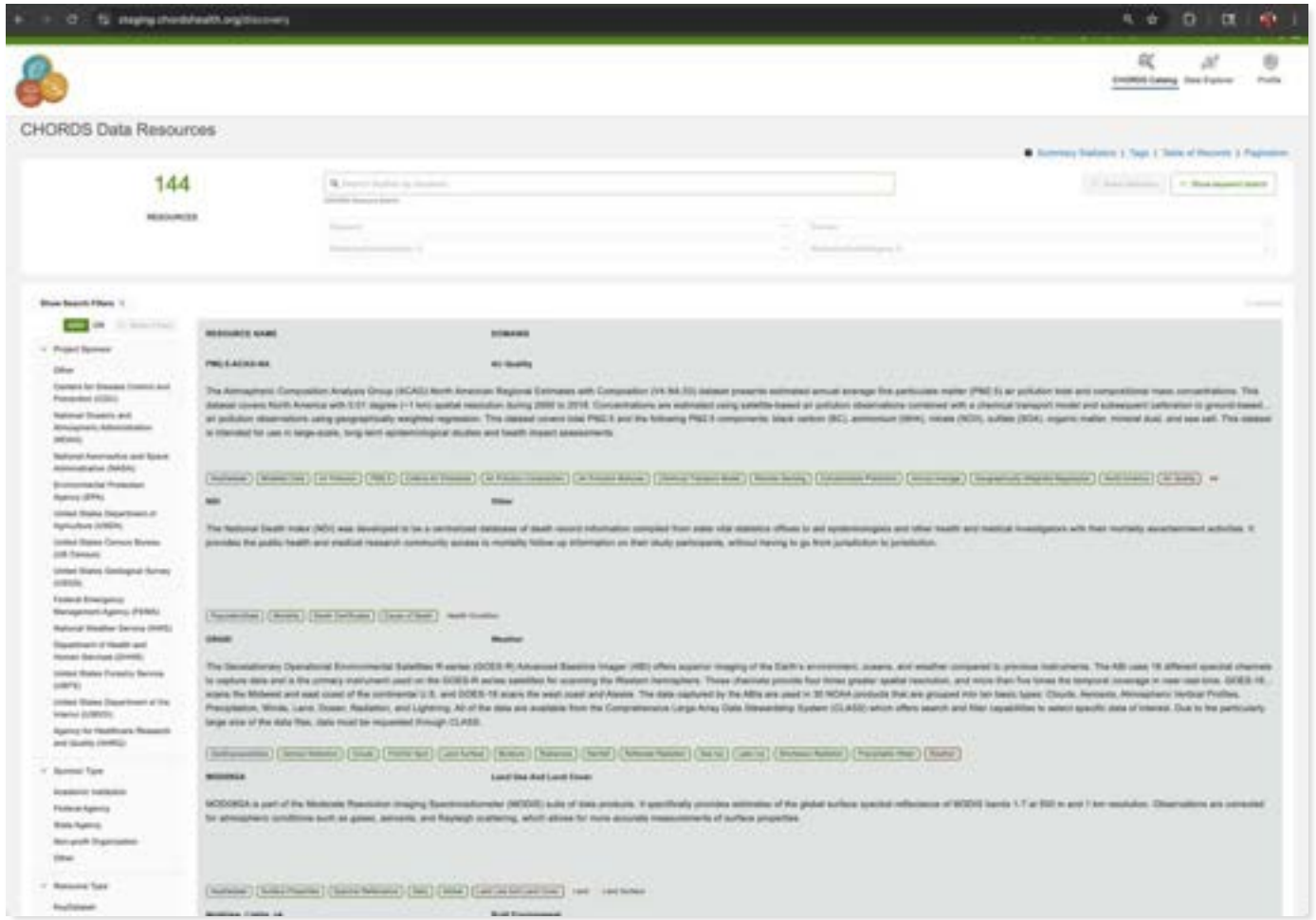
- The care and feeding of the metadata model development and curation was central
- The system needed to enable the entire workflow, from design to curation to validation to presentation
- The data itself is the catalog, and the system needed flexibility to slice, dice, repurpose and federate
- The catalog needed to fit into an expanding, federated and heterogeneous cyberinfrastructure picture

Architectural Design Principles

- Avoid building bespoke systems, use existing platforms.
- Focus on trans-NIH standards and platforms.
- Join communities (standards and open-source).
- API first (offer and utilize open API at the foundation).

The CHORDS Portal

- https://chordshealth.org/discovery
- Gen3 based
  - Open Source - https://github.com/uc-cdis
  - GA4GH Compliant
  - FedRAMP certified
- Operated by Gen3 on AWS

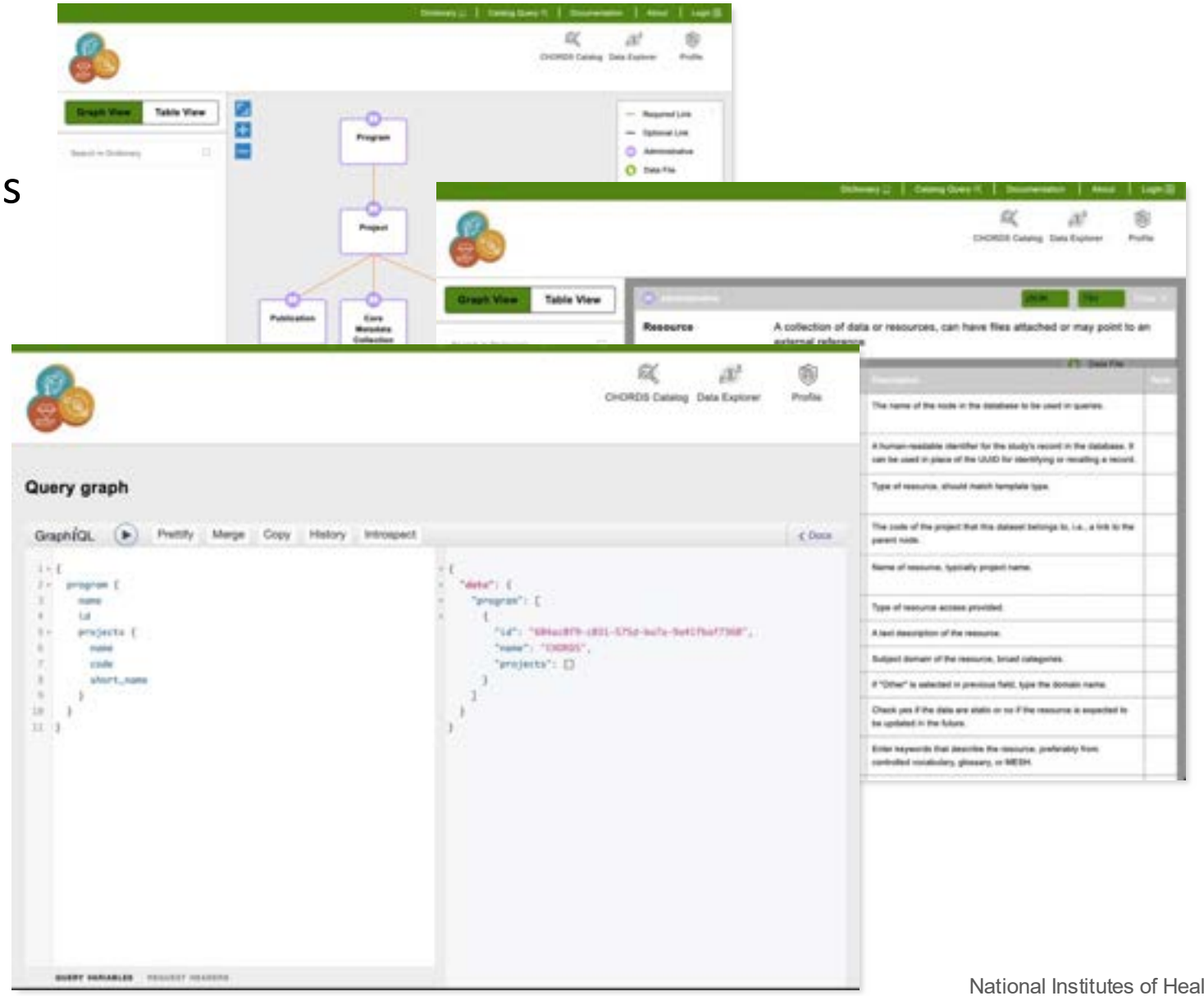


GEN3

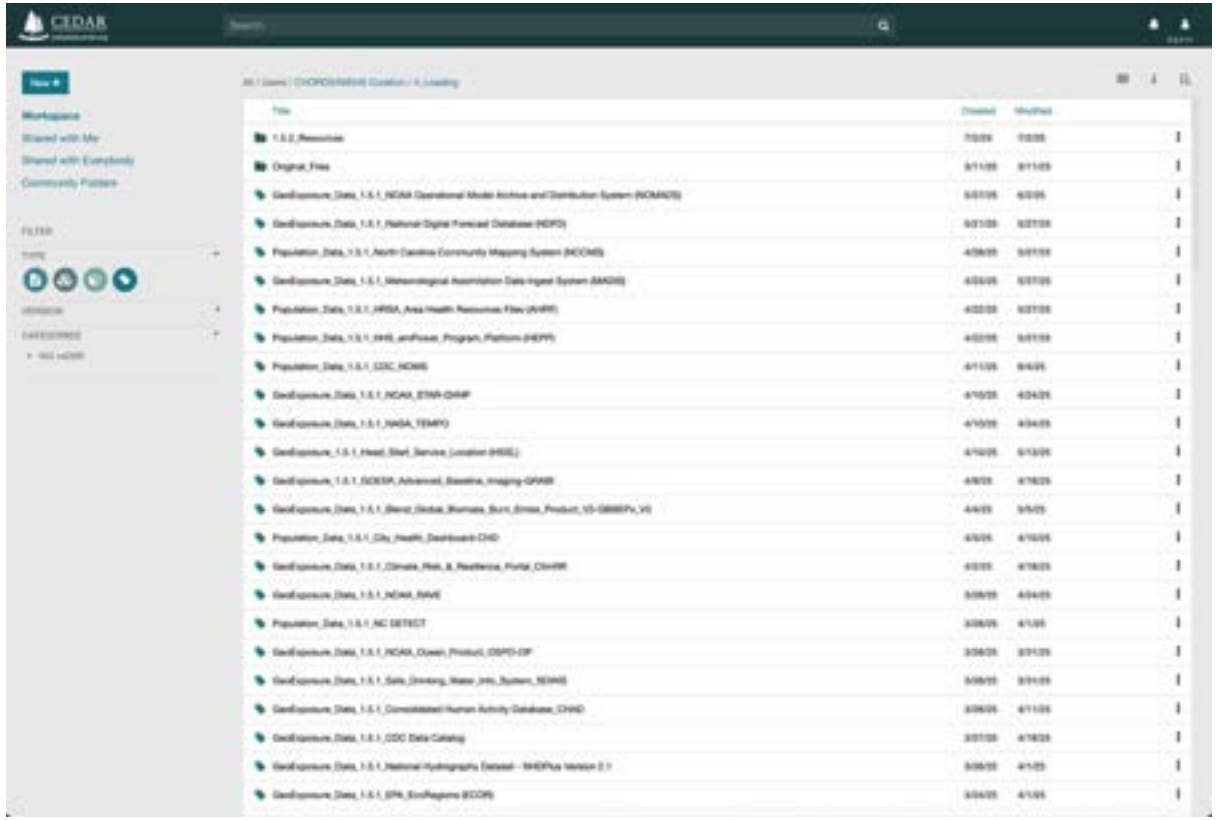
THE UNIVERSITY OF CHICAGO
CENTER FOR TRANSLATIONAL DATA SCIENCE

Data Model Centric

- Above all things, the data model was seen as the key
  - Structured
  - Flexible
  - Validated
  - Includes controlled vocabularies
  - Serializable and API Accessible



Data Curation with CEDAR

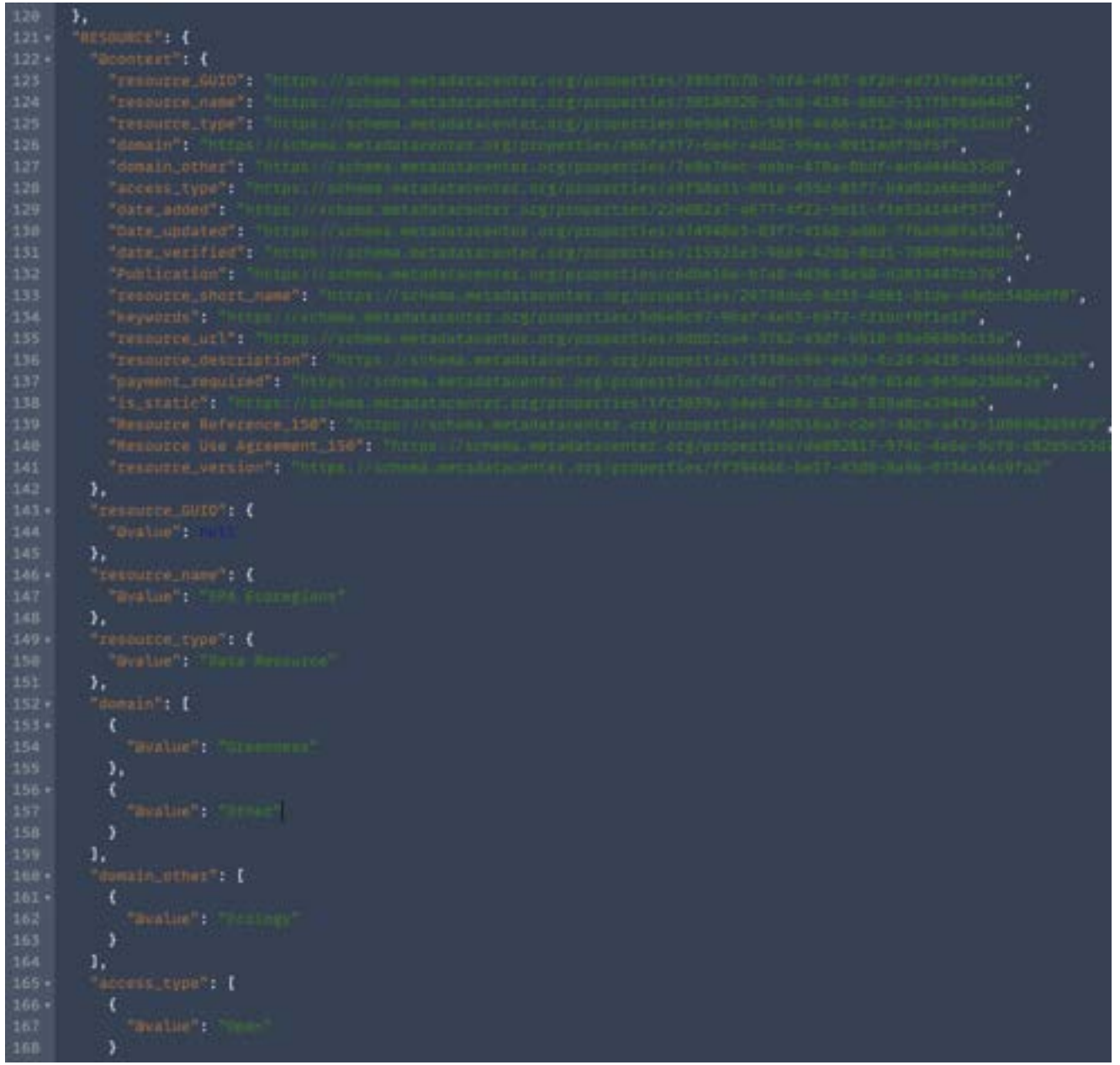


- CEDAR (https://https://cedar.metadatacenter.org/) is being used for human curation of metadata.
- CEDAR templates developed by curators, also serving as the Gen3 model specification.

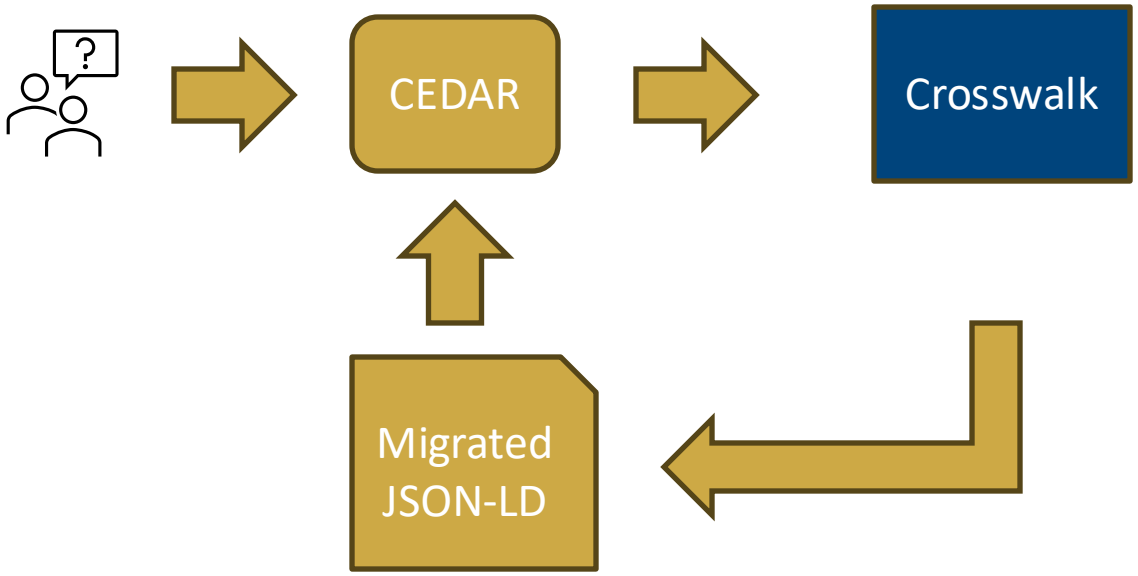
CEDAR
metadatacenter.org

Example CEDAR Template

To JSON-LD



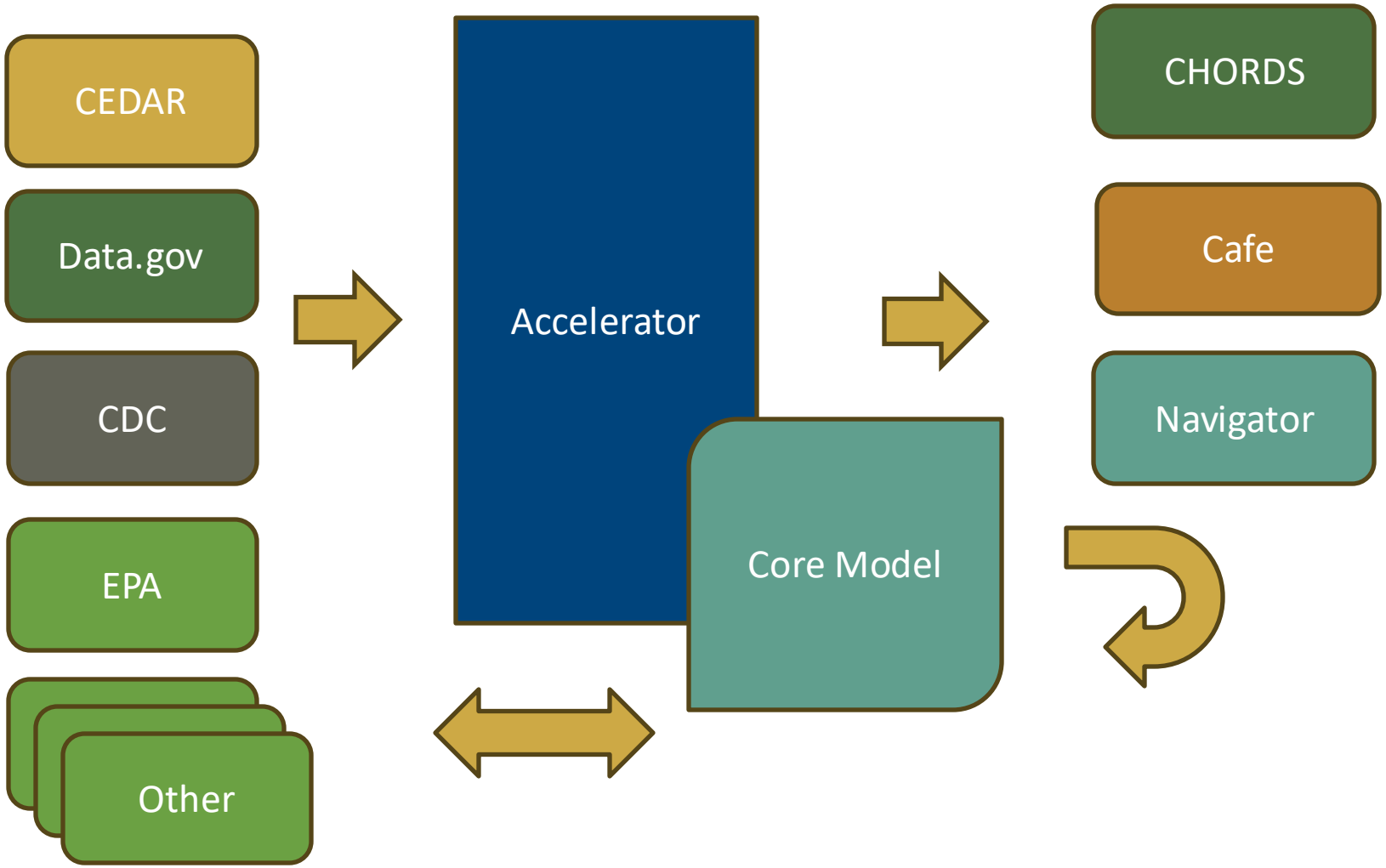
CHORDS Phase 2



Data Models Evolve

- Migration data flows have been created and must be managed
- Note the closed system, without human intervention it cannot be established which data might have disappeared or been updated
- Adding automated input from other sources and pushing data to new endpoints creates all sorts of disconnects, stale data issues

