|  |  |
| --- | --- |
| To: | Tina Laidlaw (USEPA, Region 8) |
| Cc: | DwaneYoung and Laura Shumway (both USEPA HQ) |
| From: | Jon Harcum and Erik W. Leppo (both Tetra Tech) |
| Date: | August 31, 2017 |
| Subject: | Contract EP-R8-12-04, Task Order 12: Proposed updates to USEPA Data Discovery Tool |

This technical memorandum fulfills the requirement in Contract EP-R8-12-04, Task Order 12, Subtask 4.1 to prepare a *Design Plan Memo* that identifies the proposed changes to USEPA’s Data Discovery Tool (DDT) that will be made by Tetra Tech. This memorandum was informed by a meeting with USEPA staff (Dwane Young and Laura Shumway) on July 7, 2017 and an effort to strike a balance between a tool previously developed by Tetra Tech for data extraction from the Water Quality Data Portal data and the target audience for the DDT.

The following sections of this memorandum presents steps and tasks that will were performed to make changes to the DDT. These modifications and updates are listed by the screens in the DDT. All work was done using the current version of the DDT (v1.1). The screen shots are included as examples of changes. The code for the modified DDT, as an attachment, will be returned to USEPA to integrate into the production version of the DDT.

# Additional Libraries

The code used for the DDT is written in R. As a programmy language R allows for additional capabilities through the use of add ins. These addins are called libraries or packages. The base version of the DDT ships with a number of packages. Two R packages were used to implement the modifications to the Data Discovery Tool.

* XLConnect
* DT (“editor” branch of DT package already in use in the DDT)

The package “XLConnect” (hosted on CRANN) was used to add the ability to read and save to Excel files for the QAQC Decisions file. Using Excel to store the data allows the user to save formatting and makes the files the most user friendly.

The “DT” package is already a dependency for the Data Discovery Tool. However, a specific version was used in the modifications to allow for on screen editing of tables. This version is available on GitHub at the following website; <https://github.com/rstudio/DT/tree/feature/editor>. The description of this version of DT is at: <https://github.com/rstudio/DT/issues/28>. In R, To install this branch version of DT use the following code, in R, without the starting and ending quotes ; “devtools::install\_github('rstudio/DT@feature/editor')”

The loading of the packages was added to the appropriate locations in the code but, as with the rest of the packages needed for the DDT, assumes the packages are installed on the user’s computer (or included in the R Portable distribution).

# Code Modifications

The base code of the Data Discovery Tool was left “as is” as much as possible. Rather than tweak an existing function, if feasible, a new routine or function was created to implement the additional functionality of the modifications.

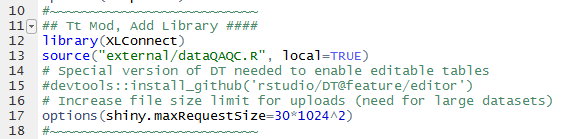
All new R code is marked with “# Tt Mod, TabName, Purpose ####” in the R files with sections set off with a string of tildes (Figure 1). The addition of this tag makes is easy to find the modification in the code and in the outline feature of RStudio (Figure 2).

Figure 1. Example of modified R code (server.R) with "Tt Mod" comment.

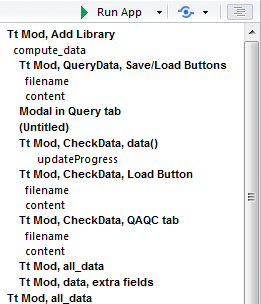
It was also necessary to increase the size of the maximum file size limit of Shiny for uploaded files (see line 17 in Figure 1). The file size limit was 5 MB. While the DDT is restricted to downloads of a maximum of 200,000 records some files that were tested were larger than 5 MB when saved and were not able to be reloaded into the tool. The maxium file size was increased to 30 MB to ensure all files could be uploaded. As a test all Region 08 data was downloaded. The number of records was 523,437 (well beyond the 200,000 limit) but the file size of the save data was only 11.7 MB (well within the new limit of 30 MB).

