Coded Data Access with R

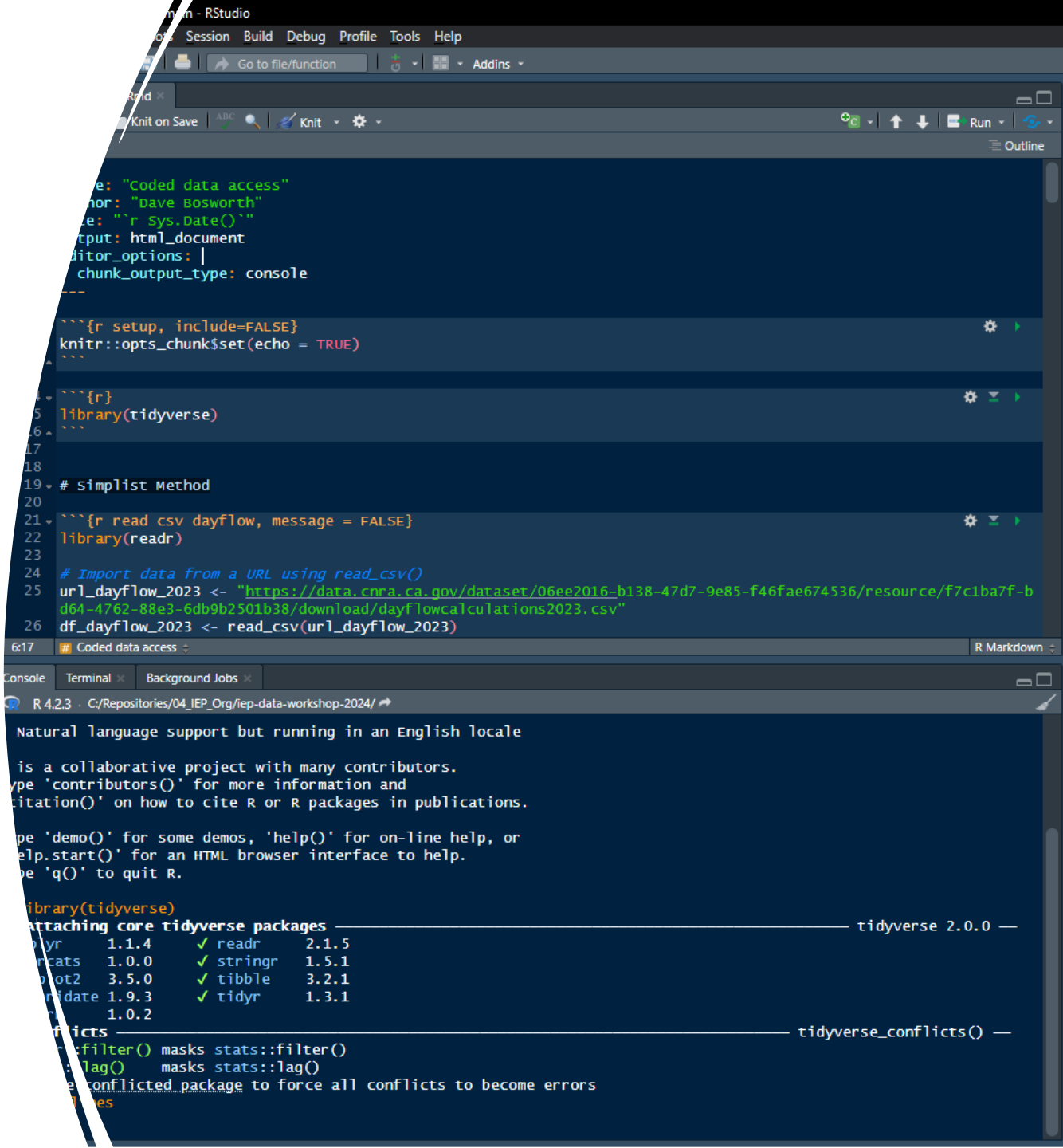
Dave Bosworth
CA Department of Water Resources (DWR)
Division of Integrated Science and Engineering (DISE)
Collaborative Science and Innovation Section (CSI)

What are we talking about here?

- Using code to import data into R from an **external** (non-local) and **open** source
- NOT: importing data from your hard drive, shared network, or SharePoint

Outline

- Simple Methods
- Dedicated R packages
- IEP Integrated datasets



Why bother?

Convenience

Efficiency

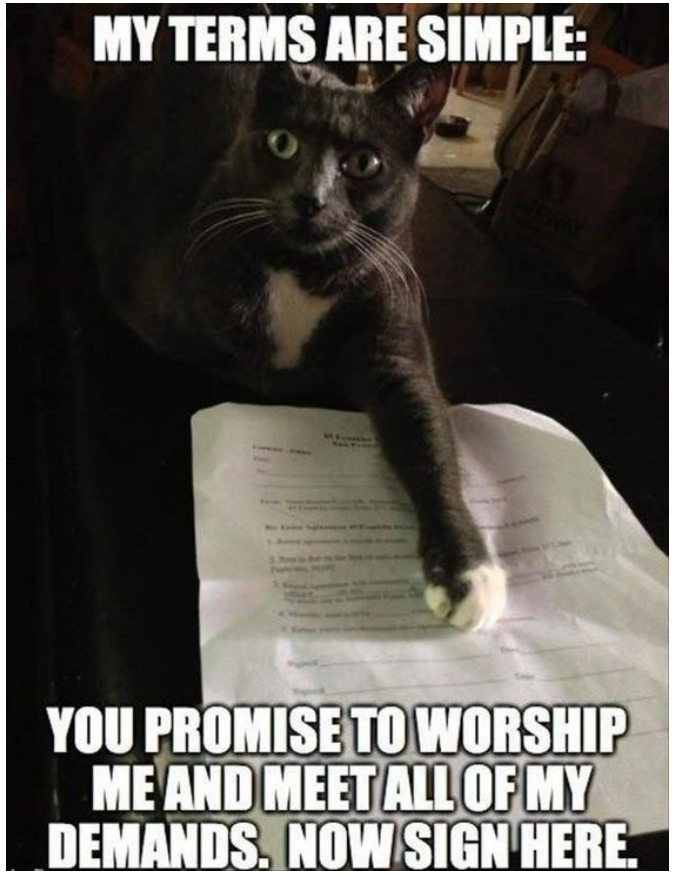
Reproducibility

Transportable

Versioning

Simple Methods

1) read_csv() or read.csv() with a URL
2) download.file() to tempdir(), then import



read_csv() with URL

CALIFORNIA NATURAL RESOURCES AGENCY

Open Data Organizations Topics Training

/ Organizations / California Department of... / Dayflow / **Dayflow Results 2023**

Dayflow Results 2023

URL: https://data.cnra.ca.gov/dataset/06ee2016-b138-47d7-9e85-f46fae674536/resource/f7c1ba7f-bd64-4762-88e3-6db9b2501b38/download/dayflowcalculations2023.csv

Dayflow results for Water Year 2022-2023.