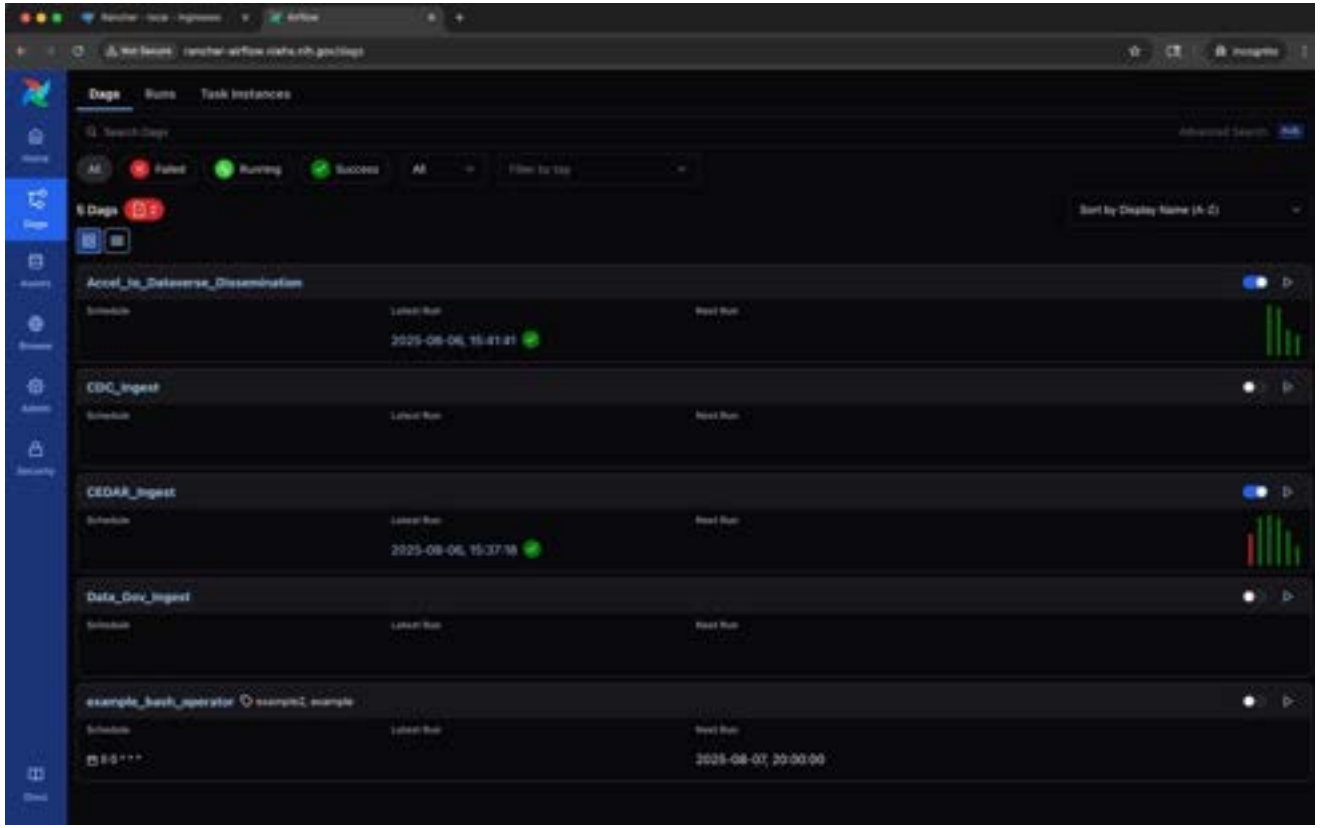
CHORDS Phase 2



- Add automated cataloging
- Create a connector architecture for adding new sources, new sinks, and migration workflows
- Automate re-validation and updates
- Project the CHORDS catalog into multiple endpoints.
- Migration workflows
- Revalidation and discovery workflows

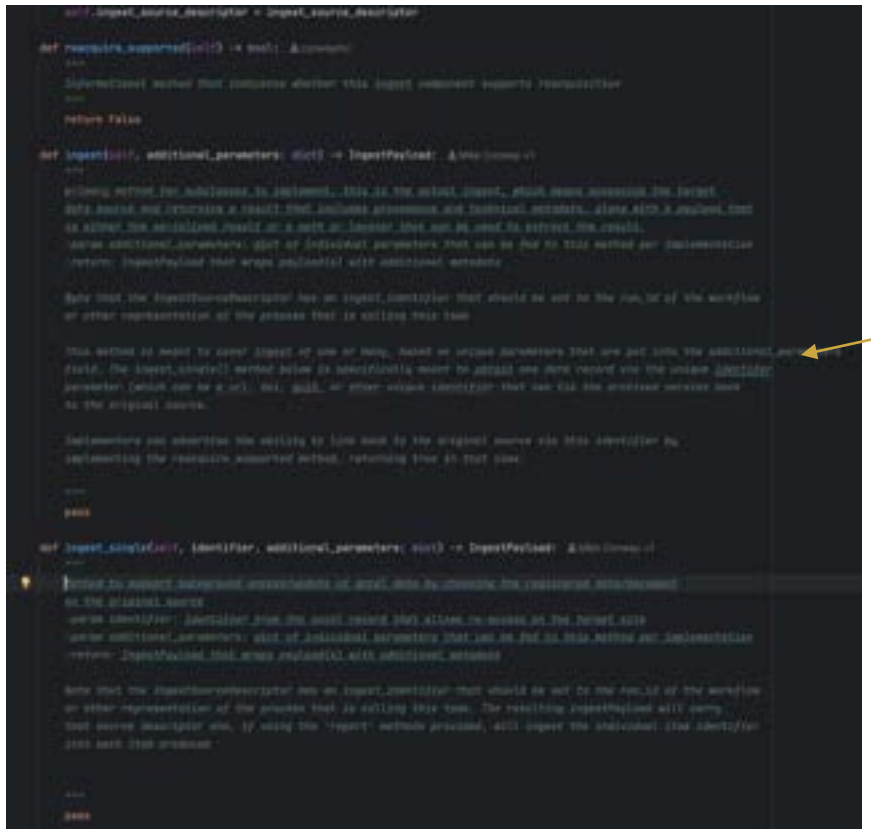
Extending CHORDS Data Federation

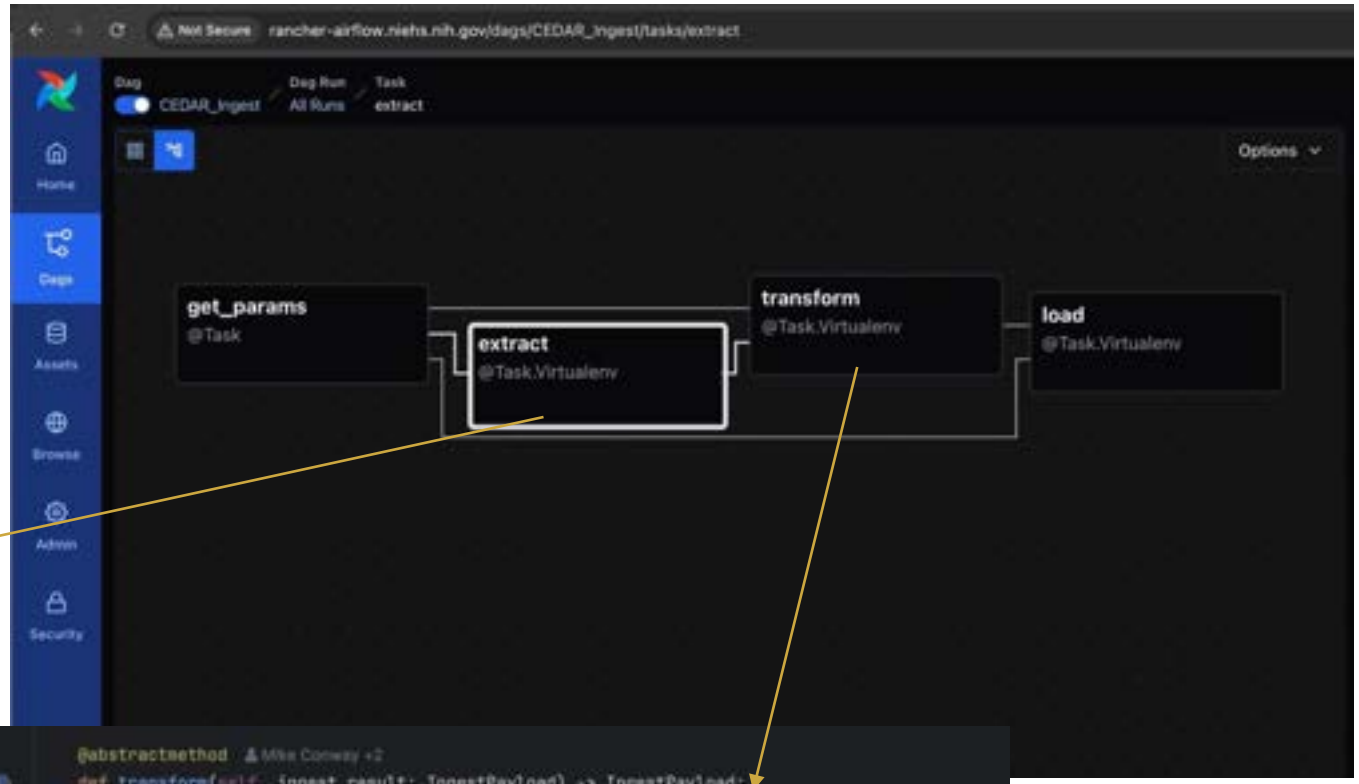
- Think about catalog federation as a commodity ETL problem
- Provide plugin frameworks for adding new sources, models, crosswalks and data sinks, easing on-boarding
- Periodic, sensor-based, and curator triggered workflows

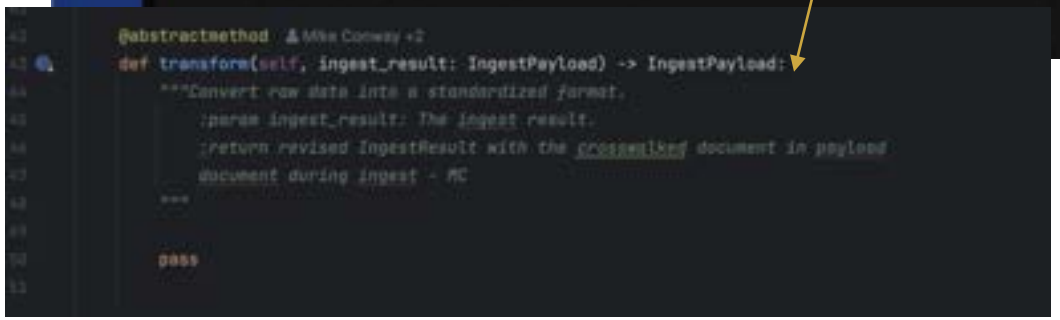


Example Ingest – CEDAR Data

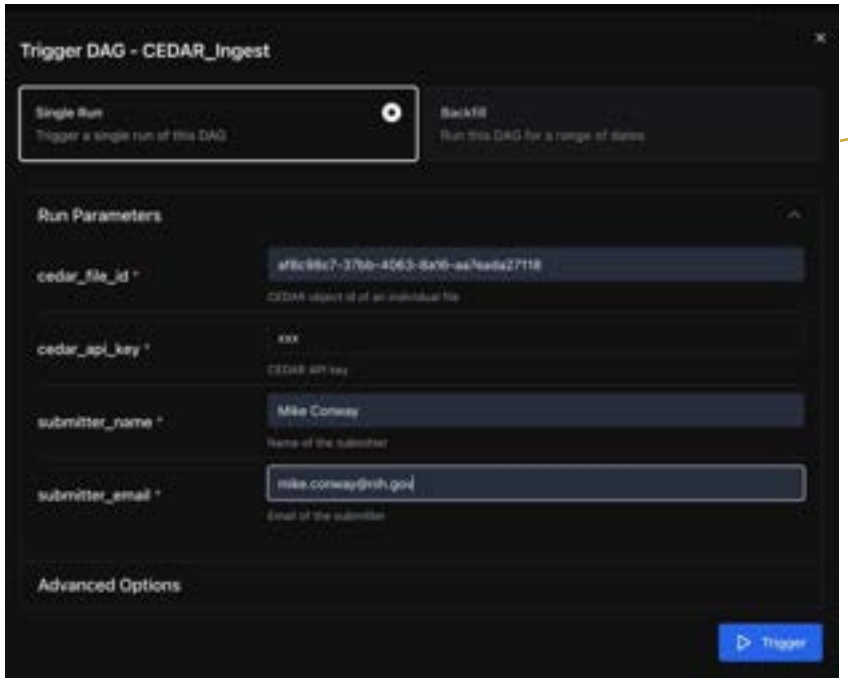
- Extract and Transform are implementations of a superclass. Simple interfaces!

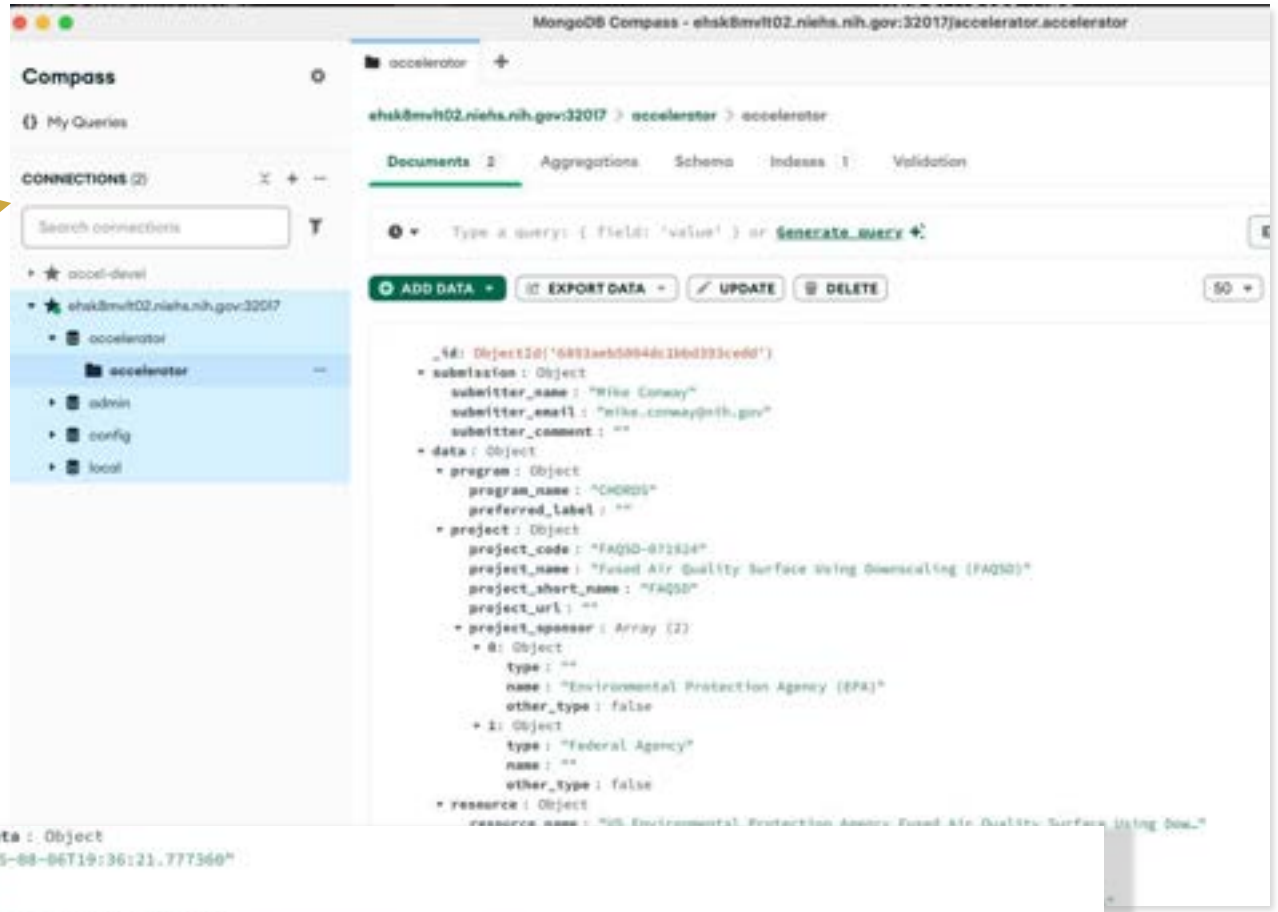






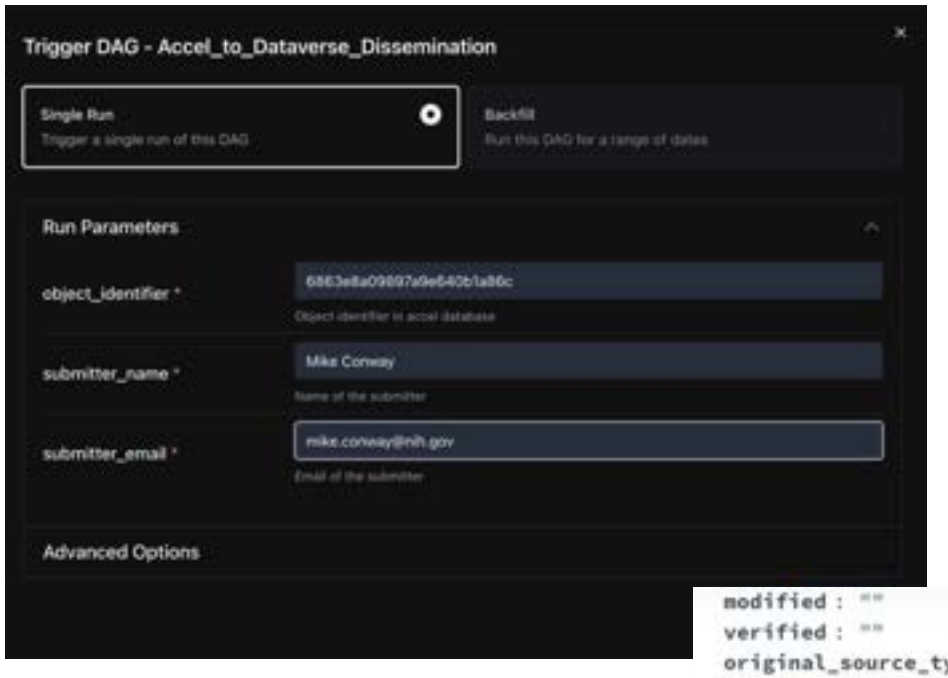
Demo – CEDAR Ingest

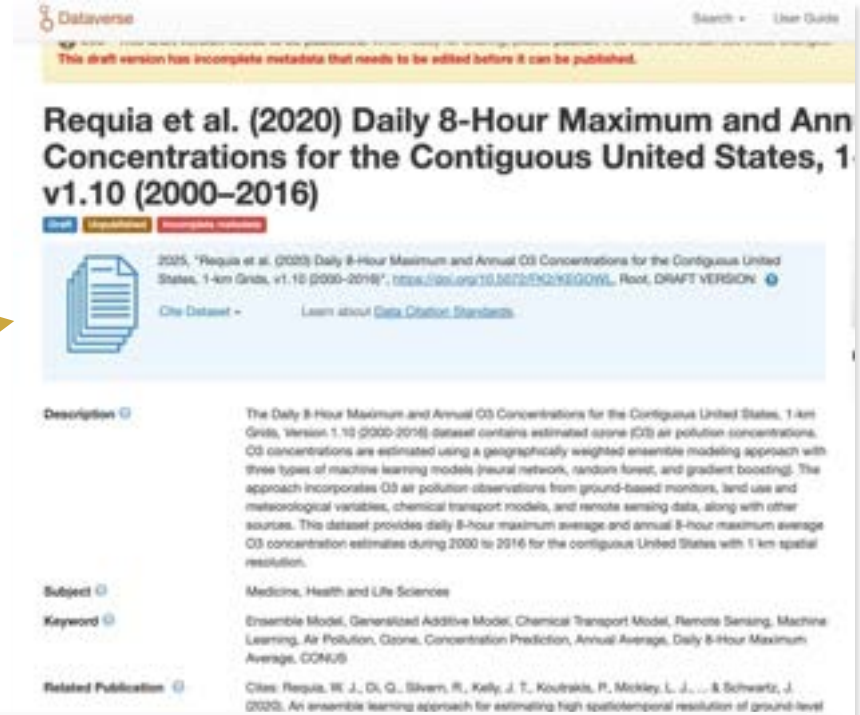




- Common model
- Knows where it came from and when
- Knows the last time it was re-validated

Demo – CHORDS Data Dissemination to CAFE





Lessons Learned

- There are lots of good data sets out there, but **data ready to use for health applications is still a challenge**
- **Machine readable metadata is not always provided**, sometimes it must be derived from web sites or other sources
- There is **spotty availability, especially at the National level**, for things like well water quality, waterway contamination, landfill location
- At the heart of it, CHORDS is about **building a data model that can reasonably catalog a diverse corpus of data**
- As we will discuss later, human curation is difficult, we need to **increase automation**, including constantly updating and re-validating the catalog

Lessons Learned

- In large part, integrating Environmental and Health data is like comparing apples to oranges. Consistent metadata is the central goal
- To better integrate data:
  - Analyte specifics/metrics for chemicals (e.g. CAS #, levels of detection, level of sensitivity)
  - Exposure metrics – average versus peak, etc.
  - Clear time frames and methods of derivation
  - Spatial metrics, how derived, levels of confidence
  - Methods for linking time and spatial metrics to each other and to health data sets

Summary

- CHORDS is about the data model
- CHORDS is about federation
- CHORDS is about human and automated curation
- Focus on open API and frameworks so that catalog data can be shared and re-purposed
- New domains and data types are welcome, and there is a sustainable method for expansion
- Core API: https://github.com/NIEHS/accelerator-core
- Helm Charts: https://github.com/NIEHS/accelerator-helm

Thanks! Get in touch with any questions!

