Session 1: Tools and Other Resources

TADA: Tools for Automated Data Analysis

Cristina Mullin, Ph.D.

EPA HQ

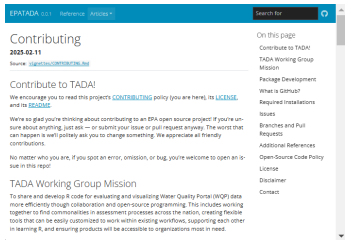
2025 National Training Workshop on Water Quality Data, Assessment, and Plans

Disclaimer

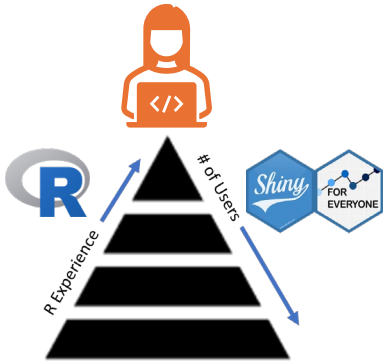
This presentation and the information contained in these slides do not represent binding requirements on the states. Such requirements are found in the Clean Water Act and EPA's implementing regulations.

1. What Is TADA?

A suite of open-source R tools (**R package and R Shiny applications**) designed to efficiently discover, compile, clean, analyze, and visualize data from the **Water Quality Portal (WQP)**. Moreover, this project provides a **hub for an open-source water quality community** to collaborate.



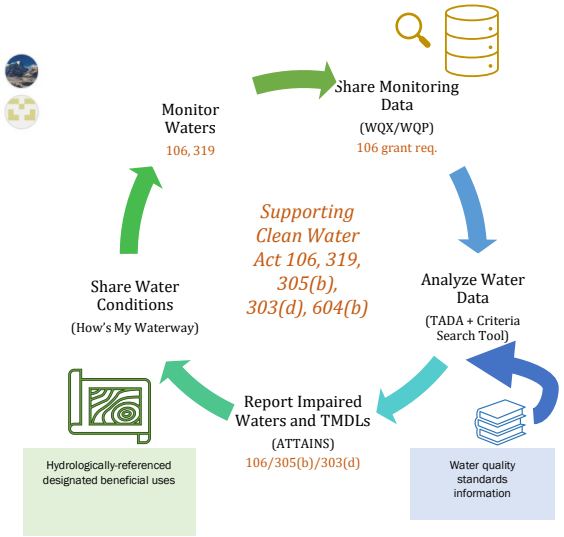
epa.gov/waterdata/TADA



NATIONAL WATER QUALITY MONITORING COUNCIL

WQX WATER QUALITY EXCHANGE

dataRetrieval USGS

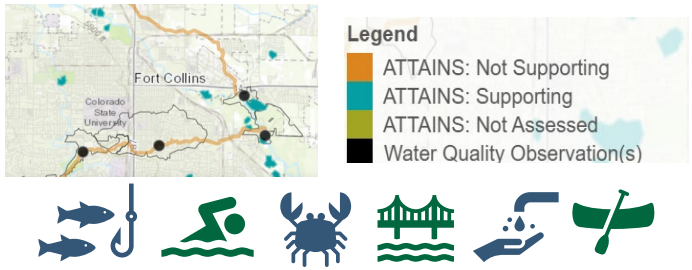


2. How Does TADA work?

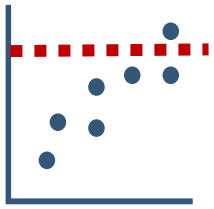
Module 1: Discover, wrangle and quality control data

- Retrieve data, correct data types, remove duplicates and filter
- Flag suspect metadata, results, and locations
- Handle censored (e.g., non-detect) data
- Harmonize synonyms, units, and speciation
- Group nearby sites and review sampling depths
- Total nitrogen and phosphorus summation

Module 2: Overlay ATTAINS assessment units, uses, and parameters with WQP monitoring locations (*in active development*)



Module 3: Analyze water quality with user supplied criteria and methodologies (*in active development)*