Data Table

dayflowcalculations2023.csv

Add Filter

Show 10 entries

Showing 1 to 10 of 365 entries

_id	Year	Mo	Date	SAC	YOLO	CSMR	MOKE	MISC	SJR	EAST	TOT	CCC	SWP	CVP	NBAQ	EXPORTS
1	2022	10	10/1/2022	7210	35	25	88	150	212	476	7721	132	494	1821	33	2480
2	2022	10	10/2/2022	7470	36	26	0	145	218	389	7895	132	499	1815	26	2471

Use this workflow when URL points to a csv file

https://data.cnra.ca.gov/dataset/dayflow/resource/f7c1ba7f-bd64-4762-88e3-6db9b2501b38

Simple Methods

read_csv("url")

Copy URL for csv file

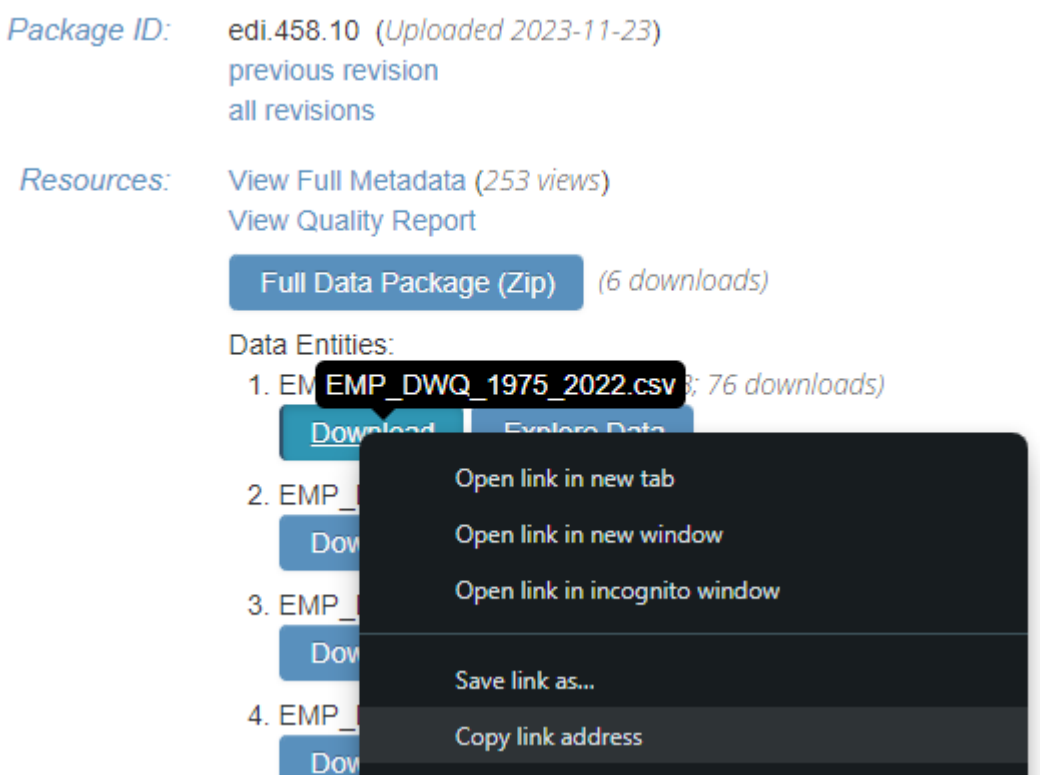
```r
library(readr)

# Import data from a URL using read_csv()
url_dayflow_2023 <- "https://data.cnra.ca.gov/dataset/06ee2016-b138-47d7-9e85-f46fae674536/resource/f7c1ba7f-bd64-4762-88e3-6db9b2501b38/download/dayflowcalculations2023.csv"
df_dayflow_2023 <- read_csv(url_dayflow_2023)
df_dayflow_2023
```

```
> df_dayflow_2023
# A tibble: 365 × 29
```

	Year	Mo	Date	SAC	YOLO	CSMR	MOKE	MISC	SJR	EAST	TOT	CCC	SWP	CVP	NBAQ
	<dbl>	<dbl>	<chr>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>	<dbl>
1	2022	10	10/1/20…	7210	35	25	88	150	212	476	7721	132	494	1821	33
2	2022	10	10/2/20…	7470	36	26	0	145	218	389	7895	132	499	1815	26
3	2022	10	10/3/20…	7550	36	26	0	147	245	418	8004	136	499	1809	25
4	2022	10	10/4/20…	7640	36	26	0	129	234	390	8065	128	498	1817	31
5	2022	10	10/5/20…	7590	36	26	0	93	222	341	7967	170	490	899	28
6	2022	10	10/6/20…	7440	35	26	0	63	213	301	7776	159	486	914	23
7	2022	10	10/7/20…	7460	33	25	0	50	216	291	7784	136	494	914	26
8	2022	10	10/8/20…	7200	33	24	0	59	248	331	7564	111	493	915	23
9	2022	10	10/9/20…	7010	32	24	0	66	308	398	7440	113	491	917	28
10	2022	10	10/10/2…	6870	32	24	0	71	305	400	7302	116	490	905	31

download.file() then import



May need to use this workflow with certain file types – xlsx, pdf, zip

https://portal.edirepository.org/nis/dataviewer?packageid=edi.458.10&entityid=cf231071093ac2861893793517db26f3

EMP discrete water quality EDI data package: https://portal.edirepository.org/nis/mapbrowse?packageid=edi.458.10