Mike Conway
NIEHS/NIH
mike.conway@nih.gov

Deep Patel
NIEHS/NIH
deep.patel@nih.gov

National Institute of Environmental Health Sciences
Your Environment. Your Health.

BUSPH-HSPH-CAFÉ

Research Coordinating Center (U24)

ISEE-ISES 2025

BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Goals of CAFÉ

Convene	• Provide platforms, opportunities, and occasions to bring the COP together.
Accelerate	• identify research data needs and define common data elements; • develop and promote software tools for processing, linking, and analyzing data; • facilitate data sharing/reuse; • provide data management guidance and cloud computing infrastructure where most needed.
Foster	• facilitate communication, engagement, and sharing of expertise; • stimulate co-production of knowledge; • promote translation of scientific advances into changes in policy and practice.
Expand	• attract a diverse community of scientists via inclusive education, training, and mentoring; • provide pilot project funding; • foster a multidisciplinary, global COP.

BUSPH-HSPH Climate Change and Health Research Coordinating Center

Greg Wellenius (BU)

Francesca Dominici (Harvard)

Amruta Nori-Sarma (Harvard)

Danielle Braun (Harvard)

Gaurab Basu (Harvard)

Rebecca Pearl-Martinez (BU)

Benjamin Sovacool (BU)

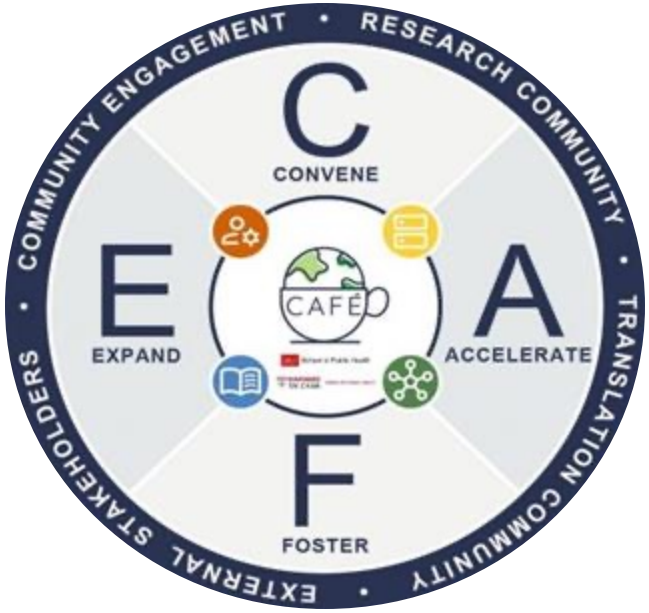
Kevin Lane (BU)

Pat Kinney (BU)

Rachel Nethery (Harvard)

Jon Levy (BU)

Pam Templer (BU)



Administrative Function

Data Management Function

Resource Function

Capacity Building Function

COMMITTED SUPPORTERS:

Harvard Data Science Institute, Harvard Climate Change Institute, BUSPH Center for Climate and Health, Microsoft, AWS, ESRI, Google, BU Events Planning, UMass Boston, Meharry Medical College, PHFI, C40 Cities, EHRA

What does CAFÉ provide?

Various resources, including:

- CAFÉ University webinars
- Free annual virtual meetings
- Funding for pilot studies
- Research translation lab
- Research Mentore-Mentee Connections
- Youtube tutorials
- Slack channels
- Monthly newsletter
- **Data and data management resources**

https://climatehealthcafe.org/

Find us on Linkedin at Climate & Health CAFÉ

Join our Community of Practice to receive our monthly newsletter and Slack channels!

BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

CAFÉ Data Management

1. Facilitate sharing and reuse of Climate Health (CCH) data (**Dataverse**)
2. Develop and promote data science and software tools for processing, linking, and analyzing common CCH data (**Github**)
3. Provide data management and dissemination guidance
4. Identify CCH research data needs and define common data elements across the community of practice (COP)

Note: CAFE is not a data processing/analysis service center! Our focus is on curating commonly used datasets and coding pipelines.

BU School of Public Health Center for Climate & Health
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

CAFE

CAFÉ: Data Repositories (Dataverse)

BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Goal: FAIR Data Principles

Source: Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

CAFE

Researchers

- Structure data clearly and apply good data management practices
- Document data and software
- Richly describe data using standardized metadata fields
- Apply license and/or clear terms of use

Repositories

- Assign persistent identifiers
- Structure metadata records according to a disciplinary standard or schema
- Index data as searchable resources
- Retrieve datasets according to an open protocol that supports authentication
- Preserve data files and metadata
- Track provenance and versioning

NIH Data Management Sharing Plans

CAFE

- Outlines how data resulting from a research project is managed in the short and long term
- Procedures for data collection, documentation, and processing
- Plan for data sharing so others can find and access materials in perpetuity
- Living document that is frequently referred to and updated as needed
- Many funders now require data sharing or data management plans as part of the conditions of funding

Source: Elements of an NIH Data Management and Sharing Plan

NIH Data Management and Sharing Policy

Submission of a Data Management and Sharing Plan (DMSP)

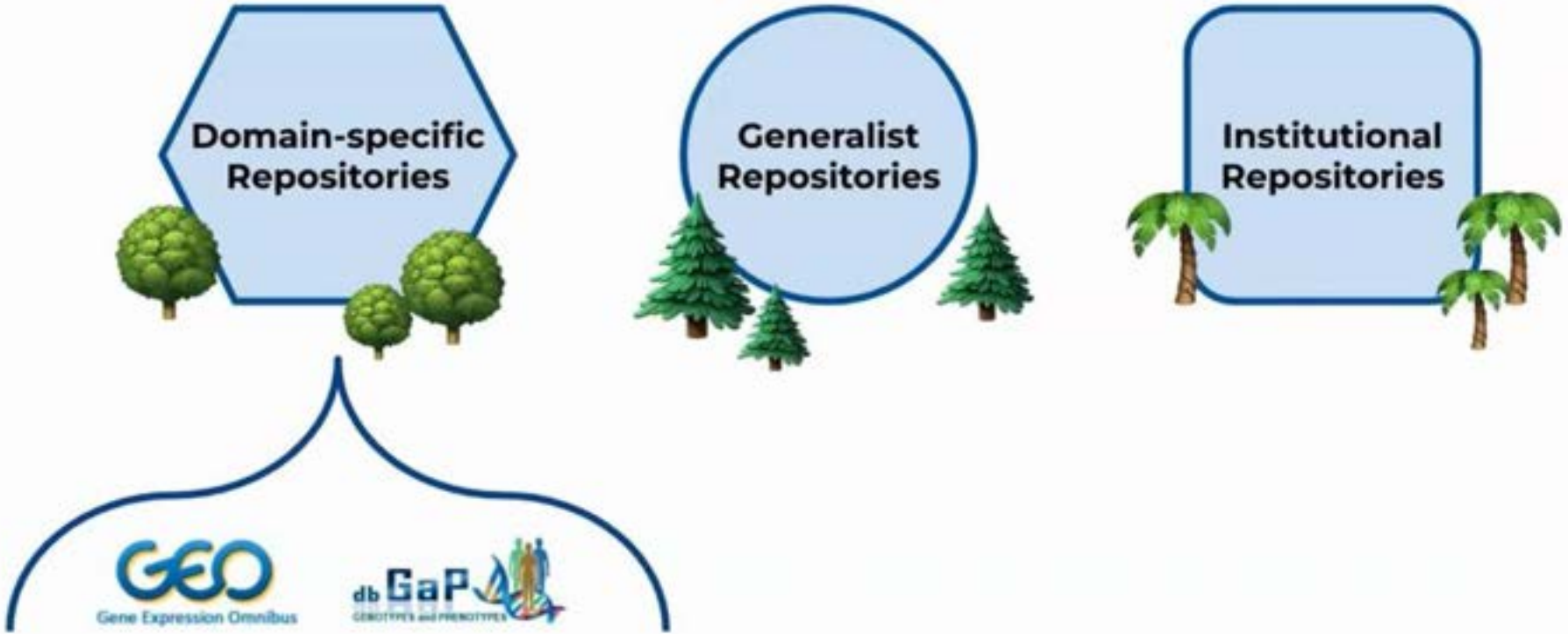
Proposed repository to be used consistent with NIH guidance

Compliance with the awardee's plan as approved by NIH ICO

Source: Data Management & Sharing Policy Overview

Data Repository Ecosystem

CAFE

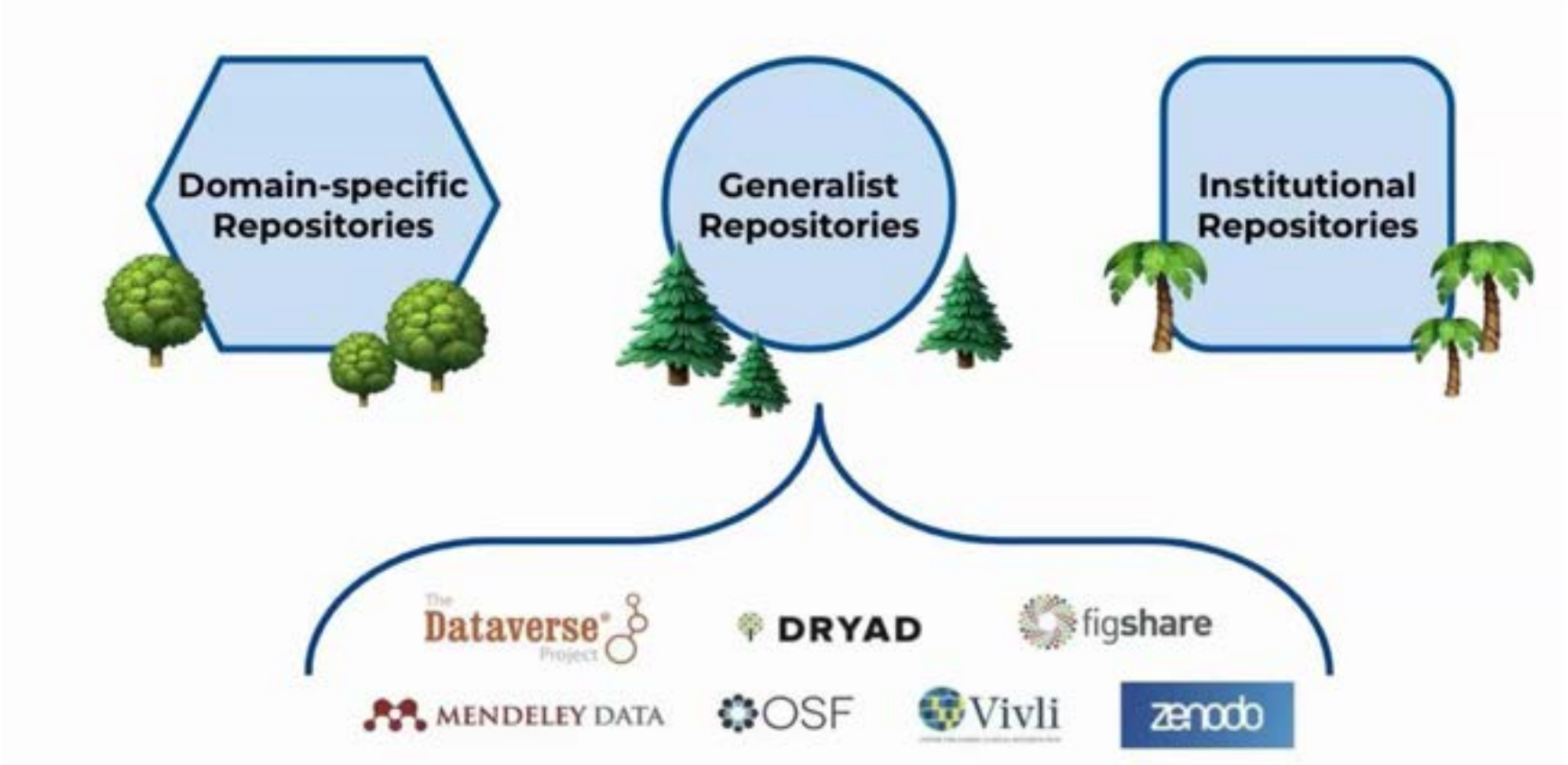


BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Data Repository Ecosystem

CAFE

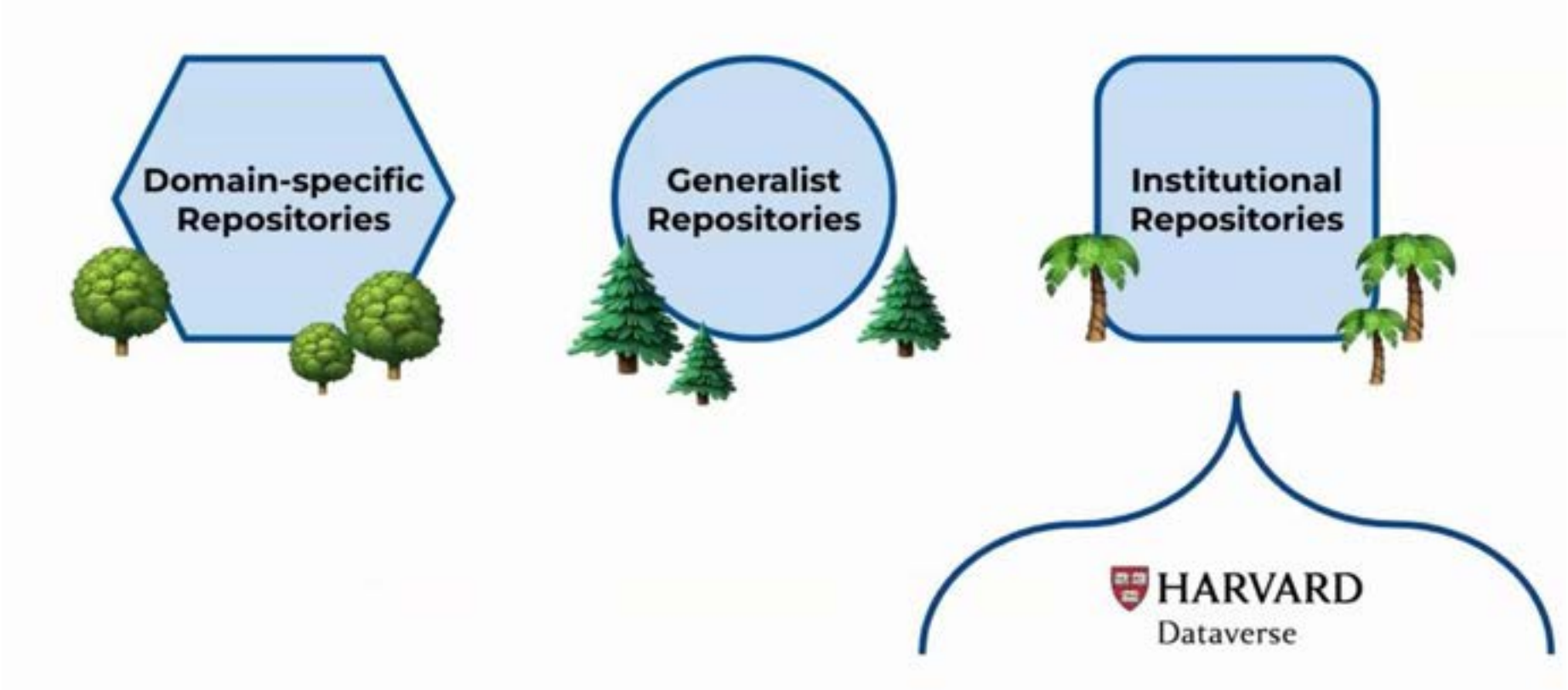


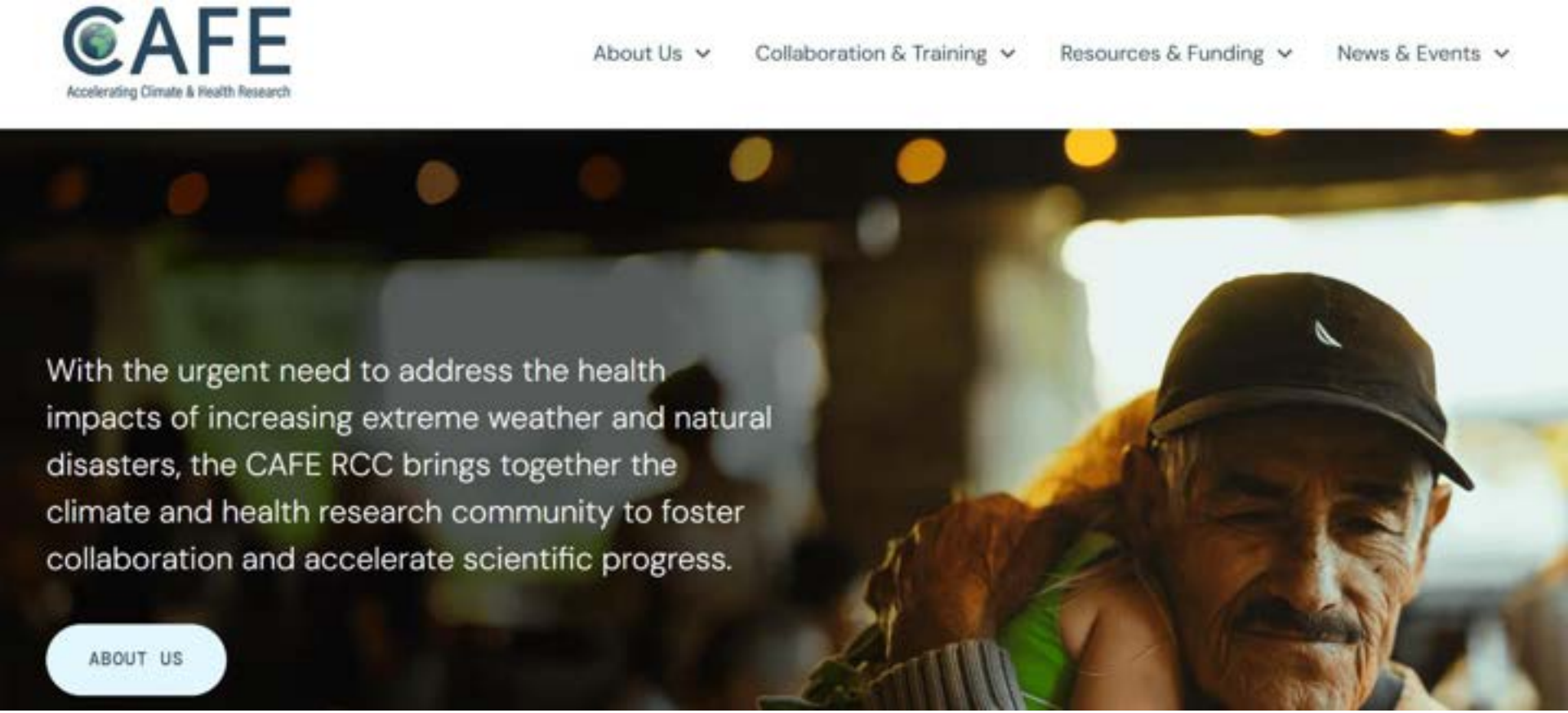
BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Data Repository Ecosystem

CAFE





https://www.climatehealthcafe.org/

BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

CAFÉ Dataverse Collection Highlights

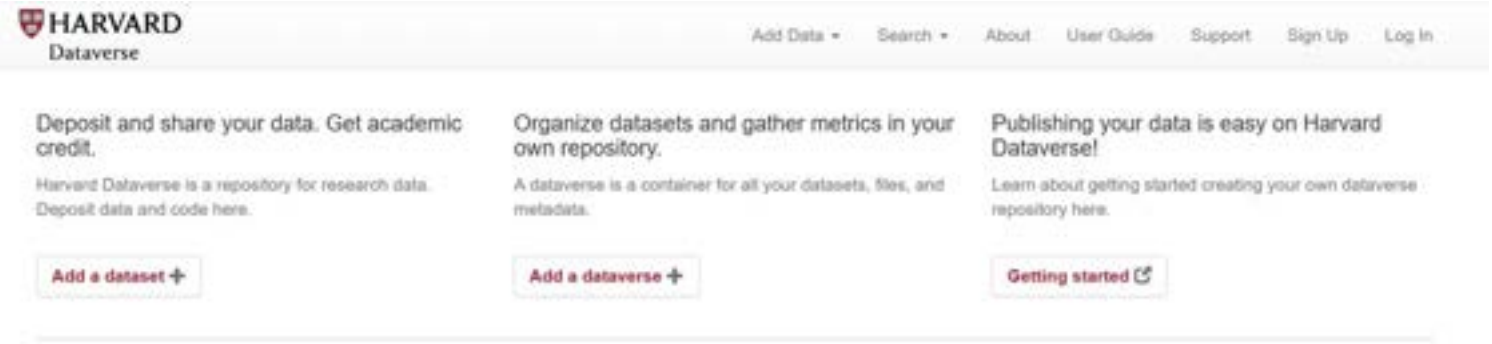
- Over **700 Data submissions** deposited in <2 years.
- Most datasets have been **downloaded between 50 - 5000 times (and increasing).**
- Most downloaded datasets include wildfire air pollution, synthetic health data, spatial aggregations and greenspace.
- New **CIESIN datasets** deposited August 2025.
- **Sub-collections**, which are a nice way for groups with multiple datasets to share data within the CAFÉ Collection in one place.

BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Dataverse Extracting Data

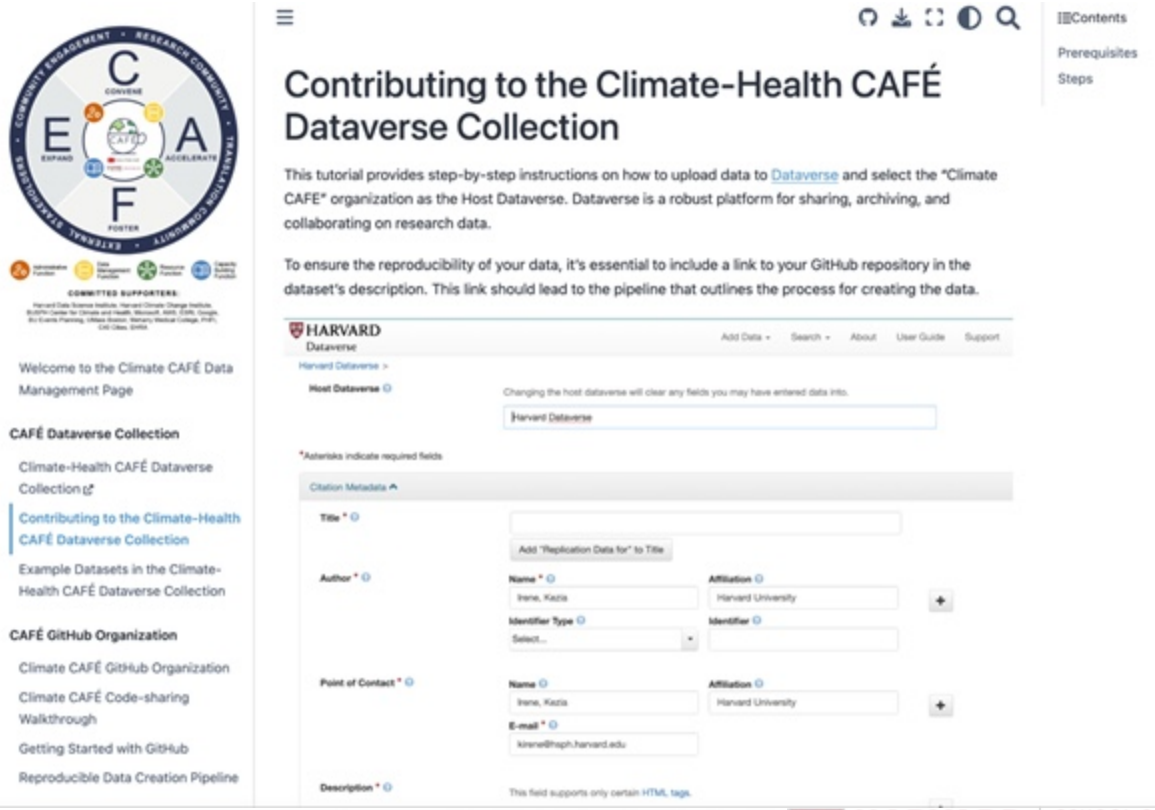
Why Create/Use a Dataverse Repository



- Can share data with collaborators or the public. Open content can be accessed directly via the UI or API, and restricted content can be requested using a "request access" feature if enabled by the data depositor.
- Assigns a DOI to every published dataset. The repository records file downloads and views and makes this information available to depositors. Depositors who create collections can also ask or require downloaders to provide information about their data re-use.
- Uses standard-compliant metadata to ensure that dataset metadata can be mapped easily to common metadata schemas to make data more preservable and interoperable.
- Provides a mechanism by which a journal's editors and reviewers can have anonymous access to a dataset or dataverse before it is made public. See Private URLs

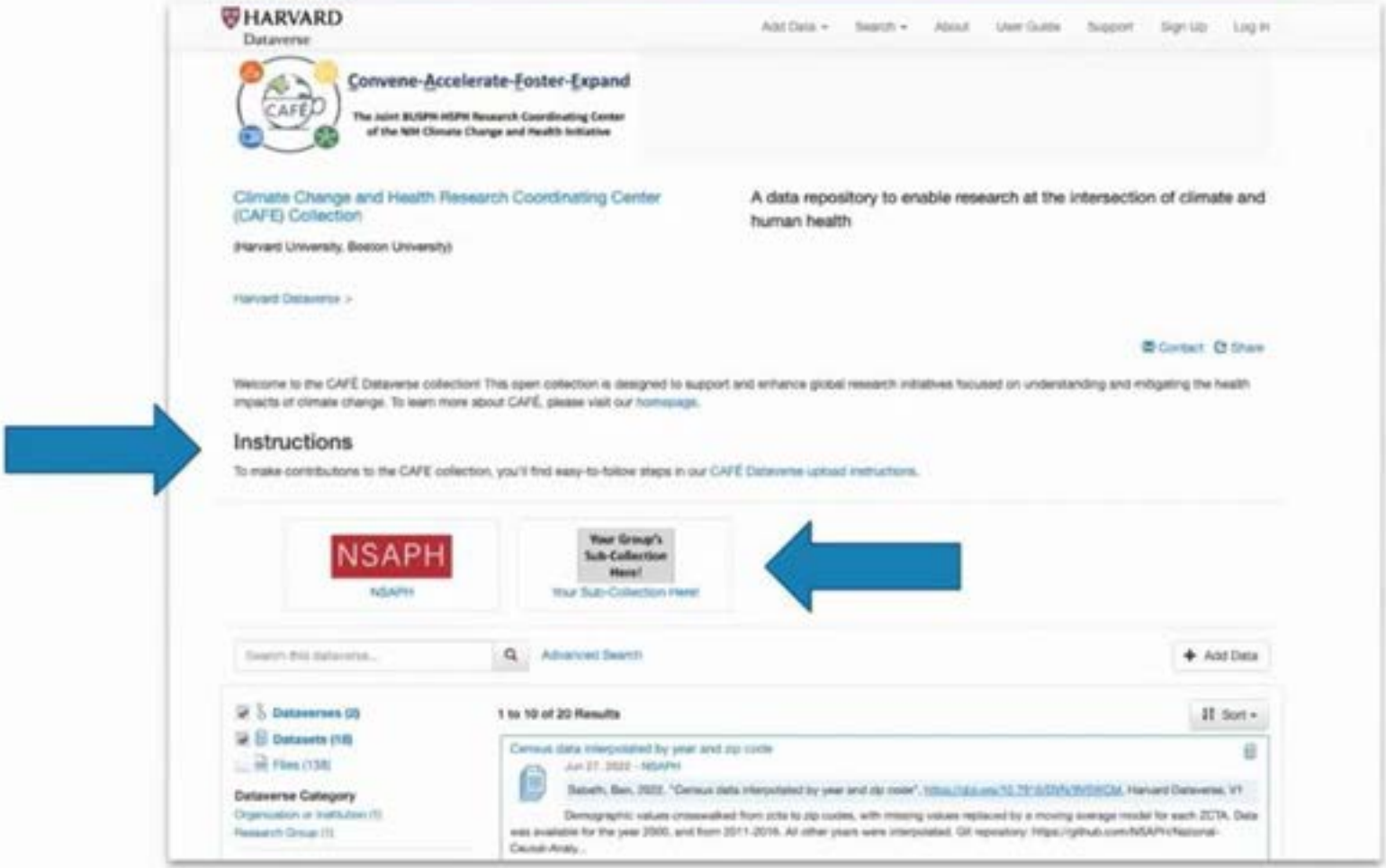
BU School of Public Health Center for Climate & Health
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

CAFÉ Dataverse Collection: Contributing Data

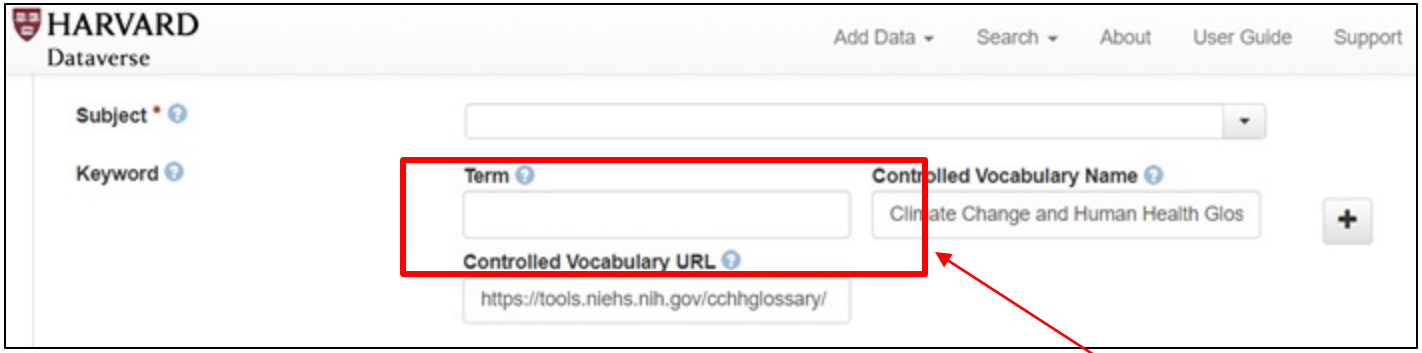


https://climate-cafe.github.io/tutorial.html

https://dataverse.harvard.edu/dataverse/CAFE



Developing your CAFÉ Dataverse Collection: Custom Metadata



NIEHS Glossary Keyword Term

https://www.niehs.nih.gov/health/topics/glossary

BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Benefits of Dataverse to Help You Meet FAIR Data Principles

CAFE

🔍 Findable
- Each dataset receives a **persistent DOI**, making it uniquely identifiable and citable.
- Rich **metadata standards** enhance searchability and indexing in global registries (e.g., DataCite, Google Dataset Search, National Library of Medicine).

📥 Accessible
- Supports **open access to data** (with options for restricted access when needed), satisfying NIH's emphasis on making data broadly available.
- Provides a stable, institutionally supported platform with **long-term data preservation**.

🔄 Interoperable
- Encourages use of **standard data formats** (e.g., CSV, JSON, Stata, SPSS) that can be reused across systems and disciplines.
- Metadata and APIs are compatible with other repositories and tools, supporting machine-readability and integration.

♻️ Reusable
- Users can attach **detailed documentation**, codebooks, and README files to promote data reuse and understanding.
- Enables application of **clear data licenses** (e.g., CC0, CC BY), fulfilling legal and ethical reuse criteria

Harvard Dataverse NIH-DMP Guidance

https://support.dataverse.harvard.edu/harvard-dataverse-nih-dmp-guidance

CAFÉ Toolkit Examples: GitHub

https://climate-cafe.github.io//intro.html

BU School of Public Health Center for Climate & Health
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Goal: FAIR Data Principles

Source: Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

Benefits of using GitHub for code sharing to meet FAIR Data Principles

CAFE

🔍 **Findable**

- GitHub repositories are indexed by search engines and can be assigned **persistent identifiers (e.g., DOIs**), making code discoverable.
- Clear project structure and metadata (e.g., README, tags, releases) improve searchability and navigation.

📥 **Accessible**

- Code is **openly accessible by default** (when using public repos), satisfying NIH expectations for open data/code sharing.
- GitHub supports easy **download, cloning, and API access** to repositories.

🔄 **Interoperable**

- GitHub supports **standard code formats** and integrates with popular tools and languages (e.g., Python, R, Jupyter, Git).
- Enables collaboration and integration with reproducibility platforms (e.g., Binder, CodeOcean).

♻️ **Reusable**

- Clear **version control and documentation** support reproducibility and reuse of code by other researchers.
- Supports **open-source licenses** (e.g., MIT, Apache), making reuse legally and ethically straightforward.
- Promotes **open science** through forks, pull requests, and issues, supporting the collaborative spirit of NIH's open data initiatives.

Supports NIH DMS Plan Compliance

- Facilitates sharing of **software, scripts, and workflows** alongside datasets, as required by NIH for reproducibility.
- Documentation (e.g., README, CONTRIBUTING.md) can describe data provenance and usage instructions.

Integration with Archival Services

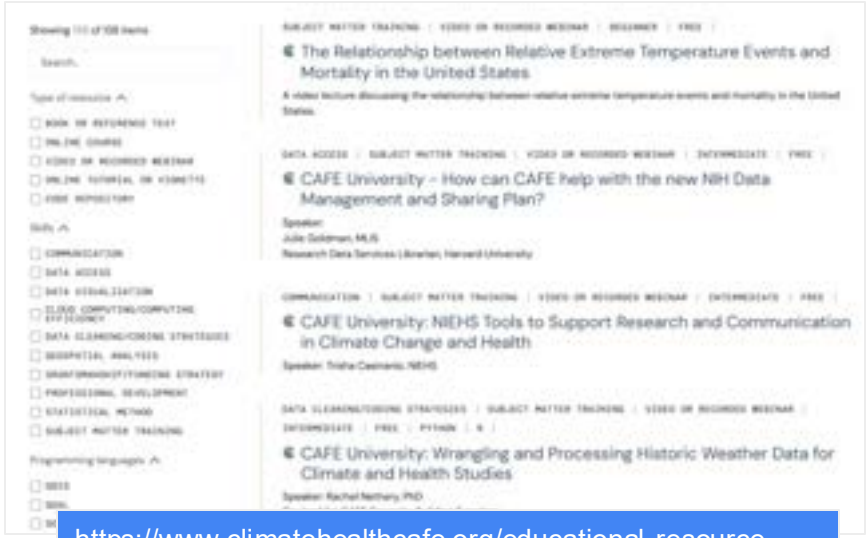
- GitHub integrates with **Zenodo** and other repositories to generate **citable DOIs**, ensuring long-term preservation and citation, aligning with FAIR's "Findable" and "Reusable" goals.

Enhances Collaboration and Community Engagement

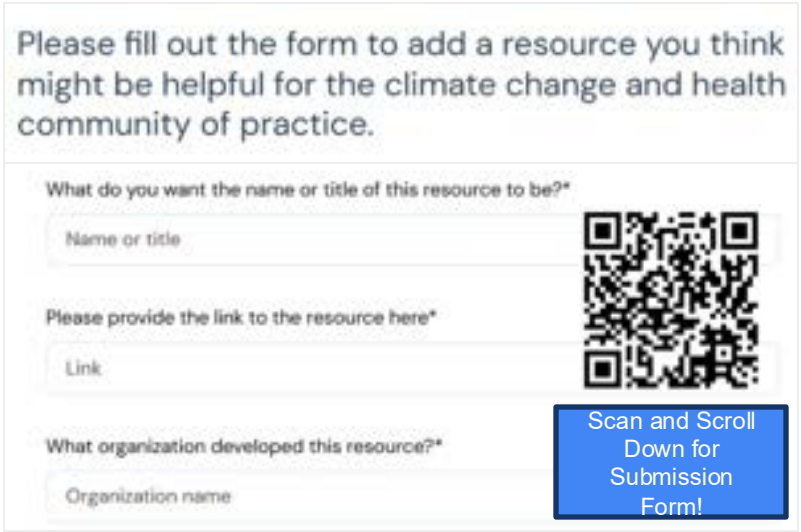
- Promotes **open science** through forks, pull requests, and issues, supporting the collaborative spirit of NIH's open data initiatives.

CAFÉ Educational Resource Hub

- Crowd-sourced, searchable catalog to centralize information across experience levels, areas of interest, programming languages, etc.
- Includes coding tips, epidemiological methods, coding tutorials, and more!



https://www.climatehealthcafe.org/educational-resource-hub



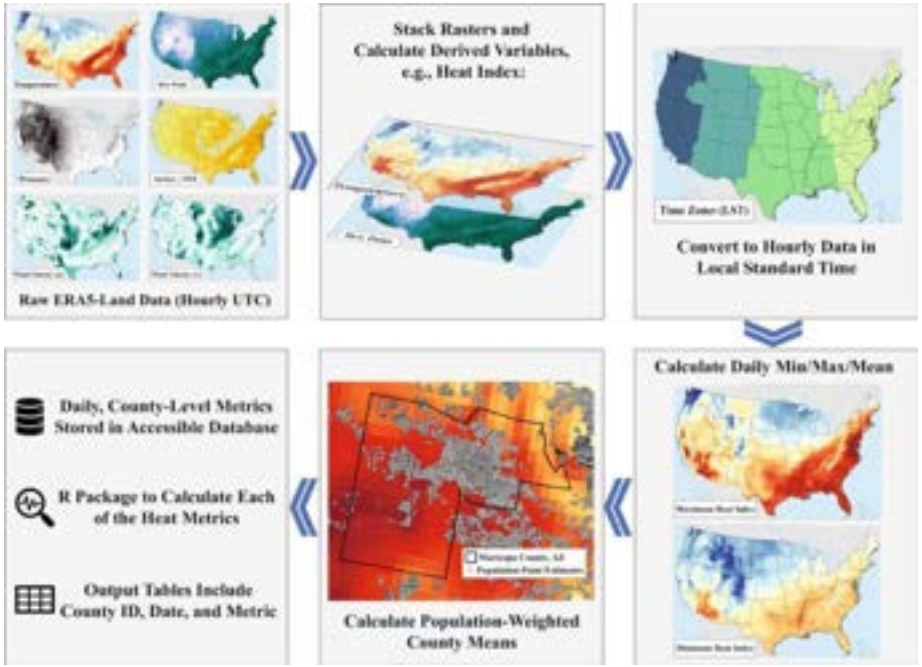
BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Wet-Bulb Globe Temperature, Universal Thermal Climate Index, and Other Heat Metrics for US Counties, 2000–2020

Keith R Spangler [1,✉], Shixin Liang [1,2], Gregory A Wellenius [1]

- ERA-5 data is available globally with a historical record.
- CAFÉ-collection has an example for working with US and Global data examples.
- Github has R and Python versions that combine multiple existing packages for processing and merging to data.



BU School of Public Health Center for Climate & Health
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

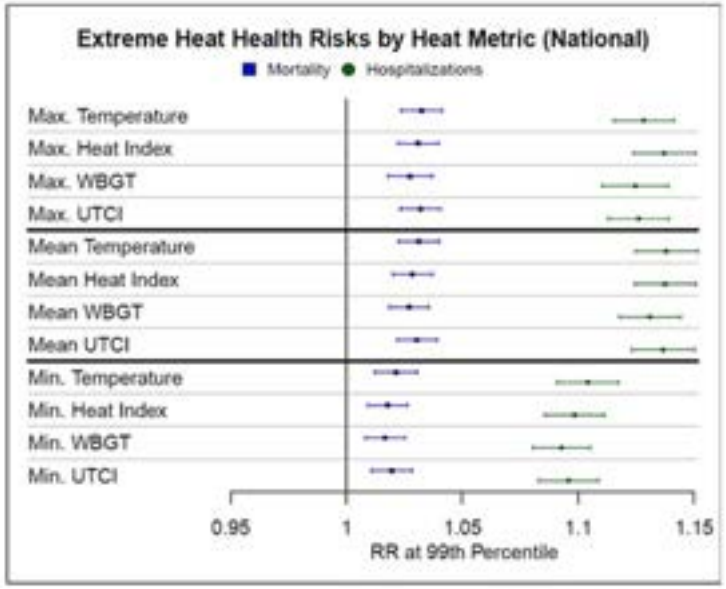
Original Research Article

ISEE ENVIRONMENTAL EPIDEMIOLOGY

OPEN

Does choice of outdoor heat metric affect heat-related epidemiologic analyses in the US Medicare population?