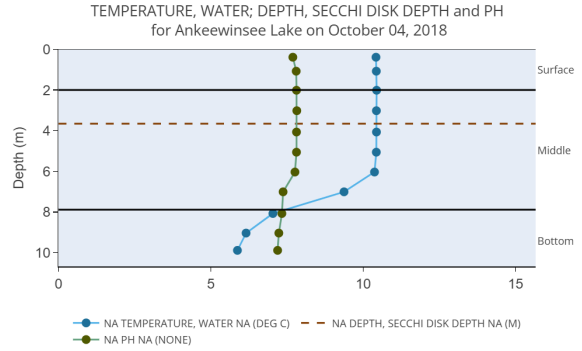


- Criteria components
  - Frequency
  - Duration
  - Magnitude
- Additional complexities
  - Seasonality
  - Site-specific
  - Acute vs. chronic exposure
  - Spatial and temporal aggregation
  - Sampling depth
  - And more!

3. Why Is TADA Useful?

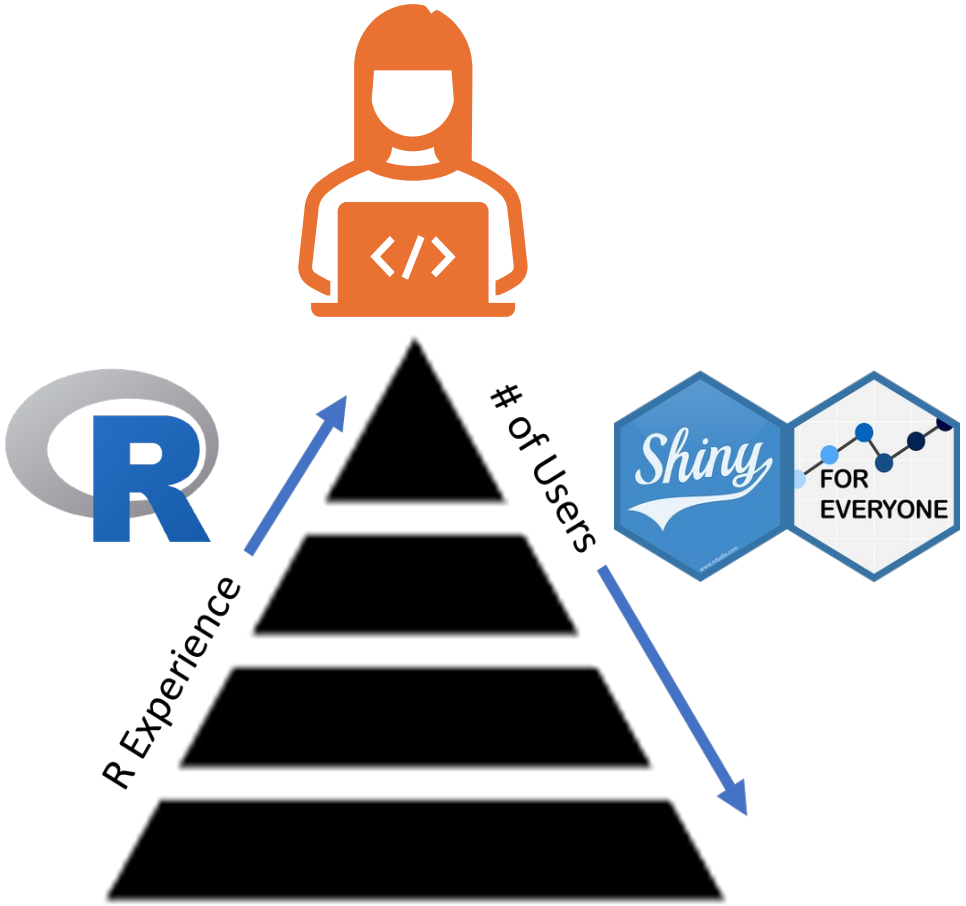
✓ Flexible design of TADA functions allows users to easily customize their existing workflows

✓ Lowers technical barriers to access and integrate disparate dataset together to support water analyses and research
✓ Centralized, interoperable and reusable data tools reduce government costs
✓ Supports users with variable R experience
✓ Discovering commonalities in assessment processes across the nation
✓ Sharing analysis processes, methods, R code, and knowledge
✓ User driven design
✓ Publicly available repository on GitHub is actively maintained and new functionality is being developed
✓ Visualizing data in interactive figures, maps, and tables