Simple Methods

download.file() then import

```r
# Import data from EDI to temporary directory
url_edi_emp_2022 <- "https://portal.edirepository.org/nis/dataviewer?packageid=edi.458.10&entityid=cf231071093ac2861893793517db26f3"
download.file(url_edi_emp_2022, file.path(tempdir(), "EMP_DWQ_1975_2022.csv"), mode = "wb")
```

```r
# Then import data using read_csv()
df_emp_2022 <- read_csv(file.path(tempdir(), "EMP_DWQ_1975_2022.csv"))
df_emp_2022
```

```
> df_emp_2022
# A tibble: 17,366 × 73
   Station Date       Time    SampleDescription Flag  FlagDescription FieldNotes Weather AirTemp
   <chr>   <date>     <time>  <chr>             <chr> <chr>           <chr>      <chr>     <dbl>
 1 D11     1975-01-07 13:00   NA                NA    NA              NA         NA         53.6
 2 D15     1975-01-07 14:00   NA                NA    NA              NA         NA         57.2
 3 D16     1975-01-07 16:00   NA                NA    NA              NA         NA         55.4
 4 D19     1975-01-07 15:00   NA                NA    NA              NA         NA         55.4
 5 D22     1975-01-07 13:00   NA                NA    NA              NA         NA         55.4
 6 D24     1975-01-07 14:00   NA                NA    NA              NA         NA         55.4
 7 D26     1975-01-07 15:00   NA                NA    NA              NA         NA         55.4
 8 D4      1975-01-07 12:00   NA                NA    NA              NA         NA         51.8
 9 D10     1975-01-08 13:00   NA                NA    NA              NA         NA         53.6
10 D12     1975-01-08 14:00   NA                NA    NA              NA         NA         55.4
```

Dedicated R packages

1) EDIutils – EDI data
2) dataRetrieval – USGS NWIS
3) cder – CDEC data

EDIutils

- REST API client for the Environmental Data Initiative (EDI)
- Allows for searching, accessing, and uploading data from R environment
- Package documentation: https://docs.ropensci.org/EDIutils/
- Available on CRAN

Dedicated R Packages

EDIutils

https://portal.edirepository.org/nis/simpleSearch

Terms used in this search: **emp**, **iep**

Displaying 1-10 of 48 matching data packages

<< < 1 2 3 4 5 > >>

Title	Creators	Publication Date	Package Id
Interagency Ecological Program: Benthic invertebrate monitoring in the Sacramento-San Joaquin Bay-Delta, collected by the Environmental Monitoring Program, 1975-2023.	Wells, Elizabeth Interagency Ecological Program	2024	edi.1036.4
Interagency Ecological Program: Discrete dissolved oxygen monitoring in the Stockton Deep Water Ship Channel, collected by the Environmental Monitoring Program, 1997-2018	Interagency Ecological Program (IEP) Lesmeister, Sarah Rinde, Jenna	2020	edi.276.2
Interagency Ecological Program: Discrete water quality monitoring in the Sacramento-San Joaquin Bay-Delta, collected by the Environmental Monitoring Program, 1975-2022	Battey, Morgan Perry, Sarah	2023	edi.458.10

```
library(EDIutils)
edi_scope <- "edi"
edi_emp_id <- 458
```

list_data_package_revisions()

```r
library(EDIutils)
edi_scope <- "edi"
edi_emp_id <- 458
```

```r
list_data_package_revisions(scope = edi_scope, identifier = edi_emp_id)
```

```
[1]  1  2  3  4  5  6  7  8  9 10
```

```r
edi_emp_rev <- list_data_package_revisions(
  scope = edi_scope,
  identifier = edi_emp_id,
  filter = "newest"
)
edi_emp_pid <- paste(edi_scope, edi_emp_id, edi_emp_rev, sep = ".")
edi_emp_pid
```

```
> edi_emp_pid
[1] "edi.458.10"
```

1. edi.458.1 (*Uplooded 2020-01-06*)
2. edi.458.2 (*Uplooded 2020-01-27*)
3. edi.458.3 (*Uplooded 2020-10-02*)
4. edi.458.4 (*Uplooded 2021-02-17*)
5. edi.458.5 (*Uplooded 2022-06-01*)
6. edi.458.6 (*Uplooded 2022-06-16*)
7. edi.458.7 (*Uplooded 2022-08-11*)
8. edi.458.8 (*Uplooded 2023-04-04*)
9. edi.458.9 (*Uplooded 2023-05-26*)
10. edi.458.10 (*Uplooded 2023-11-23*)

read_data_entity_names()

```
edi_emp_pid <- paste(edi_scope, edi_emp_id, edi_emp_rev, sep = ".")
edi_emp_pid
```

```
> edi_emp_pid
[1] "edi.458.10"
```

```
df_edi_emp_ent <- read_data_entity_names(packageId = edi_emp_pid)
df_edi_emp_ent
```

```
> df_edi_emp_ent
                          entityId                 entityName
1 cf231071093ac2861893793517db26f3          EMP_DWQ_1975_2022
2 86dd696bc3f8407ff52954094e1e9dcf EMP_DWQ_Stations_1975-2022
3 afc5b55a61e9a16d29fcaef4d802f5be          EMP_DWQ_FlagCodes
4 b399c042c893809547dc196a762b929f   EMP_DWQ_metadata_methods
```

Package ID: edi.458.10 (*Uploaded 2023-11-23*)
previous revision
all revisions

Resources: View Full Metadata (*283 views*)
View Quality Report
Full Data Package (Zip) (*8 downlo*

Data Entities:

1. EMP_DWQ_1975_2022 (*4.3 MiB; 1*
   Download Explore Data
2. EMP_DWQ_Stations_1975-2022 (*5*
   Download Explore Data
3. EMP_DWQ_FlagCodes (*183 B; 37*
   Download Explore Data
4. EMP_DWQ_metadata_methods (*1.*
   Download

read_data_entity()

```
edi_emp_pid <- paste(edi_scope, edi_emp_id, edi_emp_rev, sep = ".")
edi_emp_pid
```

```
> edi_emp_pid
[1] "edi.458.10"
```

```
df_edi_emp_ent <- read_data_entity_names(packageId = edi_emp_pid)
df_edi_emp_ent
```

```
> df_edi_emp_ent
                                entityId                 entityName
1 cf231071093ac2861893793517db26f3          EMP_DWQ_1975_2022
2 86dd696bc3f8407ff52954094e1e9dcf EMP_DWQ_Stations_1975-2022
3 afc5b55a61e9a16d29fcaef4d802f5be           EMP_DWQ_FlagCodes
4 b399c042c893809547dc196a762b929f    EMP_DWQ_metadata_methods
```

```
edi_emp_ent_id <- df_edi_emp_ent %>%
  filter(entityName == "EMP_DWQ_1975_2022") %>%
  pull(entityId)

raw_emp_2022_edi <- read_data_entity(packageId = edi_emp_pid, entityId = edi_emp_ent_id)
```

```
df_emp_2022_edi <- read_csv(raw_emp_2022_edi)
df_emp_2022_edi
```

```
> df_emp_2022_edi
# A tibble: 17,366 x 73
   Station Date       Time    SampleDescription Flag  FlagDescription FieldNotes Weather AirTemp
   <chr>   <date>     <time>  <chr>             <chr> <chr>           <chr>      <chr>     <dbl>
 1 D11     1975-01-07 13:00   NA                NA    NA              NA         NA         53.6
 2 D15     1975-01-07 14:00   NA                NA    NA              NA         NA         57.2
 3 D16     1975-01-07 16:00   NA                NA    NA              NA         NA         55.4
 4 D19     1975-01-07 15:00   NA                NA    NA              NA         NA         55.4
 5 D22     1975-01-07 13:00   NA                NA    NA              NA         NA         55.4
 6 D24     1975-01-07 14:00   NA                NA    NA              NA         NA         55.4
```