Keith R. Spangler[a,*], Quinn H. Adams[a], Jie Kate Hu[b], Danielle Braun[b,c], Kate R. Weinberger[d], Francesca Dominici[b], Gregory A. Wellenius[a]



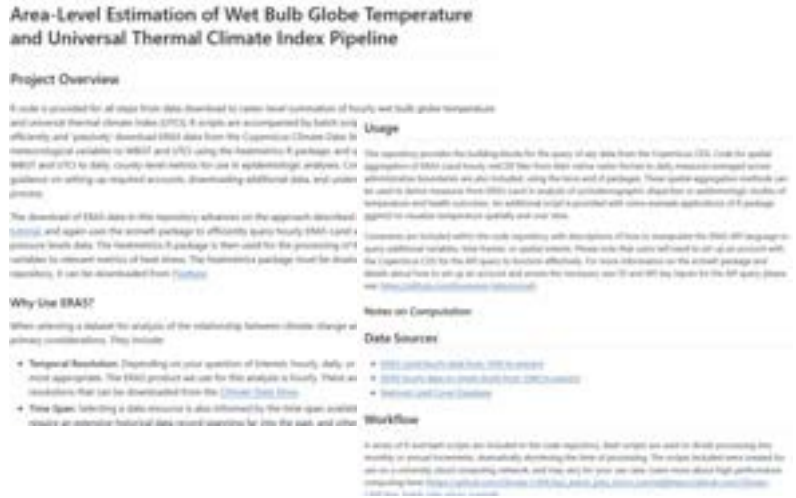
- This example has been used in health research both within the United States and Globally.

BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

Example: WBGT and UTCI ERA5 Aggregation Tutorial

- Includes template scripts for expedited batch ERA5 Download
- Applies **Spangler et al. heatmetrics R package** for step-by-step wet bulb globe and universal thermal climate index estimation
- Hourly raster to daily county aggregation, with time zone adjustment


Area-Level Estimation of Wet Bulb Globe Temperature and Universal Thermal Climate Index Pipeline

Project Overview

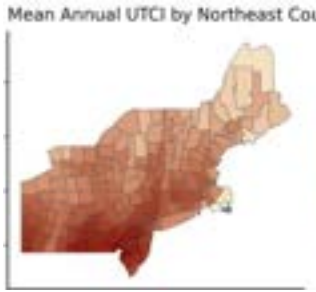
Usage

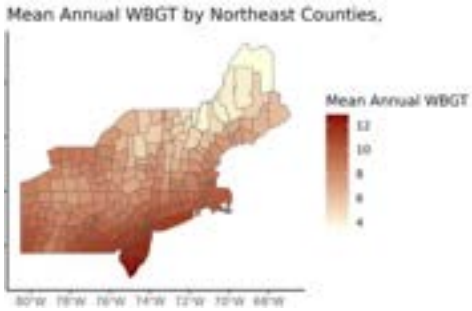
Why Use ERA5?

Notes on Computation

Data Sources

Workflow





BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

Required Input Data

- The GitHub will walk through how to use the Copernicus API to download ERA5-Land and ERA5
  - You will need to create a Copernicus account prior to use!
- There are several supporting datasets required for use of the pipeline as well
  - Find links and guidance directly in the GitHub scripts

- **ECMWF Account Creation and API Extraction**

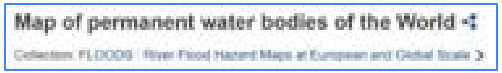
  See ecmwfr package for details!

- **National Land Cover Database**

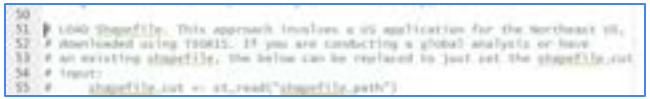
  Used to estimate urbanicity

- **JRC Permanent Water**

  Used to mask inland water bodies



- **Your Shapefile of Interest**



- **Heatmetrics R package from Figshare**

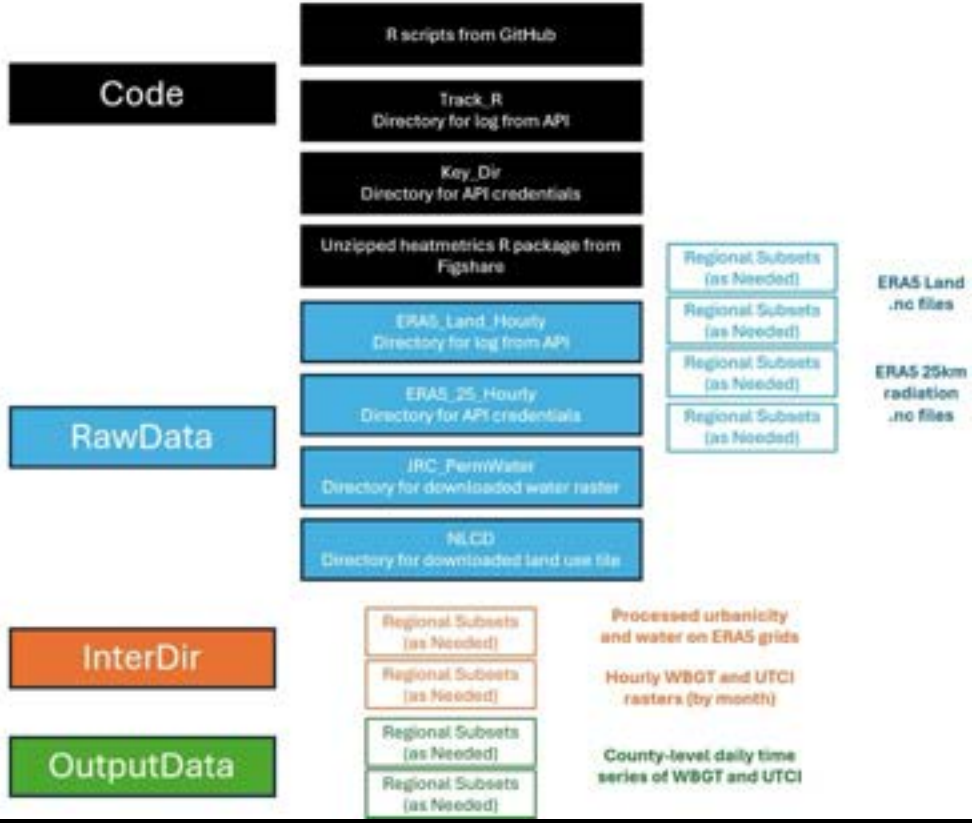
  .tar.gz file with all utci and wbgt functions

Setting Up Your Environment

- Ensuring your directory structure is created to ingest, organize, and access data is crucial at the outset
- Each R script lists where to enter your file paths at the top under 'User-Defined Parameters'

```
############################ User-Defined Parameters ##############################
# Set up directories to read in and output data
#
era_dir <- "RawData/ERA5_Hourly/"     # ERA5-Land Rasters
era_interdir <- "InterDir/"           # Intermediate where static rasters output
nlcd_dir <- "RawData/NLCD/"           # Where NLCD data is stored
water_dir <- "RawData/JRC_Permwater/" # Where JRC Permanent water is stored
```

Recommended Structure



Computational Requirements and Workarounds

- WBGT and UTCI processing involves hourly data for 8+ meteorological variables and is time and compute-intensive

- The GitHub repository includes batch scripts to divide the processing and run different regions and time frames in parallel

Unfamiliar with Batch Scripting?

See the CAFÉ microtutorial!

Climate-CAFE/hpc_batch_jobs_micro_tutorial

Example Script for Batch Submission to Subset Across Time and Space

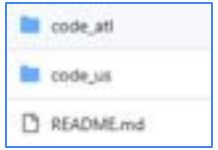
BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF Department of Biostatistics

Computational Requirements and Workarounds

- Batch scripting is an efficient way to separate large spatiotemporal processing but may not be available in all settings
- If you are processing a single county and/or a short time series (less than a year), an example processing pipeline can be found under code_atl
- Loops are used to process across months

```
# Set county This code is developed for a single US county. These could be updated to reflect
# any subdivision of your area of interest, as needed to divide processing into
# more computationally efficient steps.
#
county_fn <- 13121 # Example county is Fulton County, GA
```



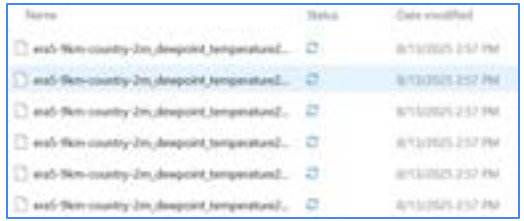
Scripts 1+2: Downloading ERA5 Data

- Smaller requests move more quickly through Copernicus download queue
- Using batch submission to submit requests will expedite process
  - To further speed up this process, the API is directed to save files to a directory that has not been created on your device
  - After all requests have been processed through the queue, the directory for saving files is created.
  - Log language from the API will be saved to a file on your device so the files can be programmatically downloaded to the newly created directory.

Example Log Text which can be used to download files from queue

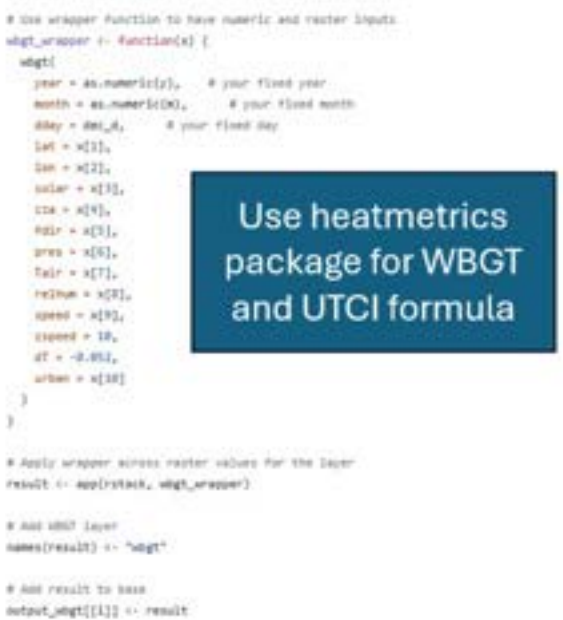
Scripts 1+2: Downloading ERA5 Data

- View the status of your requests by visiting https://cds.climate.copernicus.eu/requests
- Time from queue to completion will vary based on how many users are accessing data
- Once all requests are Complete, run script 1C to download to your device
- Downloading all variables typically takes several hours per year



Scripts 3+4: Calculating Hourly WBGT + UTCI

- Script 3 creates 'static' intermediate raster layers for percent urbanicity, percent water, etc. that are not time-varying
- Script 4 runs for every hour you are processing. It is time and computationally intensive.
  - Set up is for batch scripting to divide by month and region

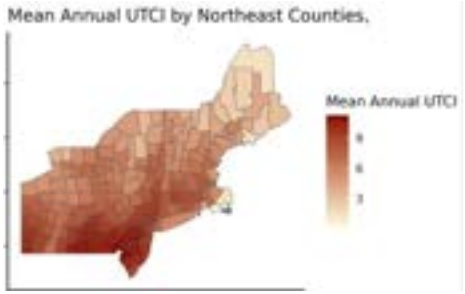


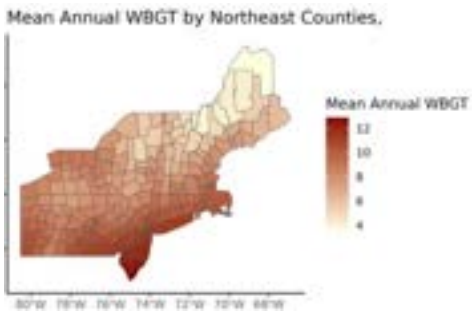
Scripts 5+6: Temporal and Spatial Aggregation

- Script 5 combines the ERA5 raster grid with the input polygon. This can then be used to extract the hourly raster values to the county

- Script 6 includes the daily and county-level aggregation, to create the final outputs for use

```
93  # %%%%%%%%%%%%%%%%%%%%%% CREATE A FISHNET GRID OF THE RASTER EXTENT %%%%%%%%%%%%%% #
94  #
95  # Here, we are making a shapefile that is a fishnet grid of the raster extent.
96  # It will essentially be a polygon of lines surrounding each ERA5 cell.
97  #
98  # Reference/credit: https://gis.stackexchange.com/a/243585
99  #
100 era_raster <- era_stack[[1]]
101 era_extent <- ext(era_raster)
102 xmin <- era_extent[1]
103 xmax <- era_extent[2]
104 ymin <- era_extent[3]
105 ymax <- era_extent[4]
```





WBGT and UTCI Next Steps

- Nationwide county products to be posted to CAFÉ Dataverse
- Update package with population weighting to county aggregation
- Enhance quality control procedures to flag values that may be more prone to error due to presence of water bodies, unexpected input values, etc. in line with Spangler et al.

	StCoFIPS	Date	Tmin_C	Tmax_C	Tmean_C	TDmin_C	TDmax_C	TDmean_C	NETmin_C	NETmax_C	NETmean_C	HImin_C	HImax_C	HImean_C	HXmin_C	HXmax_C
1	01001	2000-01-02	14.62	21.63	17.28	14.45	16.52	15.38	8.10	15.61	10.88	14.72	21.54	17.40	18.21	25.38
2	01003	2000-01-02	16.26	22.61	18.80	15.82	18.43	17.18	10.10	15.88	12.35	16.51	22.75	19.12	20.76	27.66
3	01005	2000-01-02	13.88	20.09	16.23	13.80	16.28	14.97	7.29	14.83	10.17	13.92	20.19	16.33	17.08	24.69
4	01007	2000-01-02	13.57	21.02	16.62	13.46	16.68	15.04	6.52	15.05	10.09	13.58	21.02	16.72	16.58	25.22
5	01009	2000-01-02	13.49	18.49	15.29	13.35	15.80	14.52	5.84	12.15	8.19	13.48	18.57	15.36	16.44	22.79
6	01011	2000-01-02	14.25	20.18	16.41	14.09	16.11	15.07	7.80	14.44	10.17	14.31	20.21	16.51	17.62	24.50

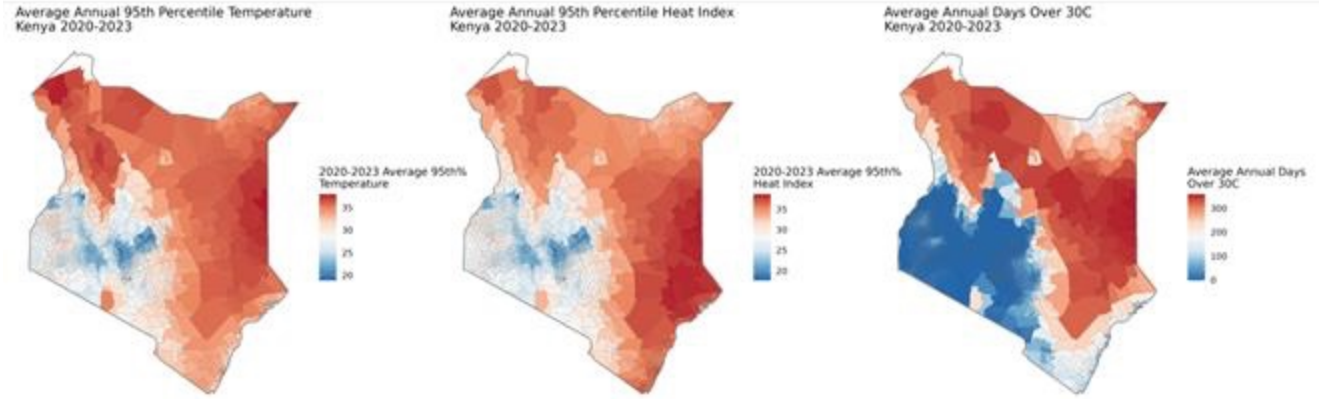
	HXmean_C	WBGTmin_C	WBGTmax_C	WBGTmean_C	UTCImin_C	UTCImax_C	UTCImean_C	Flag_T	Flag_TD	Flag_NET	Flag_HI	Flag_HX	Flag_WBGT	Flag_UTCI
1	21.41	14.21	20.49	16.73	11.98	24.88	16.57	0	0	0	0	0	0	0
2	24.13	15.78	22.33	18.65	13.32	25.46	17.92	0	0	0	0	0	0	0
3	20.12	13.40	19.67	15.88	11.59	23.10	15.76	0	0	0	0	0	0	0
4	20.56	13.17	20.20	16.15	9.71	24.62	15.55	0	0	0	0	0	0	0
5	18.91	13.10	18.63	15.08	9.77	21.11	13.55	0	0	0	0	0	0	0
6	20.36	13.78	19.51	16.02	11.84	22.45	15.75	0	0	0	0	0	0	0

ERA5 Daily Heat Aggregation

https://github.com/Climate-CAFE/era5-daily-heat-aggregation

CAFE

- Use existing R package to demonstrate process for query of ERA5 data with variable, time, and geographic extent inputs
- Across Kenya administrative boundaries, compute daily summary heat measures, including daily max temperature, heat index, and more



BU School of Public Health Center for Climate & Health

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH Department of Biostatistics

OpenStreetMap Road Metric Calculations

- **Objective**: Calculate total road lengths by type (e.g., highways, local streets) for municipalities in Mexico, adaptable to other countries
- **Data**: Uses OpenStreetMap (OSM) roads from Geofabrik, GADM shapefiles for boundaries, and R scripts to calculate and aggregate road sums.
  - Bash scripts included for fast processing.

CAFÉ GitHub: US Census Query

- Example census tutorial on CAFÉ Youtube

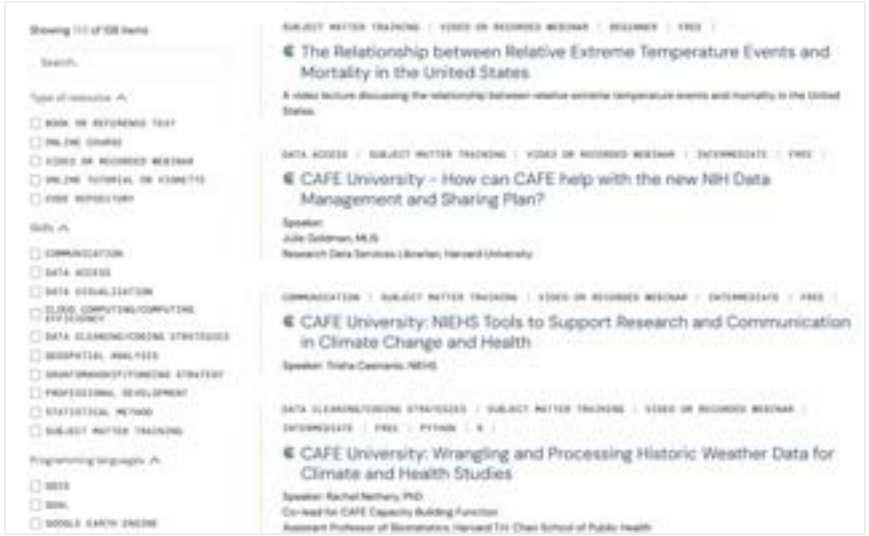
https://www.youtube.com/watch?v=zC5nA3AVMpE

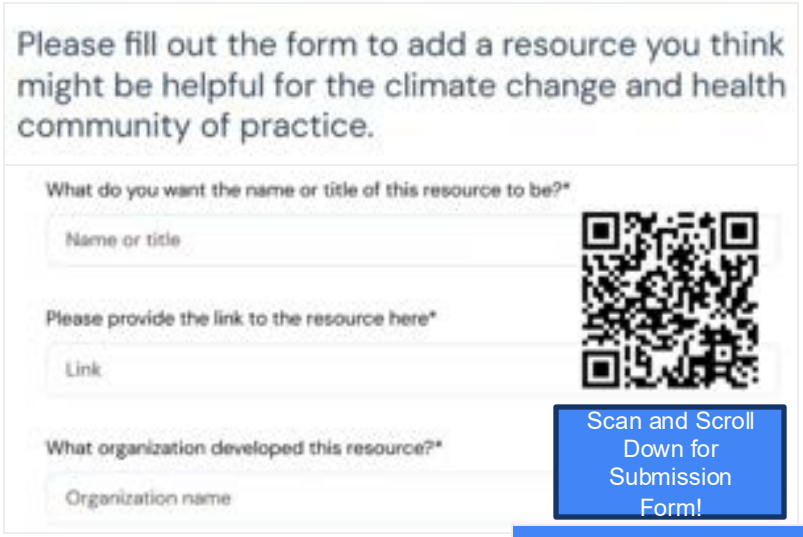
CAFÉ GitHub Organization: R & Pyhton Code Examples

Data Source	Product	Coverage
ERA5 Temperature	Example using Kenya Daily Heat Measure Statistics 2000-2023 (Example script, Dataverse entry)	Historic Weather Data, to 1950 including broader atmospheric levels
NASA FLDAS Temperature and Water Availability	Example using Iraq global administrative wards for 2000-2020 (Example script, Dataverse entry)	Measure monthly and annual temperature and drought metrics
Global Open StreetMap Road Exposure Metrics	Example using Mexico Administrative areas (Example script, Dataverse entry)	Calculate commonly used road measures into different geographical areas
PRISM Temperature	US ZCTA and County Population-Weighted 2020 Daily Max Temperature (Example script)	Historic Weather Data, to 1980 with high spatial resolution
Climate Data Online Weather Stations	US City-Level Time Series of Extreme Heat Days, 1950-Present (Example script, Dataverse entry)	Historic Weather Data, to 1940 for specific locations
CMIP6 Projected Temperature	US Tract-Level Daily Heat Threshold Measures at 2050 (Example script, Dataverse entry)	Projected Weather Data, through 2099 across SSPs
Global climate types and pm25 end-to-end pipelines	Aggregations for up to five embedded levels of geo boundaries for all countries in the globe (Pipelines, Dataverse entry)	Historical estimates, from 1998-2022

CAFÉ Educational Resource Hub

- Crowd-sourced, searchable catalog to centralize information across experience levels, areas of interest, programming languages, etc.
- Includes coding tips, epidemiological methods, coding tutorials, and more!





BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

https://www.climatehealthcafe.org/educational-resource-hub

CACHE

Center for Aging, Climate & Health

https://agingclimatehealth.org/

BU School of Public Health
Center for Climate & Health

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH
Department of Biostatistics

GECC | GATEWAY EXPOSOME COORDINATING CENTER

An NIA-funded initiative to advance life-course research on the AD/ADRD **exposome** (U24AG088894)

What is the exposome?

People experience a variety of exposures from their homes and communities. These exposures, broadly defined to capture social and behavioral exposures beyond physical and chemical, are the **exposome**. Independently and combined, they can impact people's risk for developing Alzheimer's Disease and Alzheimer's Disease Related Dementias (AD/ADRD).

GECC Aims

1. **Identify** research priorities via consensus-building among a broad range of stakeholders
2. **Develop** guidance for measuring, harmonizing, and using exposome data
3. **Create** novel exposome measures and data
4. **Disseminate** open-access resources and products to the broader research community

Domains

Process

Learn more at www.gatewayexposome.org

Thank You!!

Connect with CAFÉ

Visit us at our website
https://climatehealthcafe.org/

Find us on Linkedin at Climate & Health CAFÉ

Join our Community of Practice to receive our monthly newsletter and Slack channels!

Integration of Geospatial Data into Epidemiological and Clinical Health Research

Kyle P Messier

ISES/ISEE 2025 Pre-Meeting Workshop 06

Aug 17, 2025

Outline

- Geospatial Exposure Data
- Health Data
- Integration Challenges
- Tools to Support Geospatial Calculations and Integrations
- Amadeus Examples
- Open-Source Community
- Questions

Geospatial Data Sources

Environmental and Exposure Variables

- NASA EARTHDATA
- Multi-Resolution Land Characteristics Consortium – NLCD
- Zenodo – EU Open Research Repository
- CAFÉ Dataverse (https://dataverse.harvard.edu/dataverse.xhtml?alias=CAFE)
- CHORDS Catalog (https://chordshealth.org/discovery)

Health Data

Population and Person Level Data

- US Census
- CDC Places
- CDC Wonder

Address or Individual Level Data

- Point location of individual with health outcomes
- Epidemiological cohorts
- Electronic Health Records

- Minimum need is 1 spatial location
- Better is addresses over life course
- Best is detailed tracking
- Privacy concerns with address or time-course data

Data Integration

Spatial and Temporal Challenges

Journal of Exposure Science & Environmental Epidemiology www.nature.com/jes

REVIEW ARTICLE OPEN

A review of geospatial exposure models and approaches for health data integration

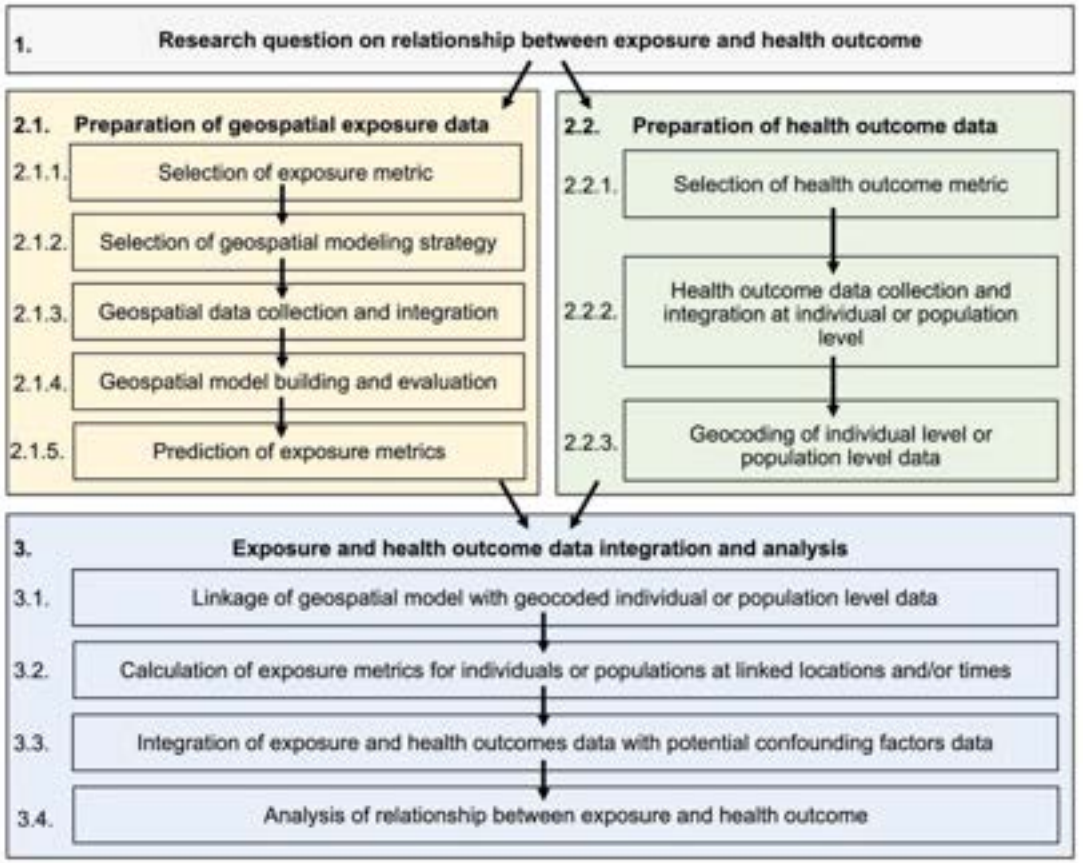
Lara P. Clark[1], Daniel Zilber[2], Charles Schmitt[1], David C. Fargo[3], David M. Reif[2], Alison A. Motsinger-Reif[4] and Kyle P. Messier[2,4]✉

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2024

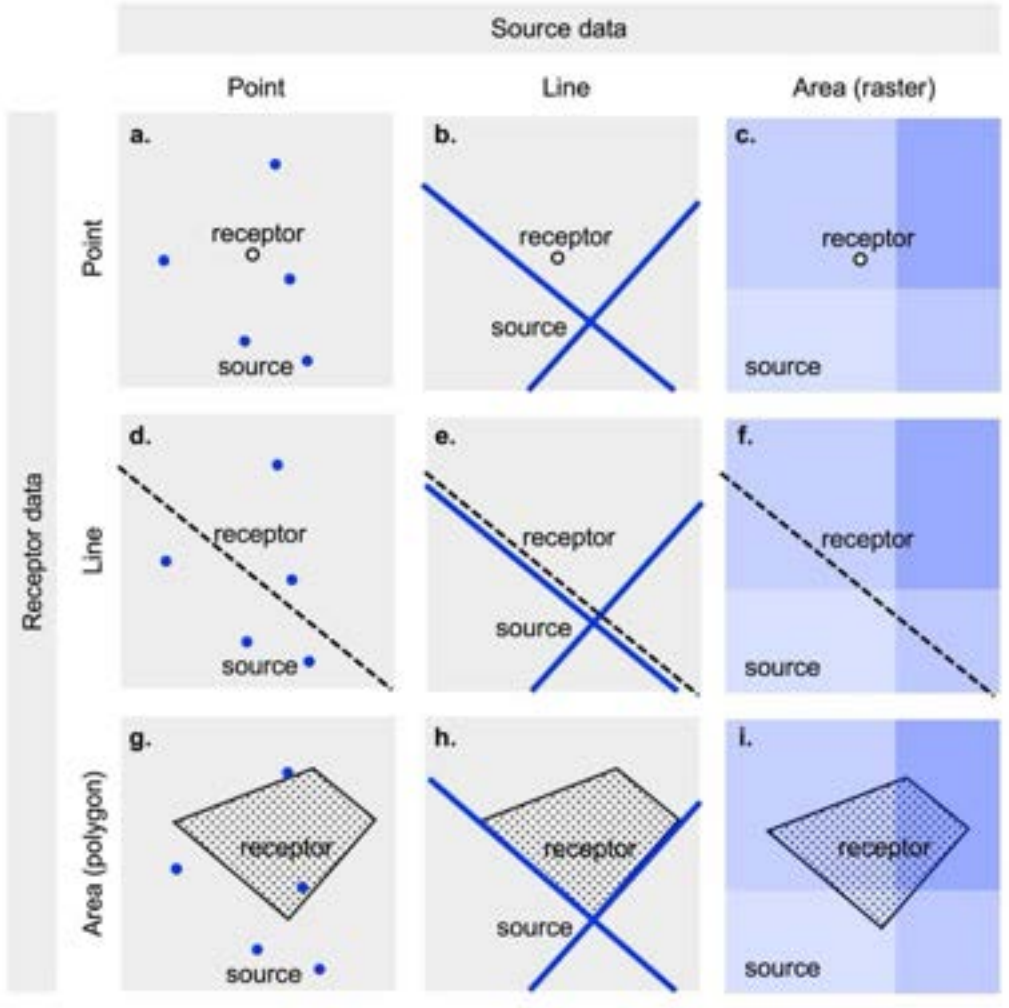
BACKGROUND: Geospatial methods are common in environmental exposure assessments and increasingly integrated with health data to generate comprehensive models of environmental impacts on public health.
OBJECTIVE: Our objective is to review geospatial exposure models and approaches for health data integration in environmental health applications.
METHODS: We conduct a literature review and synthesis.
RESULTS: First, we discuss key concepts and terminology for geospatial exposure data and models. Second, we provide an overview of workflows in geospatial exposure model development and health data integration. Third, we review modeling approaches, including proximity-based, statistical, and mechanistic approaches, across diverse exposure types, such as air quality, water quality, climate, and socioeconomic factors. For each model type, we provide descriptions, general equations, and example applications for environmental exposure assessment. Fourth, we discuss the approaches used to integrate geospatial exposure data and health data, such as methods to link data sources with disparate spatial and temporal scales. Fifth, we describe the landscape of open-source tools supporting these workflows.

Keywords: Exposome; Environmental public health; Exposure modeling; Spatiotemporal; Toxicology; Linkage

Journal of Exposure Science & Environmental Epidemiology (2025) 35:131–148; https://doi.org/10.1038/s41370-024-00712-8

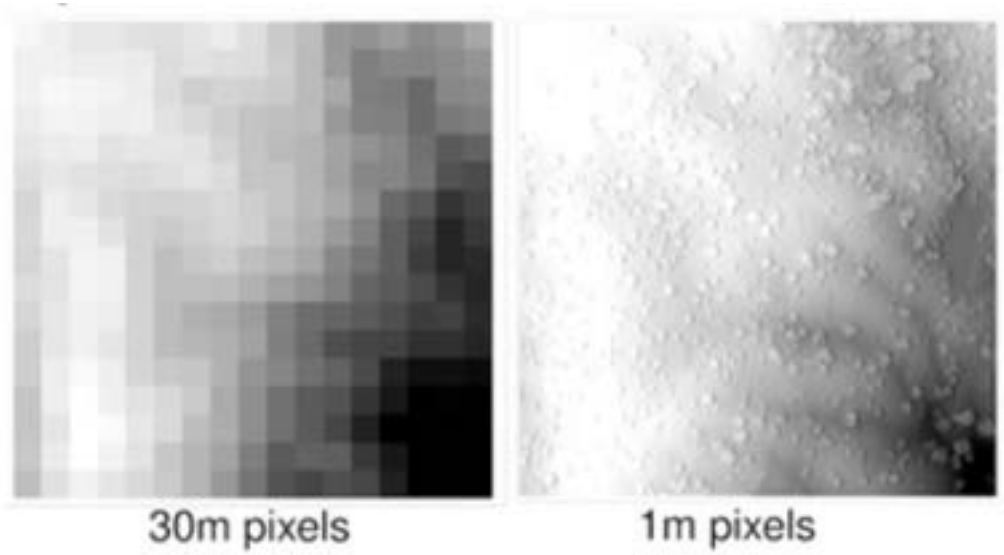


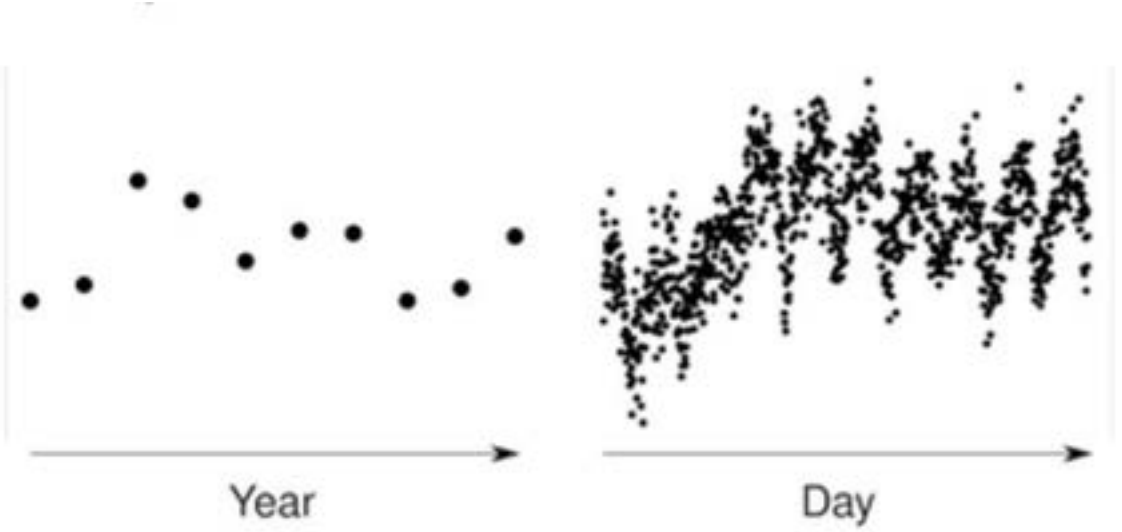
Spatial Geometry Relationships



- Source data → Geospatial and exposure data sources
- Receptor data → residential addresses, census tracts, etc.
- Integrating geospatial data to residential address → Row 1
- Integrating geospatial data to census tracts → Row 3

Spatial and Temporal Resolution





Tools to Help

- Amadeus
- Chopin
- Heat Exposure from Eva Marques
  - Samba
  - Mercury
  - Brassens
- Beethoven (in development)


amadeus
CHORDS
Samba
chopin
beethoven

amadeus: Accessing and analyzing large-scale environmental data in R

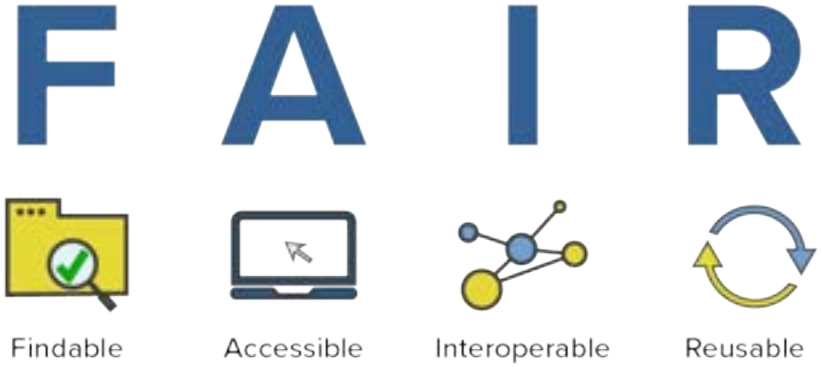
Mitchell Manware, Insang Song, Eva Marques, Mariana Alifa Kassien, Lara P Clark, Kyle P Messier

ISES/ISEE 2025 Pre-Meeting Workshop 06
Aug 17, 2025

National Institute of Environmental Health Sciences. (2024, September 23). *Area 4: Data Science and Computational Biology*. https://www.niehs.nih.gov/about/strategicplan/research/data-science
National Library of Medicine. (n.d.). *FAIR Data*. https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-200.html
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

"The development of innovative data science and data-driven approaches, such as integration of increasingly large and diverse data types and data sources, holds enormous potential for advancing key research needs across the environmental health sciences."

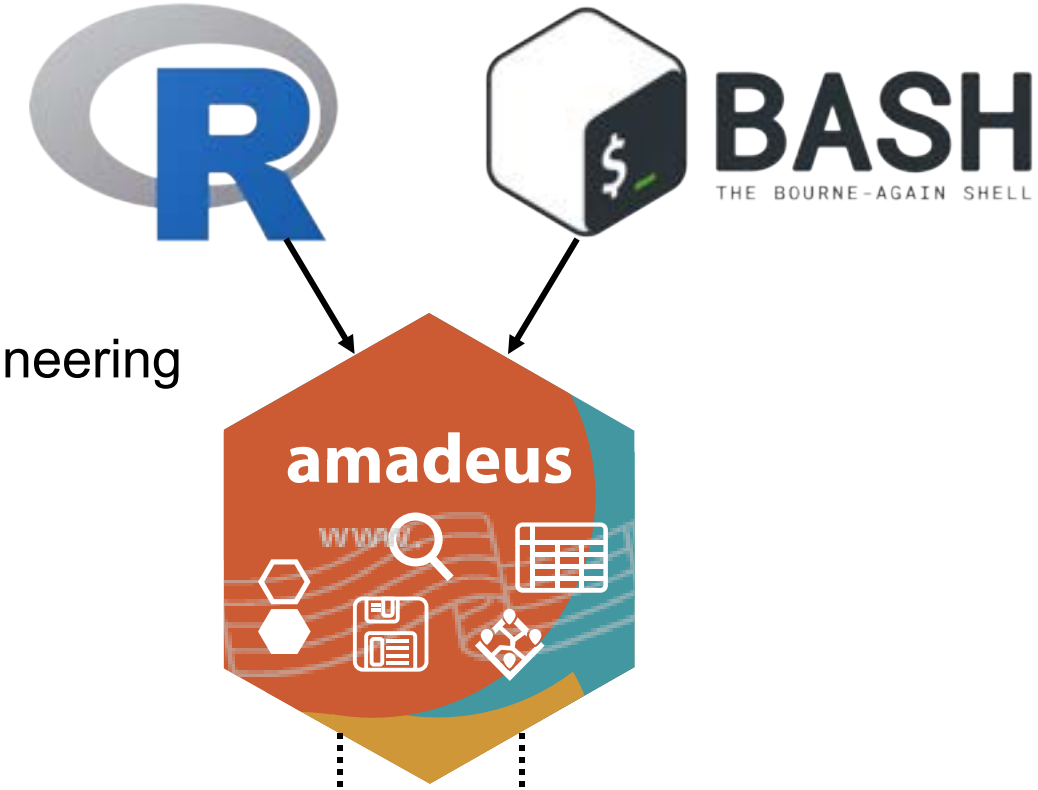
- NIEHS, Strategic Plan 2025 – 2029, Area 4: Data Science and Computational Biology

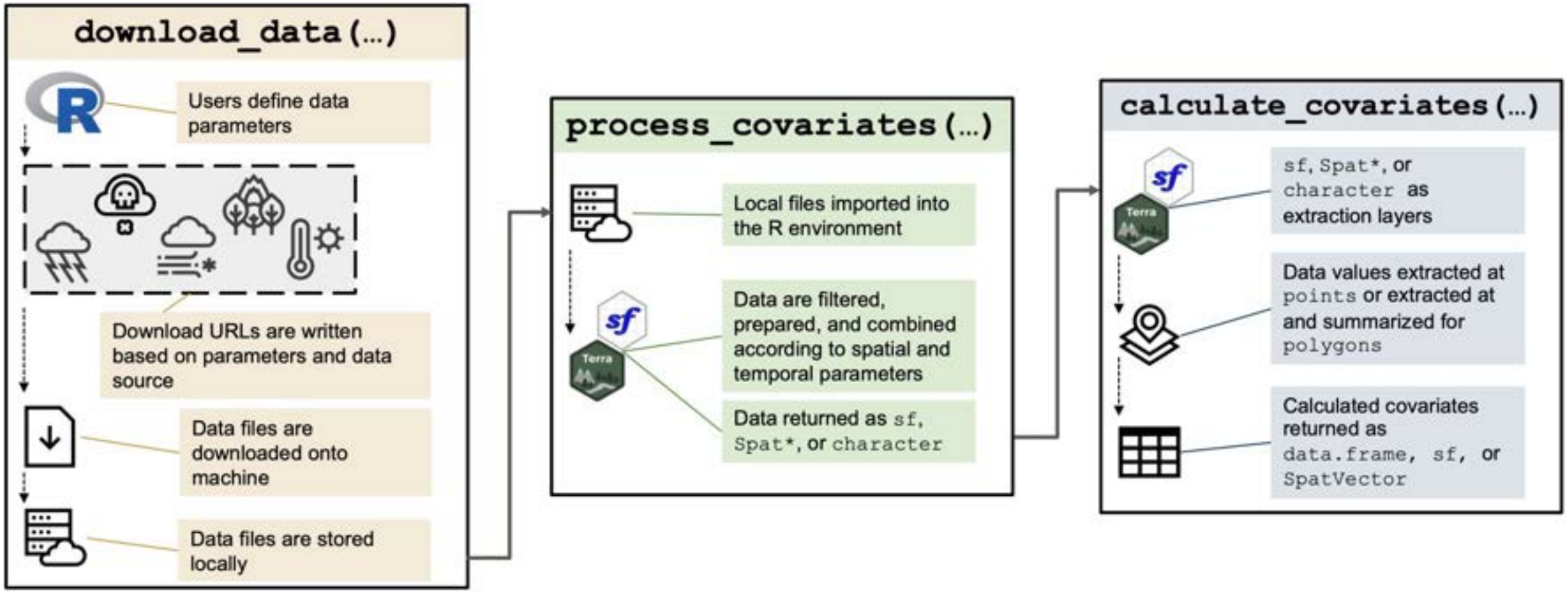


GNU Project. (2024). *Bash (Bourne Again Shell) [Computer software]*. Free Software Foundation. https://www.gnu.org/software/bash/
National Oceanic and Atmospheric Administration. (n.d.). *NOAA logo*. https://www.noaa.gov
National Aeronautics and Space Administration. (n.d.). *NASA logo*. https://www.nasa.gov
R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/
U.S. Department of the Interior. (n.d.). *DOI logo*. https://www.doi.gov
U.S. Environmental Protection Agency. (n.d.). *EPA logo*. https://www.epa.gov
U.S. Geological Survey. (n.d.). *USGS logo*. https://www.usgs.gov
University of California, Merced. (n.d.). *SNRI logo*. https://snri.ucmerced.edu

amadeus

- Data access, preparation, and covariate calculation
  - Features, model features, feature engineering
- 16 data sources; >1,000 variables
- Interoperable
- Extensible
- Test-driven development
- Open source