Open-source tools

- Free and transparent
  - Built in R programming language
  - Public GitHub repositories
    - *"a key factor in our success was the EPA's earlier decision to develop TADA using an open-source approach"*
- Different tools for different users
  - R package and R Shiny application(s)
- User driven design
  - Increases utility and use
  - Improves overall project quality



R package

EPATADA 0.0.1 Reference Articles

TADA Module 1: Water Quality Portal Data Discovery and Cleaning

TADA Team

2025-05-27

Source: vignettes/TADAModule1.Rmd

We will first use a subset of this example to demonstrate using new TADA_DataRetrieval options that allow for **spatial** or **tribe-specific** queries:

Focusing just on the "PUEBLO_POJOAQUE" organization, rerun the example above:

```
TADAProfile_single <- TADA_DataRetrieval(
  organization = "PUEBLO_POJOAQUE",
  startDate = "2018-01-01",
  endDate = "2019-01-01",
  applyautoclean = FALSE,
  ask = FALSE
)
```

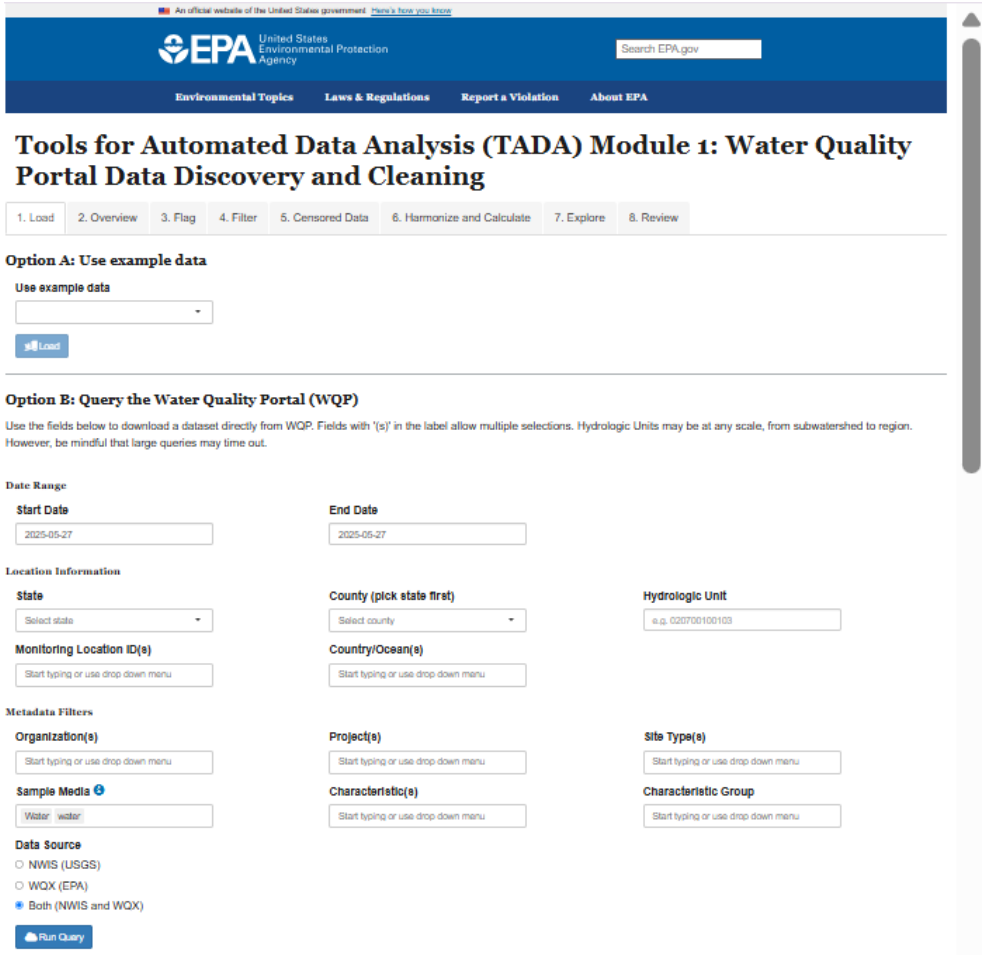
```
## [1] "Downloading WQP query results. This may take some time depending upon the query
## $startDate
## [1] "2018-01-01"
##
## $organization
## [1] "PUEBLO_POJOAQUE"
##
## $endDate
## [1] "2019-01-01"
```