NOTE:
read_data_entity() imports data as raw bytes, need to use a reader function with it

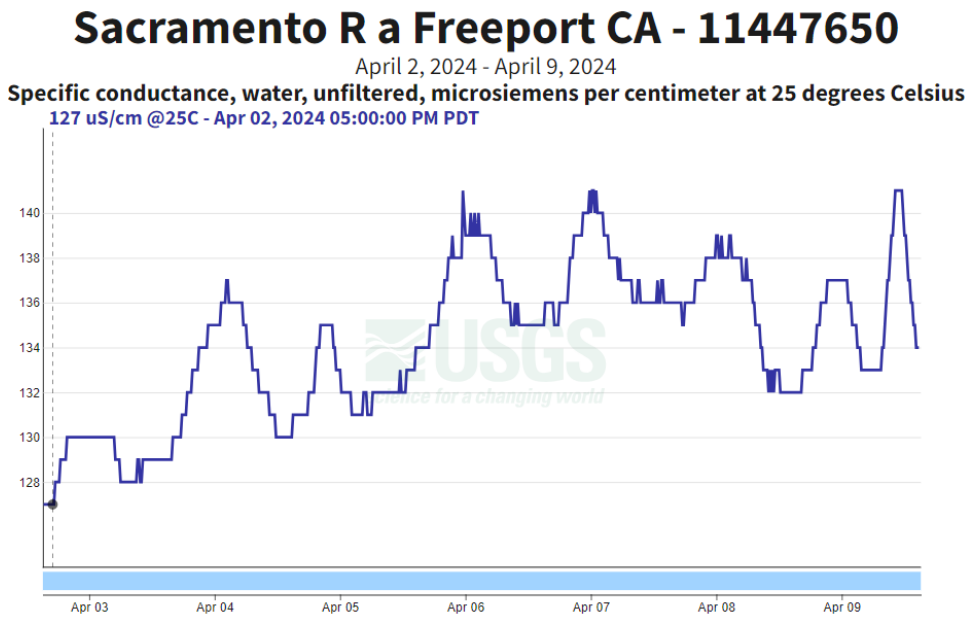
dataRetrieval

- Allows for loading USGS data into the R environment
- NWIS and WQP web services
- Package documentation: https://doi-usgs.github.io/dataRetrieval/
- Available on CRAN

Dedicated R Packages - dataRetrieval

whatNWISsites()

Provides site information from NWIS




```r
library(dataRetrieval)

df_srf_sta_info <- whatNWISsites(sites = "11447650")
glimpse(df_srf_sta_info)
```

```
> glimpse(df_srf_sta_info)
Rows: 1
Columns: 8
$ agency_cd    <chr> "USGS"
$ site_no      <chr> "11447650"
$ station_nm   <chr> "SACRAMENTO R A FREEPORT CA"
$ site_tp_cd   <chr> "ST"
$ dec_lat_va   <dbl> 38.45566
$ dec_long_va  <dbl> -121.5016
$ colocated    <lgl> FALSE
$ queryTime    <dttm> 2024-04-16 13:16:38
```

USGS Water Data: https://waterdata.usgs.gov/monitoring-location/11447650/

whatNWISdata()

Provides data availability for a site from NWIS

```r
df_srf_uv_data_avail <- whatNWISdata(siteNumber = "11447650", service = "uv")
glimpse(df_srf_uv_data_avail)
```

```
> glimpse(df_srf_uv_data_avail)
Rows: 15
Columns: 24
$ agency_cd          <chr> "USGS", "USGS", "USGS", "USGS", "USGS", "USGS", "USGS", "USGS", "USGS", "USGS", "USGS", …
$ site_no            <chr> "11447650", "11447650", "11447650", "11447650", "11447650", "11447650", "11447650", "114…
$ station_nm         <chr> "SACRAMENTO R A FREEPORT CA", "SACRAMENTO R A FREEPORT CA", "SACRAMENTO R A FREEPORT CA"…
$ site_tp_cd         <chr> "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST", "ST"
$ dec_lat_va         <dbl> 38.45566, 38.45566, 38.45566, 38.45566, 38.45566, 38.45566, 38.45566, 38.45566, 38.45566…
$ dec_long_va        <dbl> -121.5016, -121.5016, -121.5016, -121.5016, -121.5016, -121.5016, -121.5016, -121.5016, …
$ coord_acy_cd       <chr> "5", "5", "5", "5", "5", "5", "5", "5", "5", "5", "5", "5", "5", "5", "5"
$ dec_coord_datum_cd <chr> "NAD83", "NAD83", "NAD83", "NAD83", "NAD83", "NAD83", "NAD83", "NAD83", "NAD83", "NAD83"…
$ alt_va             <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA
$ alt_acy_va         <dbl> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA
$ alt_datum_cd       <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA
$ huc_cd             <chr> "18020163", "18020163", "18020163", "18020163", "18020163", "18020163", "18020163", "180…
$ data_type_cd       <chr> "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv", "uv"
$ parm_cd            <chr> "00010", "00010", "00060", "00065", "00095", "00300", "00301", "00400", "00480", "32295"…
$ stat_cd            <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA
$ ts_id              <chr> "15731", "15760", "236032", "15738", "236034", "15757", "259962", "15735", "257748", "24…
$ loc_web_ds         <chr> "[Right Bank Pump Stand]", "BGC PROJECT, [East Fender]", NA, "[PRIMARY]", "BGC PROJECT, …
$ medium_grp_cd      <chr> "wat", "wat", "wat", "wat", "wat", "wat", "wat", "wat", "wat", "wat", "wat", "wat", "wat…
$ parm_grp_cd        <chr> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA
$ srs_id             <chr> "1645597", "1645597", "1645423", "17164583", "1646694", "154302", "154302", "17028275", …
$ access_cd          <chr> "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0", "0"
$ begin_date         <date> 2007-10-01, 2013-08-30, 2007-10-01, 2007-10-01, 2013-08-30, 2013-08-30, 2013-08-30, 2013…
$ end_date           <date> 2024-04-09, 2024-04-09, 2024-04-09, 2024-04-09, 2024-04-09, 2024-04-09, 2024-04-09, 2024…
$ count_nu           <dbl> 6035, 3875, 6035, 6035, 3875, 3875, 3875, 3875, 3875, 1095, 3875, 1398, 3875, 3112, 3113
```