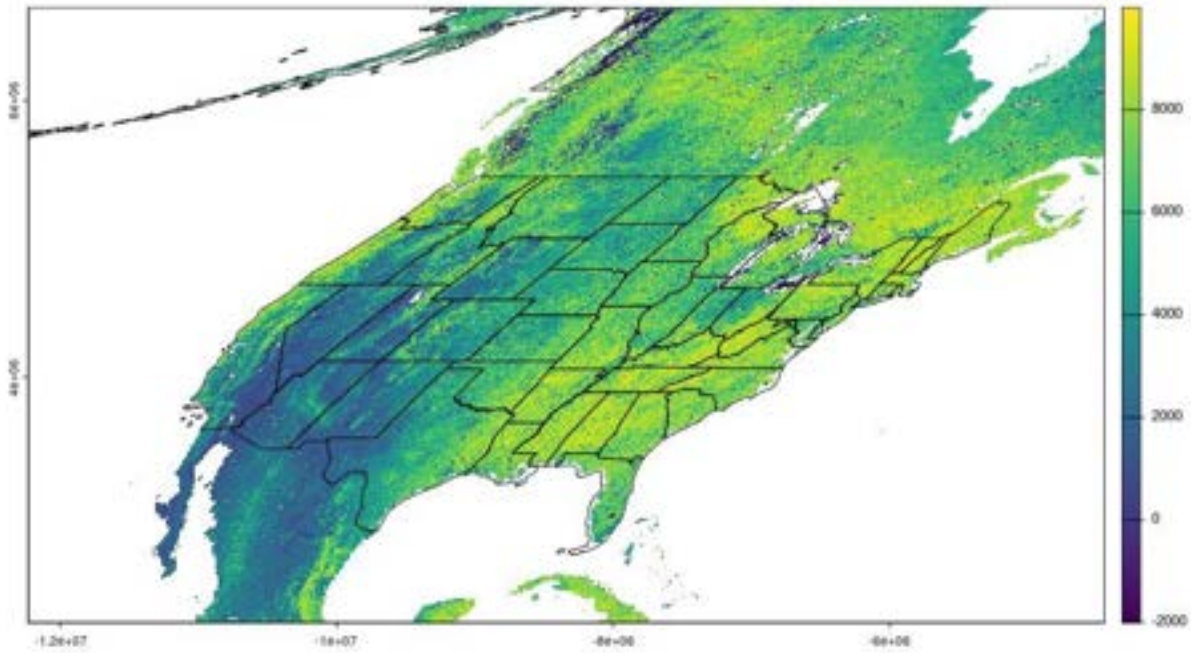

download_data(...)
Users define data parameters
Download URLs are written based on parameters and data source
Data files are downloaded onto machine
Data files are stored locally
process_covariates(...)
Local files imported into the R environment
Data are filtered, prepared, and combined according to spatial and temporal parameters
Data returned as sf, Spat*, or character
calculate_covariates(...)
sf, Spat*, or character as extraction layers
Data values extracted at points or extracted at and summarized for polygons
Calculated covariates returned as data.frame, sf, or SpatVector

Maccherone, B., and Frazier, S. (n.d.). *Moderate Resolution Imaging Spectroradiometer – Data*. https://modis.gsfc.nasa.gov/data/
NASA Earth Observing System Data and Information System. (n.d.). *MODIS Terra Vegetation Indices (MOD13A1) [Data set]*. NASA Goddard Space Flight Center. https://modis.gsfc.nasa.gov/data/

NASA Moderate Resolution Imaging Spectroradiometer (MODIS)

```
amadeus::download_modis(
  product = "MOD13Q1",
  version = "61",
  horizontal_tiles = c(7, 13),
  vertical_tiles = c(3, 6),
  nasa_earth_data_token = readLines("~/nasa_token.txt"),
  date = c("2023-06-01", "2023-06-10"),
  directory_to_save = directory_mod13q1,
  acknowledgement = TRUE,
  download = TRUE,
  remove_command = TRUE,
  hash = TRUE
)
ndvi_states <- amadeus::calculate_modis(
  from = list.files(
    directory_mod13q1, full.names = TRUE, recursive = TRUE
  ),
  locs = tigris::states(year = 2023),
  locs_id = "NAME",
  radius = 0,
  name_covariates = "NDVI_0_",
  subdataset = "(NDVI)",
  geom = "sf"
)
```

- Normalized Difference Vegetation Index (NDVI)
- Mean zonal statistics computed for U.S. states and territories
- 24 lines



Auxiliary functions

- Internal functions manage complexities
- Various file formats
- Native coordinate reference systems
- Extraction location geometries
  - Point or polygon

```
calc_prepare_locs <-
...
  #### project extraction locations to processed data
  sites_e <- process_locs_vector(
    locs,
    terra::crs(from),
    radius
  )
...
```

```
process_flatten_sds <- function(
  path = NULL,
  subdataset = NULL,
  fun_agg = "mean",
  ...
) {
    status_curv <- suppressWarnings(terra::is.rotated(terra::rast(path)))
    if (any(status_curv)) {
      stop(
        "The raster is curvilinear. Please rectify or warp
          the input then flatten it manually."
      )
    }
...
```

```
calc_worker <-
...
    #### extract layer data at sites
    if (terra::geomtype(locs_vector) == "polygons") {
      ### apply exactextractr::exact_extract for polygons
      sites_extracted_layer <- exactextractr::exact_extract(
        data_layer,
        sf::st_as_sf(locs_vector),
        progress = FALSE,
        force_df = TRUE,
        fun = fun,
        max_cells_in_memory = max_cells
      )
    } else if (terra::geomtype(locs_vector) == "points") {
      #### apply terra::extract for points
      sites_extracted_layer <- terra::extract(
        data_layer,
        locs_vector,
        method = "simple",
        ID = FALSE,
        bind = FALSE,
        na.rm = TRUE
      )
    }
...
```

Dowle, M., & Srinivasan, A. (2024). *data.table: Extension of 'data.frame'*. R package version 1.14.2. https://cran.r-project.org/package=data.table
Hijmans, R. J. (2024). *terra: Spatial data analysis package for R*. R package version 1.7-19. https://cran.r-project.org/package=terra
Miquel, M., & Dorman, M. (2024). *targets: Make dynamic and reproducible workflows*. R package version 1.0.0. https://cran.r-project.org/package=targets
Pebesma, E. J. (2024). *sf: Simple features for R*. R package version 1.0-13. https://cran.r-project.org/package=sf
Python Software Foundation. (2024). *Python programming language*. Version 3.11.5. https://www.python.org/
Walker, K. (2024). *tidycensus: Load US census data from the US Census Bureau's API*. R package version 1.0.0. https://cran.r-project.org/package=tidycensus
Walker, K. (2024). *tigris: R package for loading TIGER/Line shapefiles*. R package version 1.0.0. https://cran.r-project.org/package=tigris
Wickham, H., François, R., Henry, L., & Müller, K. (2024). *dplyr: A grammar of data manipulation*. R package version 1.1.2. https://cran.r-project.org/package=dplyr

Interoperability

- Output classes are from popular spatial packages
- Linking to U.S. Census data
- Data analysis pipelines
- Beyond R

```r
# R
amadeus::download_data(
  dataset_name = "narr",
  variable = "shum",
  year = 2021,
  directory_to_save = "./data/",
  acknowledgement = TRUE,
  download = TRUE,
  remove_command = TRUE
)
```

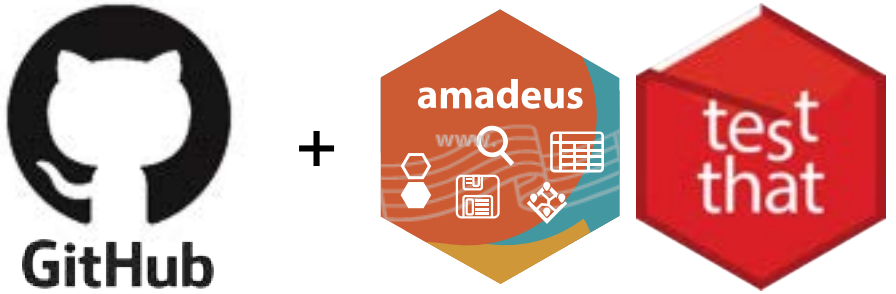
```python
# Python
import netCDF4
shum_python = netCDF4.Dataset("data/shum/shum.202101.nc")
shum_python.variables["shum"]
```

```
## <class 'netCDF4._netCDF4.Variable'>
## float32 shum(time, level, y, x)
##     GRIB_id: 51
##     GRIB_name: SPFH
##     _FillValue: 9.96921e+36
##     coordinates: lat lon
##     grid_mapping: Lambert_Conformal
##     level_desc: Pressure Levels
##     standard_name: specific humidity
```

GitHub, Inc. (n.d.). *GitHub logo*. https://github.com/logos
GitHub, Inc. (n.d.). *GitHub Invertocat logo*. https://github.com/logos
Marsner. (n.d.) *Why Test-Driven Development (TDD)*. https://marsner.com/blog/why-test-driven-development-tdd/
Wickham, H. (2024). *testthat: Get started with testing*. R package version 3.1.0. https://cran.r-project.org/package=testthat

Test-driven development

- Reliability and stability
- Error catching during development
- Interpretable error messages
- Unit tests
  - URLs return successful response status (`download_*`)
  - Correct object classes (`process_*`; `calculate_*`)
  - Identify missing or improper parameters (all functions)
- Integration tests
  - No errors when `process_*` run with live `download_*` files



CRAN 1.2.3

repo status Active

amadeus: Accessing and Analyzing Large-Scale Environmental Data

Functions are designed to facilitate access to and utility with large scale, publicly available environmental data in R. The package contains functions for downloading raw data files from web URLs (download_data()), processing the raw data files into clean spatial objects (process_covariates()), and extracting values from the spatial data objects at point and polygon locations (calculate_covariates()). These functions call a series of source-specific functions which are tailored to each data sources/datasets particular URL structure, data format, and spatial/temporal resolution. The functions are tested, versioned, and open source and open access. For sum_edc() method details, see Messier, Akita, and Serre (2012) <doi:10.1021/es203152a>.

Version:	1.2.3
Depends:	R (≥ 4.1.0)
Imports:	dplyr, sf, sftime, stats, terra, methods, data.table, httr, rvest, exactextractr, utils, stringi, testthat (≥ 3.0.0), stars, tidyr, rlang, nhdplusTools, archive, collapse, Rdpack
Suggests:	covr, withr, knitr, rmarkdown, lwgeom, FNN, doRNG, devtools, stringr, tigris, spelling
Published:	2025-01-16
DOI:	10.32614/CRAN.package.amadeus
Author:	Mitchell Manware [aut, ctb], Insang Song [aut, ctb], Eva Marques [aut, ctb], Mariana Alifa Kassien [aut, ctb], Elizabeth Scholl [ctb], Kyle Messier [aut, cre], Spatiotemporal Exposures and Toxicology Group [cph]
Maintainer:	Kyle Messier <kyle.messier at nih.gov>
BugReports:	https://github.com/NIEHS/amadeus/issues
License:	MIT + file LICENSE
URL:	https://niehs.github.io/amadeus/
NeedsCompilation:	no
Language:	en-US
Materials:	README, NEWS
CRAN checks:	amadeus results

Documentation:

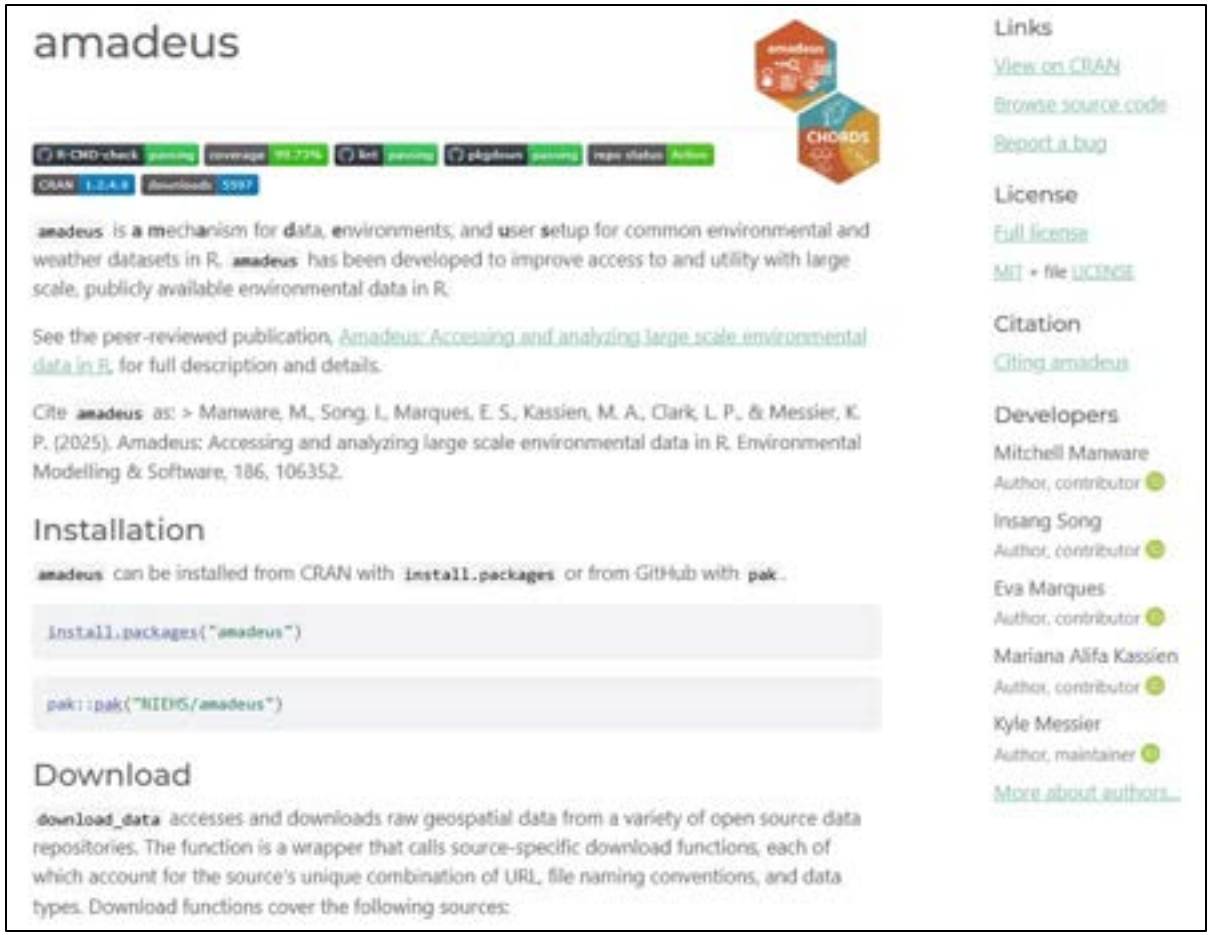
Reference manual:	amadeus.html , amadeus.pdf
Vignettes:	Climatology Lab gridMET (source, R code) North American Regional Reanalysis (NARR) (source, R code) Climatology Lab TerraClimate (source, R code)

Downloads:

Package source:	amadeus_1.2.3.tar.gz
Windows binaries:	r-devel: amadeus_1.2.3.zip, r-release: amadeus_1.2.3.zip, r-oldrel: amadeus_1.2.3.zip
macOS binaries:	r-release (arm64): amadeus_1.2.3.tgz, r-oldrel (arm64): amadeus_1.2.3.tgz, r-release (x86_64): amadeus_1.2.3.tgz, r-oldrel (x86_64): amadeus_1.2.3.tgz
Old sources:	amadeus archive

Linking:

Please use the canonical form https://CRAN.R-project.org/package=amadeus to link to this page.



amadeus

R-CMD-check passing | coverage 91.72% | lint passing | pkgdown passing | repo status Active | CRAN 1.2.4.9 | downloads 5597

amadeus is a mechanism for data, environments, and user setup for common environmental and weather datasets in R. **amadeus** has been developed to improve access to and utility with large scale, publicly available environmental data in R.

See the peer-reviewed publication, Amadeus: Accessing and analyzing large scale environmental data in R, for full description and details.

Cite **amadeus** as: > Manware, M., Song, I., Marques, E. S., Kassien, M. A., Clark, L. P., & Messier, K. P. (2025). Amadeus: Accessing and analyzing large scale environmental data in R. Environmental Modelling & Software, 186, 106352.

Installation

amadeus can be installed from CRAN with **install.packages** or from GitHub with **pak**.

```
install.packages("amadeus")
```

```
pak::pak("NIEHS/amadeus")
```

Download

download_data accesses and downloads raw geospatial data from a variety of open source data repositories. The function is a wrapper that calls source-specific download functions, each of which account for the source's unique combination of URL, file naming conventions, and data types. Download functions cover the following sources:

Links

View on CRAN

Browse source code

Report a bug

License

Full license

MIT + file LICENSE

Citation

Citing amadeus

Developers

Mitchell Manware
Author, contributor

Insang Song
Author, contributor

Eva Marques
Author, contributor

Mariana Alifa Kassien
Author, contributor

Kyle Messier
Author, maintainer

More about authors...