The same results can now be obtained using a combination of the tribal_area_type and tribe_name_parcel arguments. Both must be used together. The tribal_area_type argument indicates which one of four layer datasets ("Alaska Native Allotments", "American Indian

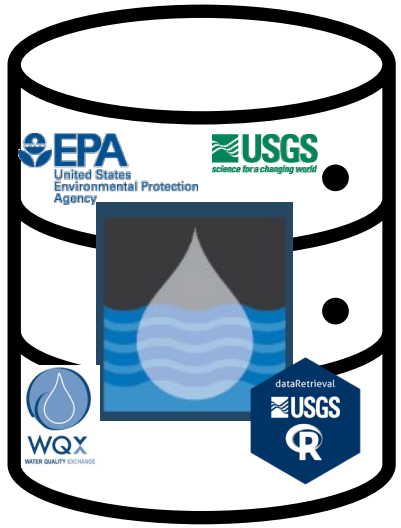
On this page

- Overview
- Install and load packages
- TADA_DataRetrieval
- USGS dataRetrieval
- Big Data Queries
- Filter data based on media type
- TADA_AutoClean
- Invalid coordinates
- Depth unit conversions
- Continuous (time series) data
- WQX Quality Assurance and Quality Control (QAQC) Service Result Flags
- WQX national upper and lower thresholds
- Potential duplicates
- Review QAPP information
- Full Dataframe Filtering
- Censored data
- Convert synonymous characteristic, fraction, speciation, and unit values to a consistent convention based on