USGS Parameter Code → parm_cd

Period of record, Sample count → begin_date, end_date, count_nu

parameterCdFile

Complete list of USGS parameter codes

```
as_tibble(parameterCdFile)
```

```
> as_tibble(parameterCdFile)
# A tibble: 24,200 × 6
   parameter_cd parameter_group_nm parameter_nm                                          casrn srsname parameter_units
   <chr>        <chr>              <chr>                                                 <chr> <chr>   <chr>
 1 00001        Information        Location in cross section, distance from right bank… " "   " "     ft
 2 00002        Information        Location in cross section, distance from right bank… " "   " "     %
 3 00003        Information        Sampling depth, feet                                  " "   " "     ft
 4 00004        Physical           Stream width, feet                                    " "   "Instr… ft
 5 00005        Information        Location in cross section, fraction of total depth,… " "   " "     %
 6 00008        Information        Sample accounting number                              " "   " "     nu
 7 00009        Information        Location in cross section, distance from left bank … " "   " "     ft
 8 00010        Physical           Temperature, water, degrees Celsius                   " "   "Tempe… deg C
 9 00011        Physical           Temperature, water, degrees Fahrenheit                " "   "Tempe… deg F
10 00012        Physical           Evaporation temperature, 48 inch pan, degrees Celsi… " "   " "     deg C
# i 24,190 more rows
```

Parameter codes:
https://help.waterdata.usgs.gov/codes-and-parameters/parameters

parameterCdFile - filtered

```
> df_srf_uv_parm_cd
# A tibble: 13 × 3
   parameter_cd parameter_nm                                                                  parameter_units
   <chr>        <chr>                                                                         <chr>
 1 00010        Temperature, water, degrees Celsius                                           deg C
 2 00060        Discharge, cubic feet per second                                              ft3/s
 3 00065        Gage height, feet                                                             ft
 4 00095        Specific conductance, water, unfiltered, microsiemens per centimeter at 25 degrees C… uS/cm @25C
 5 00300        Dissolved oxygen, water, unfiltered, milligrams per liter                     mg/l
 6 00301        Dissolved oxygen, water, unfiltered, percent of saturation                    % saturatn
 7 00400        pH, water, unfiltered, field, standard units                                  std units
 8 00480        Salinity, water, unfiltered, parts per thousand                               ppth
 9 32295        Dissolved organic matter fluorescence (fDOM), water, in situ, concentration estimate… ug/l QSE
10 32316        Chlorophyll fluorescence (fChl), water, in situ, concentration estimated from refere… ug/l
11 63680        Turbidity, water, unfiltered, monochrome near infra-red LED light, 780-900 nm, detec… FNU
12 72137        Discharge, tidally filtered, cubic feet per second                            ft3/s
13 72255        Mean water velocity for discharge computation, feet per second                ft/sec
```

Dedicated R Packages - dataRetrieval

readNWISuv()

Import instantaneous data ("uv") for one or more stations and parameters from NWIS

```
df_srf_spc_inst <- readNWISuv(
  siteNumbers = "11447650",
  parameterCd = "00095",
  startDate = "2023-01-01", endDate = "2023-12-31",
  tz = "America/Los_Angeles"
)
as_tibble(df_srf_spc_inst)
```

```
> as_tibble(df_srf_spc_inst)
# A tibble: 34,216 x 6
   agency_cd site_no  dateTime            X_00095_00000 X_00095_00000_cd tz_cd
   <chr>     <chr>    <dttm>                      <dbl> <chr>            <chr>
 1 USGS      11447650 2023-01-01 00:00:00           122 P                America/Los_Angeles
 2 USGS      11447650 2023-01-01 00:15:00           121 P                America/Los_Angeles
 3 USGS      11447650 2023-01-01 00:30:00           121 P                America/Los_Angeles
 4 USGS      11447650 2023-01-01 00:45:00           120 P                America/Los_Angeles
 5 USGS      11447650 2023-01-01 01:00:00           120 P                America/Los_Angeles
 6 USGS      11447650 2023-01-01 01:15:00           120 P                America/Los_Angeles
 7 USGS      11447650 2023-01-01 01:30:00           119 P                America/Los_Angeles
 8 USGS      11447650 2023-01-01 01:45:00           119 P                America/Los_Angeles
 9 USGS      11447650 2023-01-01 02:00:00           118 P                America/Los_Angeles
10 USGS      11447650 2023-01-01 02:15:00           117 P                America/Los_Angeles
# i 34,206 more rows
# i Use `print(n = ...)` to see more rows
```

readNWISdv() –
Import daily data ("dv") from NWIS

Same arguments with the addition of 'statCd':
'00001' – daily maximum
'00002' – daily minimum
'00003' – daily mean (default)
'00006' – daily sum
'00008' – daily median
https://help.waterdata.usgs.gov/code/stat_cd_nm_query?stat_nm_cd=%25&fmt=html

WQP Web Services

1) whatWQPsites() – Station Information

```
whatWQPsites(siteid = "USGS-11447650")
```

2) readWQPsummary() – Data availability

```
df_srf_nutr_data_avail <- readWQPsummary(
  siteid = "USGS-11447650",
  characteristicType = "Nutrient"
)
```

3) readWQPqw() – Import data

```
df_srf_nitrate <- readWQPqw(
  siteNumbers = "USGS-11447650",
  parameterCd = "Nitrate",
  startDate = "2023-01-01", endDate = "2023-12-31",
  tz = "America/Los_Angeles"
)
```

NOTE:
WQP functions require
“USGS-” prefix for site codes

cder

- Web API client for the California Data Exchange Center (CDEC)
- Allows for importing data into the R environment
- Package documentation: https://hydroecology.net/cder/index.html
- Available on CRAN

Dedicated R Packages - cder

cdec_meta()

Station information from CDEC

```
library(cder)
cdec_meta(station = "FPT")
```

SACRAMENTO RIVER AT FREEPORT

Map of surrounding area

Station ID	FPT	**Elevation**	0 ft
River Basin	SACRAMENTO RIVER	**County**	SACRAMENTO
Hydrologic Area	SACRAMENTO RIVER	**Nearby City**	FREEPORT
Latitude	38.456112°	**Longitude**	-121.500300°
Operator	US Geological Survey	**Maintenance**	None Specified

The following data types are available online. Select one of the links below to retrieve recent data.