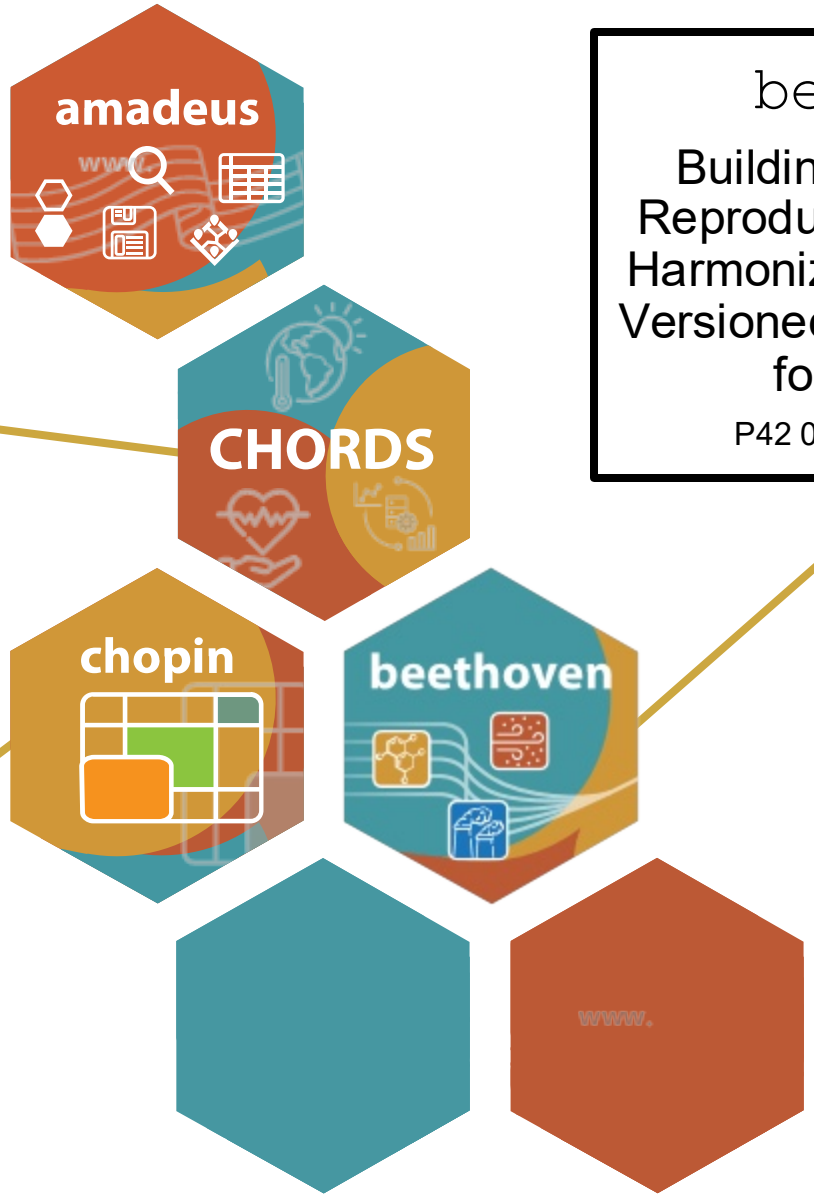
National Institute of Environmental Health Sciences. (2024, August 14). *Climate and Health Outcomes Research Data Systems (CHORDS)*. https://www.niehs.nih.gov/research/programs/chords
Song, I. & Messier, K. (2024). chopin: Computation of Spatial Data by Hierarchical and Objective Partitioning of Inputs for Parallel Processing. R package version 0.9.0. https://docs.ropensci.org/chopin
Song, I., Marques, E., Manware, M., et al. (2024). beethoven: Building an Extensible, rEproducible, Test-driven, Harmonized, Open-source, Versioned, ENsemble model for air quality. R package version 0.4.3. https://github.com/NIEHS/beethoven

Connecting Health Outcomes Research Data Systems (CHORDS)

P17 08/18/2025 15:30 – 16:15

chopin

Computation of Spatial Data by Hierarchical and Objective Partitioning Inputs for Parallel Processing



beethoven

Building an Extensible, Reproducible, Test-driven, Harmonized, Open-Source, Versioned, Ensemble Model for Air Quality

P42 08/20/25 09:00-09:45

chopin

Convenient parallel geospatial operations via data partitioning and automatic deployment

- Regular grid, hierarchical, balanced

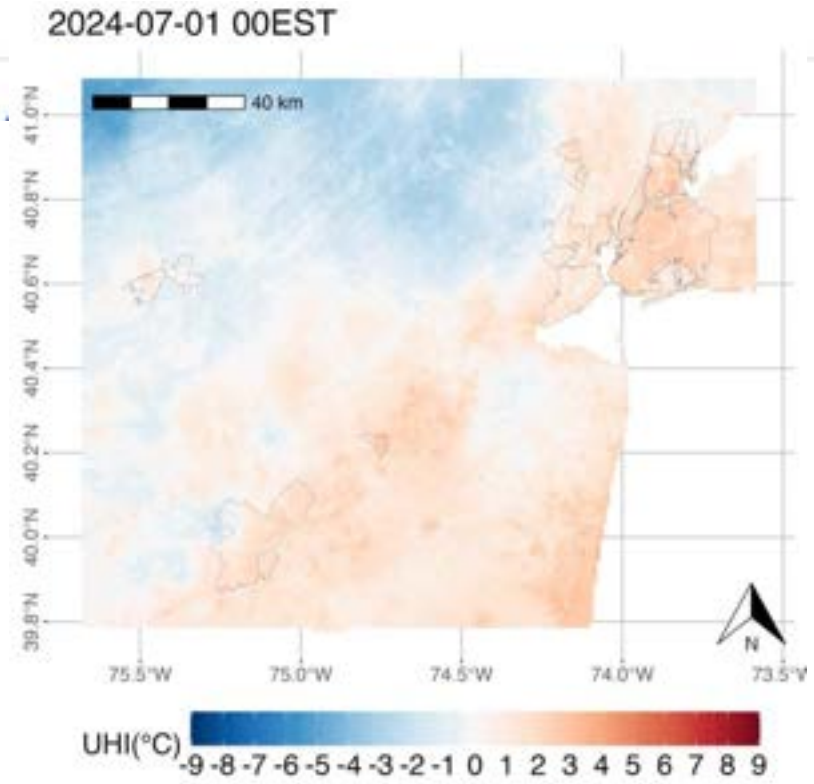
Custom and common gridding (Uber H3, DGGRID) options

Available in CRAN

Published in *SoftwareX*

lint passing DOI 10.5281/zenodo.15595978

samba: Spatiotemporal Air-temperature Model with Bayesian Approach



https://github.com/NIEHS/urbanheat_pipeline

README

A factory to generate air temperature data at hourly resolution on US cities!

This project is an assembly line using brassens, samba and mercury libraries to automatically generate 2-meters air temperature rasters on a list of US cities and observed months.

This pipeline has generate casestudies on the top largest Urban Census areas on different climatic events (heatwave, blizzard, typical weather...) and seasons. This dataset has been designed for environmental epidemiology and public health studies on (extreme) temperature exposure in US cities. It can also be leveraged by urban climatologists to improve our understanding of the spatiotemporal evolution of the urban heat island with regard to the variety of city layout and climatic region in the US.

To cite the dataset: Marques, E., & Messier, K. (2025). Urban air temperature at high spatiotemporal resolution on major US cities (soon available) [Dataset]. Harvard Dataverse. https://doi.org/10.7910/DVN/HNVCBR756

Urban heat island movie generation

If you want to create a movie from the 2m-air temperature raster:

- go on our urbanheat_pipeline Github repository.
- download `container/container_movies.def` , `create_movie.sh` and `create_maps_for_movie.R`
- open a terminal and build the container (you need to have apptainer installed). It might take a few minutes.

```
apptainer build --fakeroot container_movies.sif container_movies.def
```

Urban air temperature at high spatiotemporal resolution on major US cities

Jun 24, 2025

Marques, Eva; Messier, Kyle, 2025, "Urban air temperature at high spatiotemporal resolution on major US cities", https://doi.org/10.7910/DVN/HNVCBR, Harvard Dataverse, V2

... This dataset has been designed for environmental epidemiology and public health studies on (extreme) temperature exposure in US cities. It can also be leveraged by **urban** climatologists to improve our understanding of the spatiotemporal evolution of the **urban heat** island with regard to the variety of city layout and climatic region in the US. ...
Related Publication Citation: @article{marques2025improved, title={Improved high resolution heat exposure assessment with personal weather stations and spatiotemporal Bayesian models}, author={Marques, Eva and Messier, Kyle P}, journal={Authorea Preprints}, year={2025}, publisher={Authorea} }

National Institute of Environmental Health Sciences. (2024, September 23). *Area 4: Data Science and Computational Biology*. https://www.niehs.nih.gov/about/strategicplan/research/data-science
National Library of Medicine. (n.d.). *FAIR Data*. https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-200.html
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al*. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* 3, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

F

Findable

- CRAN
- GitHub
- Publication†
- CHORDS

A

Accessible

- Open source
- No cost

I

Interoperable

- `sf` and `terra`
- Python and other programming languages

R

Reusable

- Local files
- Re-executable code

†Manuscript published at *Environmental Modelling & Software*

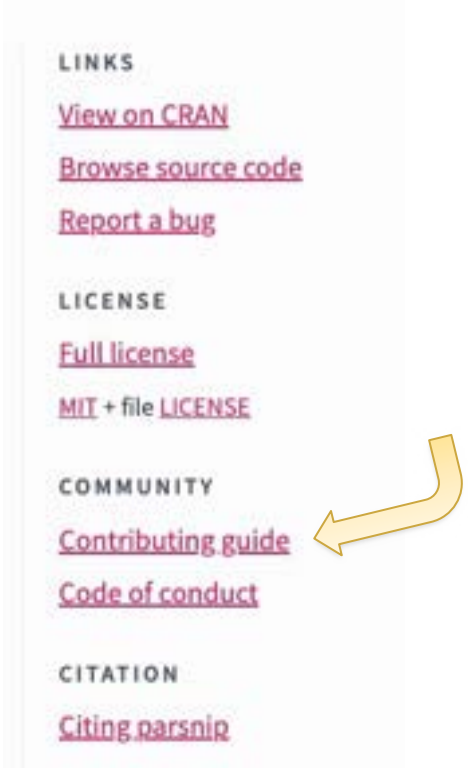
Amadeus Examples

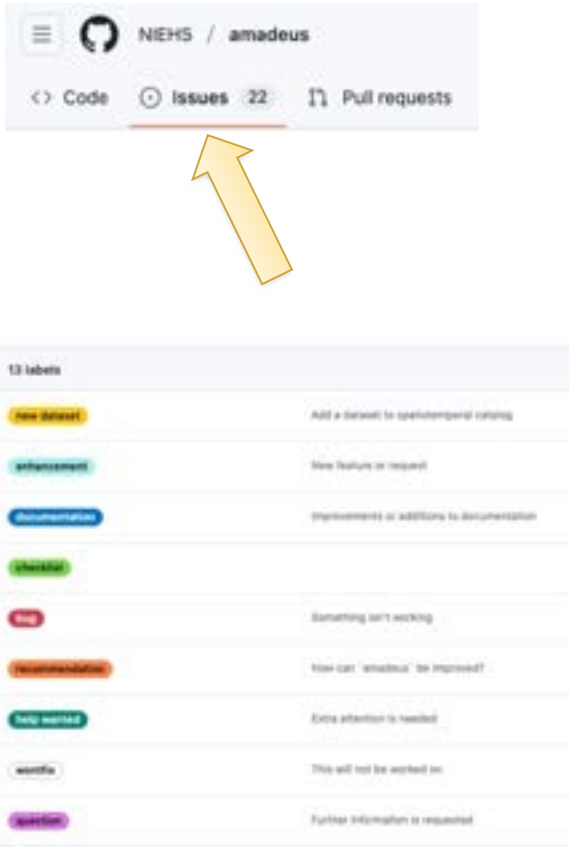
Spatial and Temporal Integration

Amadeus & Open-Source Community

That’s you!

Participating in Open-Source Projects





1. Open Issues
2. Ask Questions
3. Report Bugs
4. Help with Documentation
5. Contribute examples

Acknowledgements

- Mitchell Manware, M.S.
- Insang Song, Ph.D.
- Connecting Health Outcomes Research and Data Systems Team
  - Aubrey K Miller, M.D., M.P.H.
  - Charles P Schmitt, Ph.D.
  - David M. Reif, Ph.D.
  - Adam B. Burkholder, M.S.
  - Lara P Clark, Ph.D.
  - Alison A. Motsinger-Reif, Ph.D.
  - Ann Liu, Ph.D., M.P.H.
  - Trisha M. Castranio

Questions?

Scan for the GitHub codebase.

Dowle, M., & Srinivasan, A. (2024). data.table: Extension of 'data.frame'. R package version 1.14.2. https://cran.r-project.org/package=data.table
GitHub, Inc. (n.d.). *GitHub logo*. https://github.com/logos
GitHub, Inc. (n.d.). *GitHub Invertocat logo*. https://github.com/logos
GNU Project. (2024). *Bash (Bourne Again Shell) [Computer software]*. Free Software Foundation. https://www.gnu.org/software/bash/
Hijmans, R. J. (2024). terra: Spatial data analysis package for R. R package version 1.7-19. https://cran.r-project.org/package=terra
Maccherone, B., and Frazier, S. (n.d.). *Moderate Resolution Imaging Spectroradiometer – Data*. https://modis.gsfc.nasa.gov/data/
Marsner. (n.d.) *Why Test-Driven Development (TDD).* https://marsner.com/blog/why-test-driven-development-tdd/
Miquel, M., & Dorman, M. (2024). targets: Make dynamic and reproducible workflows. R package version 1.0.0. https://cran.r-project.org/package=targets
NASA Earth Observing System Data and Information System. (n.d.). *MODIS Terra Vegetation Indices (MOD13A1) [Data set]*. NASA Goddard Space Flight Center. https://modis.gsfc.nasa.gov/data/
National Aeronautics and Space Administration. (n.d.). *NASA logo*. https://www.nasa.gov/
National Institute of Environmental Health Sciences. (2024, August 14). Climate and Health Outcomes Research Data Systems (CHORDS). https://www.niehs.nih.gov/research/programs/chords
National Institute of Environmental Health Sciences. (2024, September 23). *Area 4: Data Science and Computational Biology*. https://www.niehs.nih.gov/about/strategicplan/research/data-science
National Library of Medicine. (n.d.). *FAIR Data*. https://www.nlm.nih.gov/oet/ed/cde/tutorial/02-200.html
National Oceanic and Atmospheric Administration. (n.d.). *NOAA logo*. https://www.noaa.gov
Pebesma, E. J. (2024). sf: Simple features for R. R package version 1.0-13. https://cran.r-project.org/package=sf
Python Software Foundation. (2024). Python programming language. Version 3.11.5. https://www.python.org/
R Core Team. (2024). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/
Song, I. & Messier, K. (2024). chopin: Computation of Spatial Data by Hierarchical and Objective Partitioning of Inputs for Parallel Processing. R package version 0.9.0. https://docs.ropensci.org/chopin
Song, I., Marques, E., Manware, M., et al. (2024). beethoven: Building an Extensible, rEproducible, Test-driven, Harmonized, Open-source, Versioned, ENsemble model for air quality. R package version 0.4.3. https://github.com/NIEHS/beethoven
U.S. Department of the Interior. (n.d.). *DOI logo*. https://www.doi.gov/
U.S. Environmental Protection Agency. (n.d.). *EPA logo*. https://www.epa.gov/
U.S. Geological Survey. (n.d.). *USGS logo*. https://www.usgs.gov/
University of California, Merced. (n.d.). *SNRI logo*. https://snri.ucmerced.edu/
Walker, K. (2024). tidycensus: Load US census data from the US Census Bureau's API. R package version 1.0.0. https://cran.r-project.org/package=tidycensus
Walker, K. (2024). tigris: R package for loading TIGER/Line shapefiles. R package version 1.0.0. https://cran.r-project.org/package=tigris
Wickham, H. (2024). *testthat: Get started with testing*. R package version 3.1.0. https://cran.r-project.org/package=testthat
Wickham, H., François, R., Henry, L., & Müller, K. (2024). dplyr: A grammar of data manipulation. R package version 1.1.2. https://cran.r-project.org/package=dplyr
Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18

National Institute of Environmental Health Sciences
Your Environment. Your Health.

Efforts to standardize geospatial data elements and to create geospatial CDEs

Maria Shatz, PhD
Office of Data Science
National Institute of Environmental Health Sciences (NIEHS)
Research Triangle Park, NC, USA
e-mail: maria.shatz@nih.gov

Conflict of Interest Statement

The author declares no conflict of interest.

Examples of geospatial measures for health

Geospatial Measure	Description	Example Data Sources	Public Health Use
Air Quality Index (AQI), PM2.5, NO_2, O_3	Measures air pollution levels by location	EPA AirNow, NASA satellite remote sensing	Identify high-risk areas for respiratory/cardiovascular illness; issue health alerts
Proximity to Pollution Sources	Distance to highways, animal feed facilities, oil wells, waste sites	National Land Cover Database, EPA Toxic Release Inventory	Assess exposure risk and plan zoning/regulatory actions
Noise Pollution Maps	Sound intensity across neighborhoods	Environmental monitoring stations noise sensors	Study effects on sleep, cardiovascular health, stress
Green Space Availability	Access to parks, tree canopy coverage	Satellite imagery, OpenStreetMap	Promote physical activity, mental health benefits
Wildfire and smoke exposure	Location based exposure to smoke and particle pollution	AirNow Fire and Smoke Map, US EPA and Forest Service	Enables the public to take protective actions against wildfire smoke.
Population Density & Demographics	Distribution of population groups (age, income, etc.)	Census data	Plan resource allocation and emergency response
Disease Incidence Heat Maps	Spatial distribution of reported cases	Public health registries, GIS geocoding	Detect clusters/outbreaks for rapid response
Flood/Heatwave/Wildfire Risk Maps	Areas prone to climate hazards	FEMA flood maps, NOAA climate data, satellite fire tracking	Prepare evacuation plans, target public health messaging
Water Quality Mapping	Location of contamination in rivers, wells, and systems	EPA Safe Drinking Water Information System	Prevent and mitigate waterborne disease outbreaks

Value of standardized data elements

Reason	Impact if standardized	Impact if Not standardized
Consistency	PM2.5 in one state means the same as PM2.5 in another	Apples-to-oranges comparisons between regions and cohorts
Data Integration	Easily merge environmental, health, and demographic datasets	Incompatible formats, loss of detail
Reproducibility	Other researchers can verify and replicate findings	Conflicting results due to method differences
Scalability	GIS tools and dashboards work in multiple jurisdictions	Costly customizations for each location
Public Trust	Clear, consistent public health communication	Confusion, misinformation, reduced engagement

Core Principles for Standardization

- **Clear Definitions** – Agree on exactly what is measured (e.g., “walkability” defined by X metrics).
- **Common Units** – Use the same measurement units (e.g., micrograms/m³ for PM2.5).
- **Temporal Alignment** – Match timeframes (e.g., daily averages vs. annual means).
- **Spatial Resolution** – Define the geographic scale (e.g., census tract, zip code, 1 km² grid).
- **Metadata Documentation** – Record data sources, collection methods, and processing steps.
- **Interoperable Formats** – Use open, standardized GIS data formats (GeoJSON, shapefiles).

Implementation Path

- **Adopt Existing Standards** – Use established frameworks (CDC's SVI, WHO air quality guidelines).
- **Collaborative Governance** – Engage diverse stakeholders in setting norms.
- **Capacity Building** – Train staff in GIS, data cleaning, and metadata creation.
- **Continuous Review** – Update standards as technology and health priorities evolve.

Challenges

- While there are standards for health data (e.g. ICD-10, LOINC, SNOMED) or for geospatial data (Open Geospatial Consortium, OGD) there are no standard on intersection of the two fields
- Multiple existing measures with no community consensus
- Researchers are not aware of the existing standards
- Variety of data sources and computational methods
- Due to the temporal element the values are dynamic;
- The presentation should be not only machine-readable but easily reused for computation with different parameters

Approaches to standardization of data elements

How do we move forward as a community to improve this situation?

NIH proposed creating and endorsing CDEs and publishing them in the CDE repository moving forward

What is a Common Data Element (CDE)?

- A CDE is a fixed representation of a variable comprising one defined question
- paired with a specified set of allowable responses including value range and units if applicable,
- mapped to semantic concepts and codes like UMLS, SNOMED AND
- that are **used in common** across multiple: studies, research sites, initiatives, datasets, etc.

Question (variable): *What is the participant's height?*

Allowable answers: *Feet and inches in whole numbers*

- For example, if all clinical studies supported by an IC were required to collect height data with this question and answer pair,

- this data element would be common to all of those studies - it would be a ***Common Data Element***

Adopted from CDE WG presentation materials

NIH CDE Governance Committee

- The CDE Governance Committee of the NIH CDE Task Force will determine whether CDEs meet **criteria that merit their designation as "NIH-endorsed CDEs."**
- The criteria:
  - Submission from a **recognized NIH body** (e.g., NIH ICO, trans-NIH committee, etc.)
  - **Clear definition** of the variable
  - For assessment instruments, documented evidence of **reliability and validity**
  - Both human- and machine-readable **formats** preferred
  - Clear **licensing** and IP status for re-use (prefer CC-by or open source)
- NIH-endorsed CDEs will be available through the NIH CDE Repository

Adopted from CDE WG presentation materials

Proposed structure for geospatial CDEs

Is change in health condition/outcome linked to an environmental factor?

Q: What change in health condition has the patient experienced?

A: List: Hospitalization/worsening of symptoms/new symptoms

Q: Which health condition/outcome was affected?

A: A standardized list of health outcomes (asthma, cardiovascular event, fatigue, etc.)

Q: What was patient's main exposure location during period of interest

A: Location of residential or work or recreation, address or zip code. How do we justify and describe our choice of "main" location?

Q: What environmental factor has the patient been exposed to?

A: A standardized list of environmental factors (direct measurement of allergens, smoke, pesticides, air pollution, weather extremes, or modeling based on proximity to animal feed facilities, agricultural fields). Do we use individual factors or some aggregated indices?

Q: For how long the patient was exposed during period of interest

A: Duration of exposure standard during period of interest (period of interest can be anything from last week to second trimester of pregnancy, duration can be a single number or a range - >1h,2-5h, 6-12h etc.)

Proposed structure for geospatial CDEs

Q: What was the intensity of the exposure?

A: Numeric value

Units

Value type: Computed or directly measured

If computed

Source of data

Model used (algorithm, meteorological data source if used)

Calculated value is: range, mean, median, max, min etc. during indicated above ***period of interest***

If directly measured

Method,

Instrument

Beyond metadata schema

- Can we propose a core set of environmental geospatial CDEs, perhaps bundled by disease? Reminder: CDEs are **used in common** across multiple: studies, research sites, initiatives, datasets, etc.
- Ongoing effort – conversion PEGS (Personalized Environment and Genes Study) surveys developed in coordination with NHANES studies into CDEs – funded by ODSS and in coordination with NCI team

Existing Efforts on EHS Standards

- **ESIP Geo4Health cluster** – recent session ***Communicating through Bio Geo divide*** during the annual meeting
- **Gravity Project** is part of HL7 initiative that seeks to identify data elements and associated value sets to represent SDOH information documented in electronic health records (EHRs) across clinical activities. They recently started an effort on identifying **Community Level Domains: Food Access, Neighborhood Safety, Access to Green Space** data elements.
- **GA4GH** recently created **Human Exposome Data Standards Group** and are exploring ways to support geolocation and exposure data in **Phenopackets,** will be hosting a Human Exposome session titled Bridging Location and Phenotype
- **American Medical Informatics Association** (AMIA) **Climate, Health and Informatics WG**
- **Network for Exposomics in the United States (**Nexus) is building a **Geospatial Sciences Hub** to promote Understanding of Place-Based Exposures
- **Gateway Exposome Coordinating Center** (GECC) aims to develop guidance for measuring, harmonizing, and using exposome data