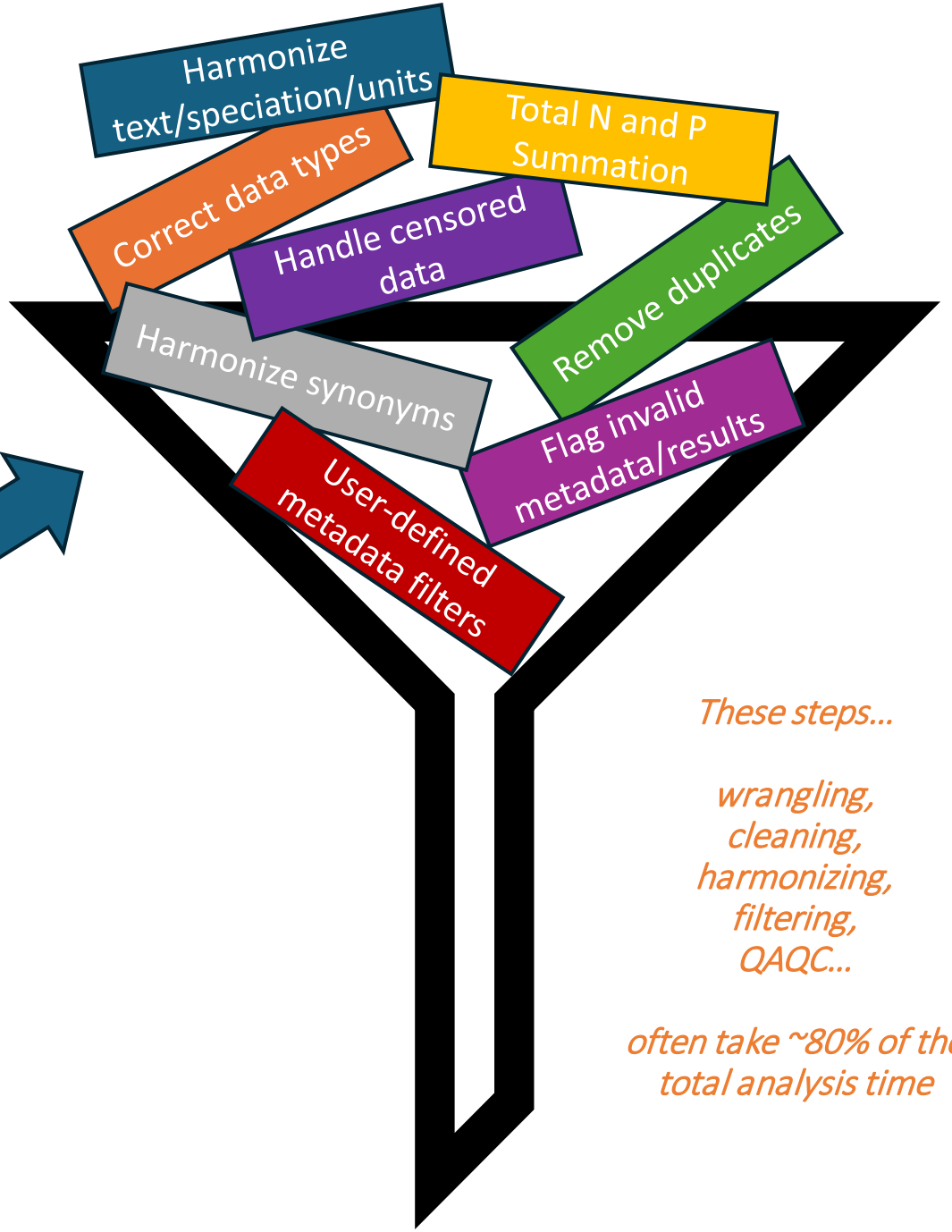
R Shiny app



Access and process WQP data



Construct a unified dataset containing key metadata



These steps…

wrangling, cleaning, harmonizing, filtering, QAQC…

often take ~80% of the total analysis time

- Is it of sufficient quality?
- Is it relevant?
- Does it include key metadata?
- Can it be harmonized and grouped in a way that makes sense?

TADA 0.0.1 Reference Articles

Contributing

2023-06-07

Source: vignettes/CONTRIBUTING.Rmd

Contribute to TADA!

We encourage you to read this project's CONTRIBUTING policy (you are here), its LICENSE, and its README.

We're so glad you're thinking about contributing to an EPA open source project! If you're unsure about anything, just ask — or submit your issue or pull request anyway. The worst that can happen is we'll politely ask you to change something. We appreciate all friendly contributions.

No matter who you are, if you spot an error, omission, or bug, you're welcome to open an issue in this repo!

TADA Working Group Mission

To share and develop R code for evaluating and visualizing Water Quality Portal (WQP) data more efficiently though collaboration and open-source programming. This includes working together to find commonalities in assessment processes across the nation, creating flexible tools that can be easily customized to work within existing workflows, supporting each other in learning R, and ensuring products will be accessible to organizations most in need.

Review required
At least 1 approving review is required by reviewers with write access. Learn more

Some checks haven't completed yet
5 successful and 1 in progress checks

- R-CMD-check / macos-latest (release) (pull_request) Successful in 52m
- test-coverage / test-coverage (pull_request) Successful in 10m
- R-CMD-check / windows-latest (release) (pull_request) Successful in 43m
- R-CMD-check / ubuntu-latest (devel) (pull_request) *In progress — This check*
- R-CMD-cl
- R-CMD-cl