Sensor Description	Sensor Number	Duration	Plot	Data Collection	Data Available
FLOW, RIVER DISCHARGE, CFS	**20**	(daily)	(FLOW)	COMPUTED	10/01/1948 to present
WATER, TURBIDITY FNU, FNU	**221**	(daily)	(TURB WF)	COMPUTED	12/04/2009 to present
FLOW, RIVER DISCHARGE, CFS	**20**	(event)	(FLOW)	DATA XCHG-USGS	10/29/2004 to present
RIVER STAGE, FEET	**1**	(event)	(RIV STG)	DATA XCHG-USGS	05/06/2013 to present
WATER, TURBIDITY FNU, FNU	**221**	(event)	(TURB WF)	DATA XCHG-USGS	12/03/2009 to present
WATER, VELOCITY, FT/SEC	**21**	(event)	(VLOCITY)	DATA XCHG-USGS	05/06/2013 to present

Dedicated R Packages - cder

cdec_query()

Import data for one or more stations and sensors from CDEC

Options for ‘durations’ argument:
‘E’ – event
‘H’ – hourly
‘D’ – daily
‘M’ – monthly

```r
df_fpt_turb <- cdec_query(
  stations = "FPT", sensors = 221, durations = "E",
  start.date = "2024-03-01", end.date = "2024-03-31"
)
df_fpt_turb
```

```
> df_fpt_turb
# A tibble: 2,877 x 9
   StationID Duration SensorNumber SensorType DateTime            ObsDate             Value DataFlag SensorUnits
   <chr>     <chr>           <int> <chr>      <dttm>              <dttm>              <dbl> <chr>    <chr>
 1 FPT       E                 221 TURB WF    2024-03-01 00:00:00 2024-03-01 00:00:00  23.2 ""       FNU
 2 FPT       E                 221 TURB WF    2024-03-01 00:15:00 2024-03-01 00:15:00  22.7 ""       FNU
 3 FPT       E                 221 TURB WF    2024-03-01 00:30:00 2024-03-01 00:30:00  22.9 ""       FNU
 4 FPT       E                 221 TURB WF    2024-03-01 00:45:00 2024-03-01 00:45:00  23.3 ""       FNU
 5 FPT       E                 221 TURB WF    2024-03-01 01:00:00 2024-03-01 01:00:00  23.5 ""       FNU
 6 FPT       E                 221 TURB WF    2024-03-01 01:15:00 2024-03-01 01:15:00  22.9 ""       FNU
 7 FPT       E                 221 TURB WF    2024-03-01 01:30:00 2024-03-01 01:30:00  23.1 ""       FNU
 8 FPT       E                 221 TURB WF    2024-03-01 01:45:00 2024-03-01 01:45:00  24.2 ""       FNU
 9 FPT       E                 221 TURB WF    2024-03-01 02:00:00 2024-03-01 02:00:00  23.3 ""       FNU
10 FPT       E                 221 TURB WF    2024-03-01 02:15:00 2024-03-01 02:15:00  24.1 ""       FNU
# i 2,867 more rows
```

IEP Integrated datasets

Interagency Ecological Program

1) deltafish
2) zooper
3) discretewq
4) deltamapr

Need to use “devtools::install_github()” function to install these packages

Also, for Windows computers, need to have RTools installed: https://cran.r-project.org/bin/windows/Rtools/

deltafish

- Provides easy query access to the very large published EDI dataset of IEP fish abundance and length data
- 1959-2021 from 9 surveys: Bay Study, FMWT, EDSM, 20mm, SLS, SKT, DJFMP, Suisun Marsh, STN
- Available on GitHub: https://github.com/Delta-Stewardship-Council/deltafish
- EDI data repository: https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=1075

Build and load database

1) Build and cache database – set 'update' to TRUE to rebuild to latest version

```
# install.packages("devtools")
# devtools::install_github("Delta-Stewardship-Council/deltafish")
library(deltafish)
# Build the database - this takes a while, use update = TRUE to
  # re-build cached database
create_fish_db()
```

2) Open two data files in database

```
# Open two data files
surv <- open_survey()
fish <- open_fish()
```

Build and run query

3) Build query by using ‘dplyr’ functions

```
# Filter for sources and taxa of interest and join them together
surv_FMWT <- surv %>% filter(Source == "FMWT") %>% select(SampleID, Date)

fish_smelt <- fish %>%
  filter(Taxa %in% c("Dorosoma petenense", "Morone saxatilis", "Spirinchus thaleichthys"))

df_fish <- left_join(surv_FMWT, fish_smelt)
```

4) Run query and import data into R workspace using ‘dplyr::collect()’

```
# Collect the resulting data frame - collect executes
  # the SQL query and gives you a table
df_fish_c <- collect(df_fish)
df_fish_c
```

```
> df_fish_c
# A tibble: 175,009 x 6
   SampleID Date       Length Count Notes_catch Taxa
   <chr>    <date>      <dbl> <dbl> <chr>       <chr>
 1 FMWT 1   1992-01-10     NA     0 NA          Spirinchus thaleichthys
 2 FMWT 1   1992-01-10     NA     0 NA          Morone saxatilis
 3 FMWT 1   1992-01-10     NA     0 NA          Dorosoma petenense
 4 FMWT 2   1992-02-07     NA     0 NA          Spirinchus thaleichthys
 5 FMWT 2   1992-02-07     NA     0 NA          Morone saxatilis
 6 FMWT 2   1992-02-07     NA     0 NA          Dorosoma petenense
 7 FMWT 3   1992-03-18     NA     0 NA          Spirinchus thaleichthys
 8 FMWT 3   1992-03-18     NA     0 NA          Morone saxatilis
 9 FMWT 3   1992-03-18     NA     0 NA          Dorosoma petenense
10 FMWT 4   1992-09-16     NA     0 NA          Spirinchus thaleichthys
# i 174,999 more rows
```

zooper

- Allows for downloading and integrating IEP zooplankton data
- 1972-2021 from 7 surveys: 20mm, DOP, FRP, EMP, STN/FMWT, YBFMP, IEP zooplankton study
- Available on GitHub: https://github.com/InteragencyEcologicalProgram/zooper
- EDI data repository: https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=539