Figure 2. RStudio Outline feature showing links to modified code in the file server.R.

# New Files

Some new files were added to implement the modifications. All files are in the “external” folder.

* DDT\_QAQC\_BLANK.xlsx
* DDT\_QAQC\_Default.xlsx
* QAQC.R (still used?)
* dataQAQC.R

paragraph on each.

# Files Generated During Operation

The modifications to the DDT included the ability to save and load at different stages of the work flow in the Data Discovery Tool. These included Queries, Filters, and the Data. All files have a default format based on the data being saved along with the date and time. The user has the ability to change the location and name of the files but not the file type (RDS). Files are saved as RDS files to provide smaller files (as they are compressed) and avoid user manipulation. In R the benefit is that the information read from RDS files retains all characteristics of the original object in R when saved.

DDT\_Abc\_YYYYMMDD\_HHMMSS.rds

* Abc = short name of what type of information is being saved.
  + Data
  + Filters
  + Query
* YYYY = year (4 digit)
* MM = month (2 digit)
* DD = day (2 digit)

The user also has the option of saving the QAQC Decision Excel file. The default name for these files are “DDT\_QAQC\_Default.xlsx”. Again, the user has the ability to change the location and name of the file (but not the file type).

The only new “final” data output of the DDT (i.e., the all data table with QAQC decisions applied) is in the pre-existing TSV (tab separated) format.

# Bugs in Base Code

During testing of the modifications to the DDT there were three bugs that became evident in the DDT. These were detected during testing of the modifications but were found to also be present in the base code version of the DDT before any modifications were performed. These bugs were passed along to USEPA (Laura Shumway and Duane Young) but fixing them was beyond the scope of this project. These bugs are enumerated below.

## Bug 1. Revising Data Used in the DDT

If a user queries data the user cannot query new data to replace the existing data. That is, if a user has a set of data that they are using in the DDT they cannot work with a different dataset without first exiting the DDT. This bug is also manifested in the modification to load data from a saved file.

This can be observed by querying and retrieving any set of data. Going to the “View Data” tab to view the map. Then querying and retrieving a different set of data (easier to detect if in a different region/state). Then going back to the map on the “View Data” tab the original data will still be displayed.

## Bug 2. Filter for Date Range

The date range filter on the “View Data” tab appears to be using less than and greater than to subset the data. This is excluding the data that falls on the min and max dates.

This can be observed by retrieving data for the site “CROWCRKS-SITE 21” and parameters “temperature, water” and “pH”. On the “View Data” tab the table will include all 28 records that were enumerated during the data retrieval. After clicking “Submit!” for the default date range filter (2006-06-26 to 2007-10-30) the table results in only 24 records. Expanding the date range by one day for both min and max (2006-06-25 and 2007-10-31) and clicking “Submit!” returns the missing 4 records. If the user sorts the table by ActivityStartDate the first date is 2006-06-26 and the last date is 2007-10-30.

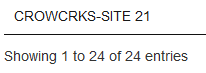
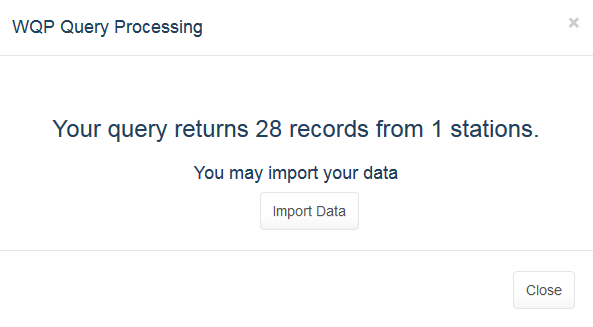
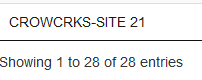


Figure 3



## 

## Bug 3. Filter for Sample Fraction Value of NA

The Filters do not properly handle the “NA” value for Sample Fraction. If remove “NA” from Sample Fraction it is not possible to get it back. It was not investigated if this bug (removing and reading NA values) applies to other data fields.