Contributors 21

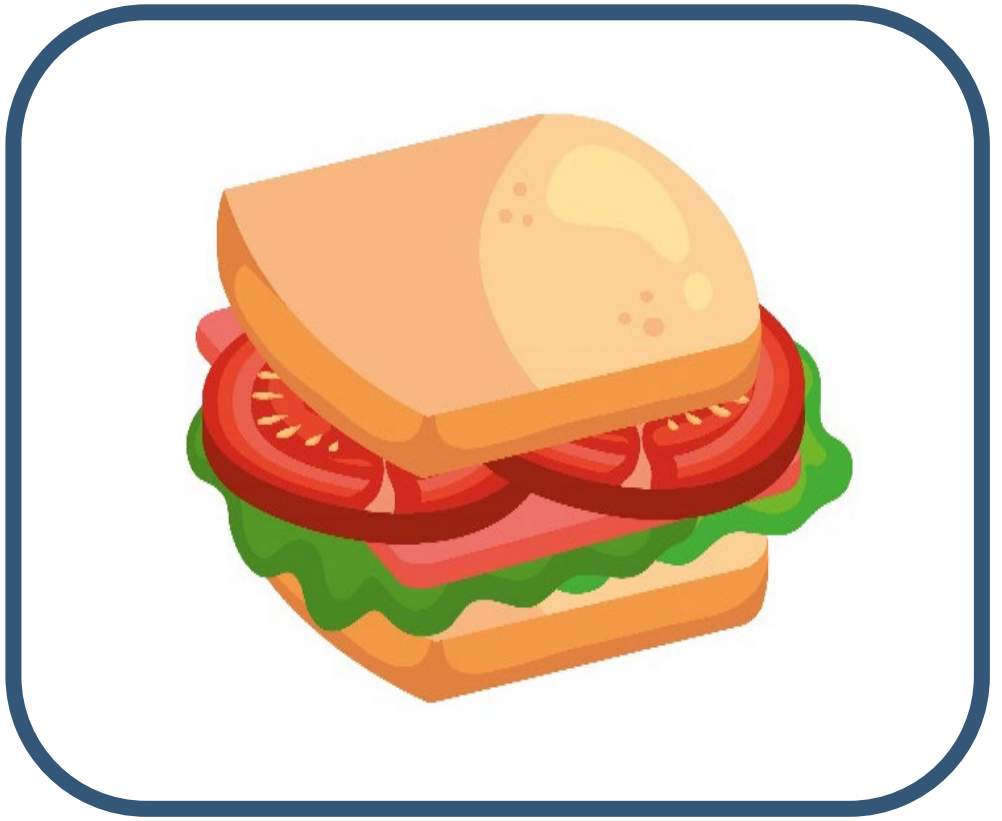
+ 7 contributors

TADA Working Group: An enthusiastic, engaged, open-source community

- Discovering commonalities in assessment processes
- Crowd sourcing solutions to analysis and assessment related data processing challenges
- Setting priorities for TADA development
- Facilitating knowledge sharing and code collaboration
- Established in 2020, meeting every other month

