HBR\_watershed\_10

###Step 1: Load packages

library(metajam)   
library(udunits2)

## udunits system database read

# For wrangling the data  
library(readr)  
library(tidyr)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(purrr)  
library(stringr)  
library(metajam)  
library(tidyverse) # for convenience

## ── Attaching packages ──────────────────────────────────────────────── tidyverse 1.3.0 ──

## ✓ ggplot2 3.3.0 ✓ forcats 0.5.0  
## ✓ tibble 3.0.1

## ── Conflicts ─────────────────────────────────────────────────── tidyverse\_conflicts() ──  
## x dplyr::filter() masks stats::filter()  
## x dplyr::lag() masks stats::lag()

library(here) # for file path management

## here() starts at /home/kpeach/R/SI\_river\_data/lterwg-si-export

library(stringdist) # for first pass of naming-matching  
library(vctrs) # for joining tables  
library(readxl) # for reading the template excel sheet

####Step 2: Find the link to the dataset Go to the web address for the dataset and find the download button for the data.

In our case the link is: <https://portal.edirepository.org/nis/dataviewer?packageid=knb-lter-nwt.103.12&entityid=413c3eb5a1a52e3ce2f9e52b3565aadd>

#### Step 3: Choose where you want the files to be saved

In our case, we’ll just put it into the metajam\_example folder.

#eg desired\_path\_to\_data <- "~/Desktop"  
desired\_path\_to\_data <- "~/SI\_river\_data"  
  
# create the folder if it does not exist yet  
dir.create(desired\_path\_to\_data, showWarnings = FALSE)

#### Step 4: Download the data by pasting the link you just copied

my\_data\_url <- "https://portal.edirepository.org/nis/dataviewer?packageid=knb-lter-nwt.103.12&entityid=413c3eb5a1a52e3ce2f9e52b3565aadd"  
  
# this will download the data into a folder and save the path to that folder  
downloaded\_data <- download\_d1\_data(data\_url = my\_data\_url, path = desired\_path\_to\_data)

## Warning in check\_version(data\_url, formatType = "data"): Several identifiers are  
## associated with 413c3eb5a1a52e3ce2f9e52b3565aadd

##   
## Downloading metadata https://pasta.lternet.edu/package/metadata/eml/knb-lter-nwt/103/12 ...

## Download metadata complete

##   
## Downloading data https://pasta.lternet.edu/package/data/eml/knb-lter-nwt/103/12/413c3eb5a1a52e3ce2f9e52b3565aadd ...

## Warning in download\_d1\_data(data\_url = my\_data\_url, path =  
## desired\_path\_to\_data): This dataset has already been downloaded. Please delete  
## or move the folder to download the dataset again.

#### Step 5: Now read in the data (with all the metadata)

my\_data <- read\_d1\_files(downloaded\_data)

## Parsed with column specification:  
## cols(  
## attributeName = col\_character(),  
## attributeLabel = col\_character(),  
## attributeDefinition = col\_character(),  
## storageType = col\_character(),  
## definition = col\_character(),  
## measurementScale = col\_character(),  
## domain = col\_character(),  
## formatString = col\_character(),  
## minimum = col\_character(),  
## maximum = col\_double(),  
## missingValueCode = col\_character(),  
## missingValueCodeExplanation = col\_character(),  
## unit = col\_character(),  
## numberType = col\_character(),  
## metadata\_pid = col\_character()  
## )

## Parsed with column specification:  
## cols(  
## name = col\_character(),  
## value = col\_character()  
## )

## Parsed with column specification:  
## cols(  
## .default = col\_character(),  
## year = col\_double(),  
## date = col\_date(format = "")  
## )

## See spec(...) for full column specifications.

albisolu\_NWT\_data <- my\_data$data  
  
my\_data$attribute\_metadata

## # A tibble: 45 x 15  
## attributeName attributeLabel attributeDefini… storageType definition  
## <chr> <chr> <chr> <chr> <chr>   
## 1 LTER\_site LTER\_site Niwot Ridge LTE… string any text   
## 2 local\_site local\_site sample location string any text   
## 3 year year year (yyyy) dateTime <NA>   
## 4 date date date (yyyy-mm-d… dateTime <NA>   
## 5 time time time collected … dateTime <NA>   
## 6 pH pH pH; NP-not perf… float <NA>   
## 7 cond cond conductance float <NA>   
## 8 ANC ANC acid neutralizi… float <NA>   
## 9 acid acid acidity; NP-not… float <NA>   
## 10 alkal alkal alkalinity; NP-… float <NA>   
## # … with 35 more rows, and 10 more variables: measurementScale <chr>,  
## # domain <chr>, formatString <chr>, minimum <chr>, maximum <dbl>,  
## # missingValueCode <chr>, missingValueCodeExplanation <chr>, unit <chr>,  
## # numberType <chr>, metadata\_pid <chr>