IEP Integrated datasets - zooper

Zoopsynther()

```r
# install.packages("devtools")
# devtools::install_github("InteragencyEcologicalProgram/zooper")
library(zooper)

df_zoop <- Zoopsynther(
  Data_type = "Community", Response = c("CPUE", "BPUE"),
  Sources = c("EMP", "FRP", "FMWT"), Size_class = "Meso",
  Date_range = c("1990-10-01", "2000-09-30")
)
df_zoop
```

```
> df_zoop
# A tibble: 151,478 × 35
   Source SizeClass Volume Lifestage Taxname  Phylum Class Order Family Genus Species Taxlifestage SampleID  CPUE
   <chr>  <chr>      <dbl> <chr>     <chr>    <chr>  <chr> <chr> <chr>  <chr> <chr>   <chr>        <chr>    <dbl>
 1 EMP    Meso        10.6 Adult     Acantho… Arthr… Cope… Cycl… Cyclo… Acan… NA      Acanthocycl… EMP NZE… 11.3
 2 EMP    Meso        10.6 Adult     Acartia… Arthr… Cope… Cala… Acart… Acar… NA      Acartia_UnI… EMP NZE…  1.89
 3 EMP    Meso        10.6 Adult     Acartie… Arthr… Cope… Cala… Acart… Acar… Acarti… Acartiella … EMP NZE…  5.67
 4 EMP    Meso        10.6 Adult     Asplanc… Rotif… Euro… Ploi… Aspla… Aspl… NA      Asplanchna_… EMP NZE…  0
 5 EMP    Meso        10.6 Adult     Bosmina… Arthr… Bran… Clad… Bosmi… Bosm… Bosmin… Bosmina lon… EMP NZE…  0
 6 EMP    Meso        10.6 Adult     Calanoi… Arthr… Cope… Cala… NA     NA    NA      Calanoida_U… EMP NZE…  0
 7 EMP    Meso        10.6 Adult     Cladoce… Arthr… Bran… Clad… NA     NA    NA      Cladocera_U… EMP NZE…  7.56
 8 EMP    Meso        10.6 Adult     Cyclopo… Arthr… Cope… Cycl… NA     NA    NA      Cyclopoida_… EMP NZE… 49.2
 9 EMP    Meso        10.6 Adult     Daphnia… Arthr… Bran… Clad… Daphn… Daph… NA      Daphnia_UnI… EMP NZE…  5.67
10 EMP    Meso        10.6 Adult     Diaphan… Arthr… Bran… Clad… Sidid… Diap… NA      Diaphanosom… EMP NZE…  0
# i 151,468 more rows
# i 21 more variables: BPUE <dbl>, Undersampled <lgl>, Date <dttm>, Station <chr>, Chl <dbl>, Secchi <dbl>,
#   Temperature <dbl>, BottomDepth <dbl>, Tide <chr>, TowType <chr>, Datetime <dttm>, Turbidity <dbl>, pH <dbl>,
#   DO <dbl>, Microcystis <chr>, Year <dbl>, AmphipodCode <chr>, SalSurf <dbl>, SalBott <dbl>, Latitude <dbl>,
#   Longitude <dbl>
# i Use `print(n = ...)` to see more rows
```

Use ‘Data_type’ argument to choose between two approaches to resolving differences in taxonomic resolution:

‘Taxa’ - all available data on given Taxa

‘Community’ - to conduct a community analysis

discretewq

- Provides an integrated dataset of IEP water quality data
- 1959-2022 from 16 surveys: Bay Study, FMWT, EDSM, 20mm, SLS, SKT, SDO, EMP, YBFMP, DJFMP, USBR-SDWSC, Suisun Marsh, STN, USGS-SFBS, USGS-CAWSC, DWR-NCRO
- Available on GitHub: https://github.com/InteragencyEcologicalProgram/discretewq
- EDI data repository: https://portal.edirepository.org/nis/mapbrowse?scope=edi&identifier=731

IEP Integrated datasets -discretewq

wq()

```r
# install.packages("devtools")
# devtools::install_github("InteragencyEcologicalProgram/discretewq")
library(discretewq)

df_dwq <- wq(
  Sources = c("EMP", "NCRO", "USGS_CAWSC", "USGS_SFBS"),
  Start_year = 2020, End_year = 2022
)
df_dwq
```

```
> df_dwq
# A tibble: 5,176 x 78
   Source Station Latitude Longitude Field_coords Date                Datetime            Notes       Depth Tide
   <chr>  <chr>      <dbl>     <dbl> <lgl>        <dttm>              <dttm>              <chr>       <dbl> <chr>
 1 EMP    C10A        37.7     -121. FALSE        2020-01-14 00:00:00 2020-01-14 12:45:00 Station to…  1.49 High…
 2 EMP    C3A         38.4     -122. FALSE        2020-01-14 00:00:00 2020-01-14 09:00:00 NA           4.97 High…
 3 EMP    C9          37.8     -122. FALSE        2020-01-14 00:00:00 2020-01-14 10:55:00 NA           4.75 High…
 4 EMP    D12         38.0     -122. FALSE        2020-01-15 00:00:00 2020-01-15 08:45:00 Fluorescen… 11.6  High…
 5 EMP    D19         38.0     -122. FALSE        2020-01-15 00:00:00 2020-01-15 10:00:00 Bottom mea… 10.5  High…
 6 EMP    D28A        38.0     -122. FALSE        2020-01-15 00:00:00 2020-01-15 10:35:00 NA           5.91 High…
 7 EMP    D16         38.1     -122. FALSE        2020-01-16 00:00:00 2020-01-16 08:15:00 NA          14.0  High…
 8 EMP    D26         38.1     -122. FALSE        2020-01-16 00:00:00 2020-01-16 09:05:00 NA          13.0  High…
 9 EMP    MD10A       38.0     -121. FALSE        2020-01-16 00:00:00 2020-01-16 10:20:00 NA           3.60 High…
10 EMP    P8          38.0     -121. FALSE        2020-01-16 00:00:00 2020-01-16 11:50:00 NA           8.81 High…
# i 5,166 more rows
# i 68 more variables: Microcystis <dbl>, Chlorophyll_Sign <chr>, Chlorophyll <dbl>, Secchi <dbl>,
#   Temperature <dbl>, Temperature_bottom <dbl>, Conductivity <dbl>, Conductivity_bottom <dbl>,
#   DissolvedOxygen <dbl>, DissolvedOxygen_bottom <dbl>, DissolvedOxygenPercent <dbl>,
#   DissolvedOxygenPercent_bottom <dbl>, pH <dbl>, pH_bottom <dbl>, TurbidityNTU <dbl>,
#   TurbidityNTU_bottom <dbl>, TurbidityFNU <dbl>, TurbidityFNU_bottom <dbl>, Pheophytin_Sign <chr>,
#   Pheophytin <dbl>, TotAlkalinity_Sign <chr>, TotAlkalinity <dbl>, TotAmmonia_Sign <chr>, TotAmmonia <dbl>, …
# i Use `print(n = ...)` to see more rows
```

deltamapr

- Provides spatial data for the SF Bay-Delta
- Data objects stored in ‘sf’ format as four data types:
  - Waterways (WW)
  - Regions (R)
  - Habitats (H)
  - Stations/Points (P)
- See GitHub repository for list of data available: https://github.com/InteragencyEcologicalProgram/deltamapr