Flexible, reproducible, efficient

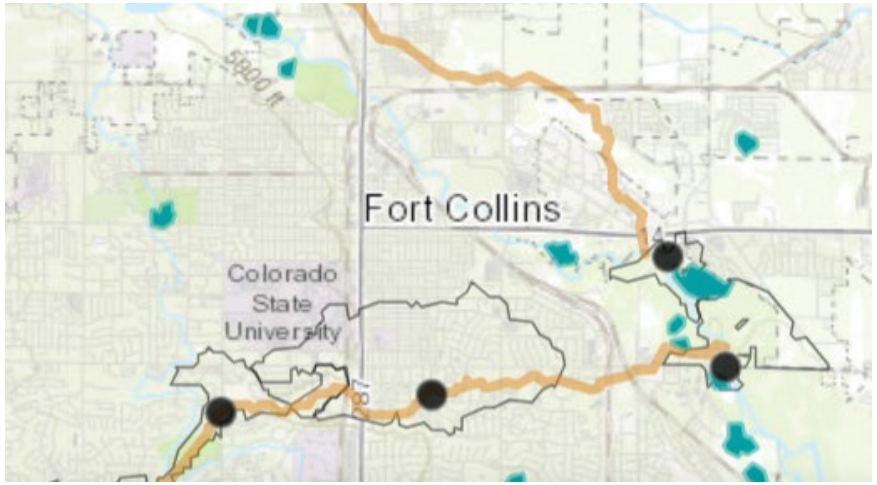
TADA Functions

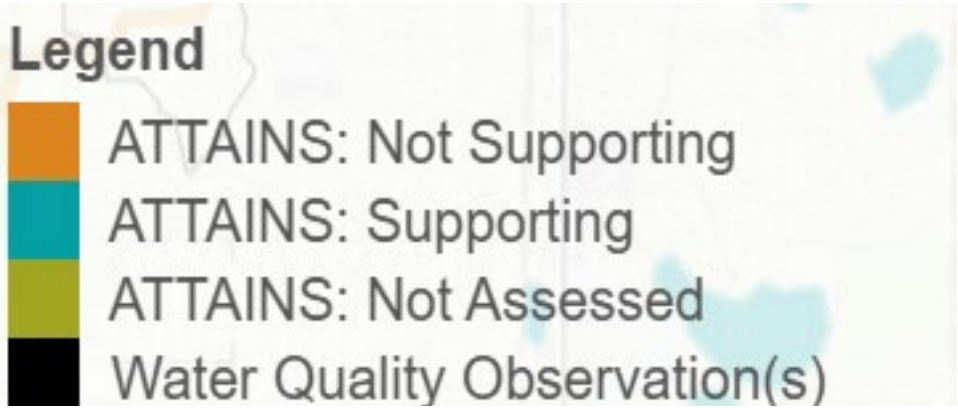


Your Analysis

Geospatial tools integrating ATTAINS and NHDPlusV2

- Assessment unit, site, and use associations
- Collaborated with ROSSyndicate Lab at Colorado State University
- rExpertQuery





Training video on new features: https://www.youtube.com/watch?v=7zzdGo58dvA

Challenges and limitations

- Bug fixes, maintenance, and improvements to existing functions
- Scale and scope creep
- Data and information availability
- System integration
- Changing staff, budget and priorities

Arsenic	Nickel
Boron	Nitrate
Chlorophyll a	Total Nitrogen, mixed forms
Chromium	pH
Chromium(VI)	Total Phosphorus, mixed forms
Cadmium	Depth, Secchi disk depth
Copper	Selenium
Dissolved oxygen (DO)	Silver
Dissolved oxygen saturation	Temperature, water
Escherichia coli	Total suspended solids
Lead	Chromium(III)
Mercury	Zinc

Opportunities for expansion

Criteria

Magnitude

Upper or lower limit

Range

Equation based

hardness, pH, or T dependent magnitude

common for ammonia and toxic metals

Duration

Assessment period

Characteristic specific assessment start and end dates

Temporal aggregation

n-day mean

n-day mean maximum or mean minimum

n-hour mean

geometric mean

arithmetic mean

n-day rolling average

Frequency

Never exceed

10% rule

1-in-3 years rule

applied using binomial test or percentile

Assessment methodologies & additional complexities

- Spatial aggregation (x, y, z)
  - Assessment unit and/or station level analysis
  - Sampling depth
- Acute vs chronic exposure
- Saltwater vs freshwater
- Water type
- Data requirements
  - Minimum number of samples
- Unique spatial variability
  - Ecoregion, warm vs. cold water fisheries, upstream vs. downstream
- Species-based, life-stage dependent
- Seasonality
- And more!