###Step 6: data cleanup

my\_data$attribute\_metadata

## # A tibble: 45 x 15  
## attributeName attributeLabel attributeDefini… storageType definition  
## <chr> <chr> <chr> <chr> <chr>   
## 1 LTER\_site LTER\_site Niwot Ridge LTE… string any text   
## 2 local\_site local\_site sample location string any text   
## 3 year year year (yyyy) dateTime <NA>   
## 4 date date date (yyyy-mm-d… dateTime <NA>   
## 5 time time time collected … dateTime <NA>   
## 6 pH pH pH; NP-not perf… float <NA>   
## 7 cond cond conductance float <NA>   
## 8 ANC ANC acid neutralizi… float <NA>   
## 9 acid acid acidity; NP-not… float <NA>   
## 10 alkal alkal alkalinity; NP-… float <NA>   
## # … with 35 more rows, and 10 more variables: measurementScale <chr>,  
## # domain <chr>, formatString <chr>, minimum <chr>, maximum <dbl>,  
## # missingValueCode <chr>, missingValueCodeExplanation <chr>, unit <chr>,  
## # numberType <chr>, metadata\_pid <chr>

# I can see that the column Year\_Month includes the info from the column in the template 'Sampling Date' so I want to rename that column   
  
albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 rename('Sampling Date' = 'date')  
  
albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 rename('Site/Stream name' = 'local\_site')  
  
albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 rename('LTER' = 'LTER\_site')  
  
albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 rename('Time' = 'time')  
  
albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 rename('Conductivity' = 'cond')  
  
albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 rename('alkalinity' = 'alkal')  
  
  
  
#albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
# rename('NO3' = 'NO3-')  
  
#albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
# rename('PO4' = 'PO4---')  
  
#albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
# rename('NH4' = 'NH4+')  
  
#albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
 # rename('Na' = 'Na+')  
  
#albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
# rename('K' = 'K+')  
  
#albisolu\_NWT\_data <- albisolu\_NWT\_data %>%   
# rename('Mg' = 'Mg++')

Read in template

template <- read\_excel(here("metajam\_example", "Stream\_Data\_Template.xlsx"),   
 sheet = "Raw Data",  
 col\_types = "text"  
 ) %>%  
 mutate(`Sampling Date` = as.Date(`Sampling Date`))

Fuzzy match

# Start by matching by closest name as a first pass.   
 # Note that we match the lower case names  
 # Note that the weight i= 0.1 says that we will be more likely to match if watershed 1 = template + extra  
(fuzzy\_match <- tibble(template = names(template)) %>%  
 mutate(watershed10 = names(albisolu\_NWT\_data)[amatch(tolower(template), tolower(names(albisolu\_NWT\_data)), maxDist = 1, weight = c(d=1,i=0.1,s=1,t=1))])  
)

## # A tibble: 39 x 2  
## template watershed10   
## <chr> <chr>   
## 1 LTER LTER   
## 2 Site/Stream Name Site/Stream name  
## 3 Sampling Date Sampling Date   
## 4 Time Time   
## 5 Land Use <NA>   
## 6 Treatment <NA>   
## 7 Q (Discharge) <NA>   
## 8 alkalinity alkalinity   
## 9 ANC ANC   
## 10 pH pH   
## # … with 29 more rows

close\_matches\_albisolu\_NWT\_data <- my\_data$attribute\_metadata %>%   
 filter(attributeName %in% c('NO3-','PO4---','NH4+','NH4+','Na+','K+','Mg++'))  
  
print(close\_matches\_albisolu\_NWT\_data)

## # A tibble: 6 x 15  
## attributeName attributeLabel attributeDefini… storageType definition  
## <chr> <chr> <chr> <chr> <chr>   
## 1 NH4+ NH4+ NH4+; NP-not pe… float <NA>   
## 2 Mg++ Mg++ Mg++; NP-not pe… float <NA>   
## 3 Na+ Na+ Na+; NP-not per… float <NA>   
## 4 K+ K+ K+; NP-not perf… float <NA>   
## 5 NO3- NO3- NO3-; NP-not pe… float <NA>   
## 6 PO4--- PO4--- NP-not performed float <NA>   
## # … with 10 more variables: measurementScale <chr>, domain <chr>,  
## # formatString <chr>, minimum <chr>, maximum <dbl>, missingValueCode <chr>,  
## # missingValueCodeExplanation <chr>, unit <chr>, numberType <chr>,  
## # metadata\_pid <chr>

write.csv(close\_matches\_albisolu\_NWT\_data, file = "close\_matches\_albisolu\_NWT\_data.csv", row.names = FALSE)