IEP Integrated datasets - deltamapr

WW_Delta

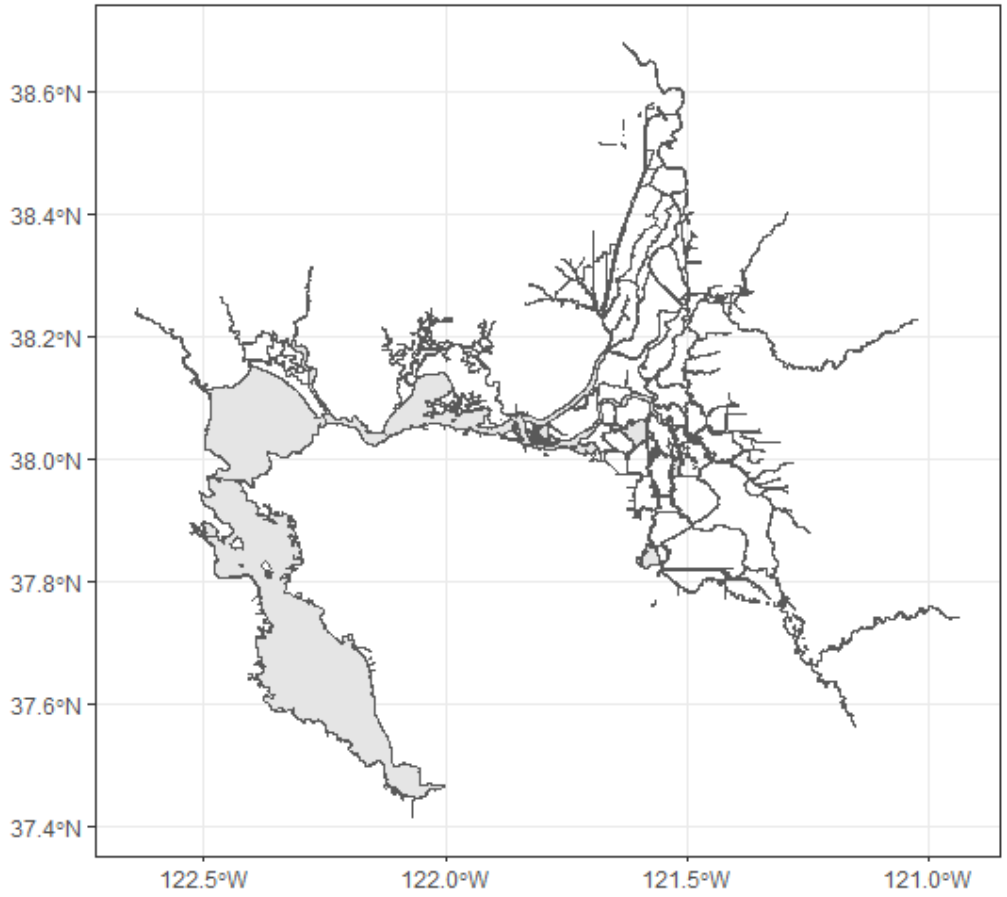
```r
# install.packages("devtools")
# devtools::install_github("InteragencyEcologicalProgram/deltamapr")
library(deltamapr)
library(sf)
WW_Delta
```

```
> WW_Delta
Simple feature collection with 282 features and 9 fields
Geometry type: POLYGON
Dimension:     XY
Bounding box:  xmin: -122.6408 ymin: 37.41522 xmax: -120.9357 ymax: 38.67781
Geodetic CRS:  NAD83
# A tibble: 282 × 10
      AREA PERIMETER HYDRO_POLY HYDRO_PO_1 HYDRO_24K_ TYPE  HNAME Shape_Leng Shape_Area                    geometry
     <dbl>     <dbl>      <int>      <int>      <int> <chr> <chr>      <dbl>      <dbl>                 <POLYGON [°]>
 1  7.35e7   1033340        791        797        798 MR    SACR…      2.45     3.48e-3 ((-121.5099 38.24711, -1…
 2  8.76e4      3319.      1965       1963       1964 S     W          0.0357   9.06e-6 ((-121.5673 38.57437, -1…
 3  7.92e6     87428.      1967       1965       1966 C     SACT…      0.829    8.17e-4 ((-121.5238 38.56153, -1…
 4  1.04e5      2719.      1970       1969       1970 L     GREE…      0.0264   1.07e-5 ((-121.6011 38.55476, -1…
 5  1.06e5      2798.      1977       1974       1975 L     LAKE…      0.0283   1.10e-5 ((-121.5456 38.55452, -1…
 6  1.59e5      3392.      1982       1978       1979 S     W          0.0314   1.65e-5 ((-121.6305 38.55163, -1…
 7  4.26e4      1003.      1992       1989       1990 S     W          0.00952  4.40e-6 ((-121.6298 38.52384, -1…
 8  5.65e3       498.      2001       2008       2009 MR    SOUT…      0.00548  5.84e-7 ((-121.6384 38.51655, -1…
 9  4.14e3       502.      2006       2012       2013 MR    SOUT…      0.00536  4.28e-7 ((-121.6315 38.51578, -1…
10  9.78e4      6630.      2008       2011       2012 MR    SOUT…      0.0746   1.01e-5 ((-121.6825 38.5156, -12…
# i 272 more rows
# i Use `print(n = ...)` to see more rows
```

IEP Integrated datasets - deltamapr

WW_Delta

```
ggplot(WW_Delta) + geom_sf() + theme_bw()
```



Other Resources

- rvest package – Web scraping
  https://rvest.tidyverse.org/index.html
- pdftools package – Extracting data from .pdf file
  https://docs.ropensci.org/pdftools/
- contentid package – work with external data through content identifiers
  https://cboettig.github.io/contentid/
- Code from this tutorial
  https://github.com/InteragencyEcologicalProgram/iep-data-workshop-2024/blob/main/coded_data_access/coded_data_access.Rmd