Popularity

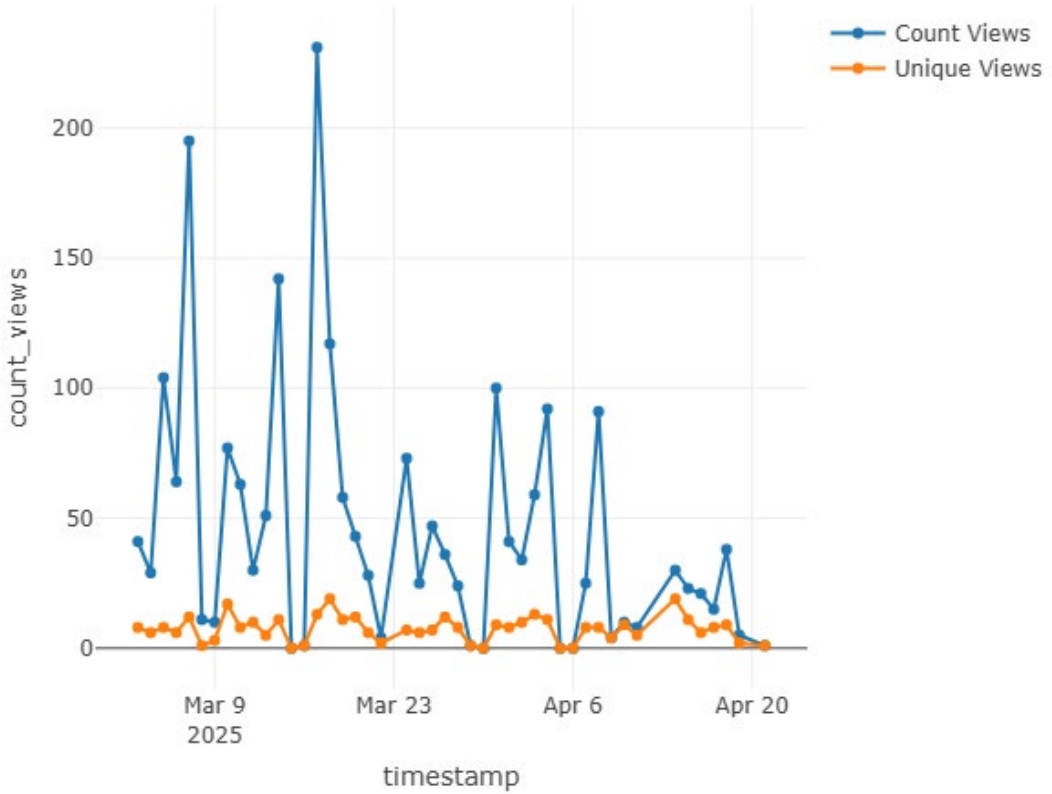
	Name	Rank	Description	Stars	Forks	Issues
1	EPATADA	1	This R package can be used to compile and evaluate Water Quality Portal (WQP) data for samples collected from surface water monitoring sites on streams and lakes. It can be used to create applications that support water quality programs and help states, tribes, and other stakeholders efficiently analyze the data.	52	23	76
2	CompTox-ToxCast-tcpl	2	US EPA's Toxicity Forecaster (ToxCast) Pipeline. More information on the ToxCast program available here: https://www.epa.gov/comptox-tools/toxicity-forecasting-toxcast	36	19	31
3	useeior	3	Estimating potential environmental impacts of goods and services in the US economy	36	29	13
4	StreamCatTools	4	R package to work with the StreamCat API within R and access the full suite of StreamCat catchment and watershed scale metrics for all NHDPlusV2 stream reaches and catchments.	35	9	17
5	LakeTrophicModelling	5	Applying data-mining approaches to predict lake health	31	17	0

EPATADA Public

This R package can be used to compile and evaluate Water Quality Portal (WQP) data for samples collected from surface water monitoring sites on streams and lakes. It can be used to create applications that support water quality programs and help states, tribes, and other stakeholders efficiently analyze the data.

R 52 CC0-1.0 23 74 (8 issues need help) 2

Updated 48 minutes ago



Broader Impacts

Use of TADA has potential to greatly reduce total government costs

- State and tribal agencies
- EPA regions
- EPA HQ
- USGS, other federal agencies

Efficient, transparent, and reproducible assessments

- Frees up time for other important tasks
- May facilitate assessing more waters
- Assists tribal onboarding to ATTAINS
- Helps discover and share commonalities in assessment processes nationally

Supports cross organization data integration

- Improves interoperability across WQX/WQP, ATTAINS, and the Criteria Search Tool (CST)
- Facilitates use of other organizations data in State or Tribal assessments
- Makes the WQX QAQC service available on the WQP side
- Helps find and address data quality issues in WQX/WQP

Thank you

Questions or suggestions?

TADA website

mullin.cristina@epa.gov

Related sessions this week – please remember to install R, RStudio, and required packages before coming!

- Beginner R user: introduction to using the **EPATADA R package** to retrieve, clean, analyze, and visualize WQP data
- How to use the Tools for Automated Data Analysis (TADA) **R Shiny app** to retrieve, clean, and visualize Water Quality Portal (WQP) data (no R experience required)
- rExpertQuery: How to access **ATTAINS information in R using Expert Query** web services