This bug can be observed using the same dataset as in Bug#2 (see previous section for dataset particulars). Removing the value “NA” from the filter and the clicking “Submit!”. Changes the record count from 28 to 0. Adding “NA” back to the filter and clicking “Submit!” will not return the 28 records. The sample fraction for this dataset is blank (NA) for all records. The sample fraction can be observed in the QAQC Advanced table on the modified “View Data” tab.

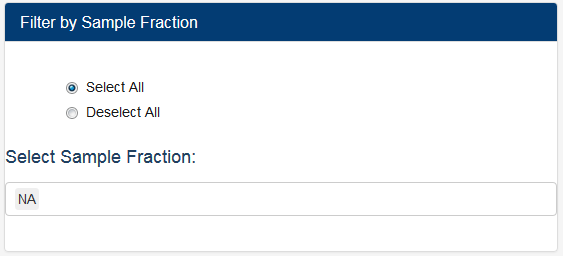


Figure 4. Sample Fraction filter on View Data tab showing only “NA”.

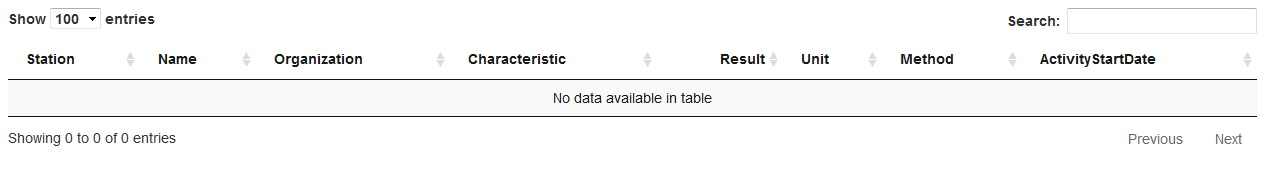


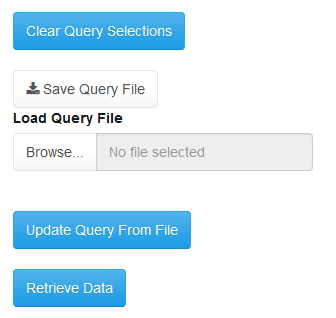
Figure 5. Table on View Data tab after Sample Fraction = NA filter turned off then on again; no records.

# Modifications by Section

## Query Data

Save and load query.

Clear query.



## Check Data

Save/Load data.

QAQC Decisions– 3 tabs

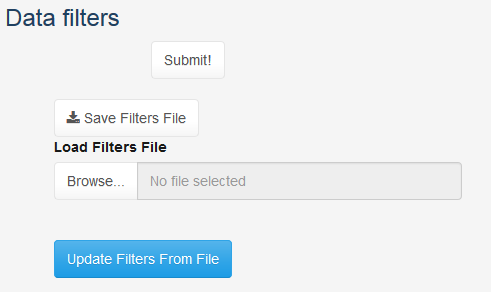


## View Data

Additional filters (3). Activity Type, Equipment, and StatusID (ActivityTypeCode, SampleCollectionEquipmentName, ResultStatusIdentifier).

Save/Load Filters

Data Summary – plots/tables and ability to save to file

JUST COPIED FROM PROPOSAL FROM HERE OUT

# Knowledge Base

The knowledge base is a series of reviewed QA/QC manipulations of parameter names, units, and sample fractions so that matching data can be combined by the users for subsequent analyses. Updates to the knowledge base will be performed using data extracted from the target audience for this project; Region 8 Tribes using water quality data (Table 1). This update will allow users to use the QA/QC “out of box” with minimal effort.

The Knowledge Base will be in Excel and can be edited offline by users.

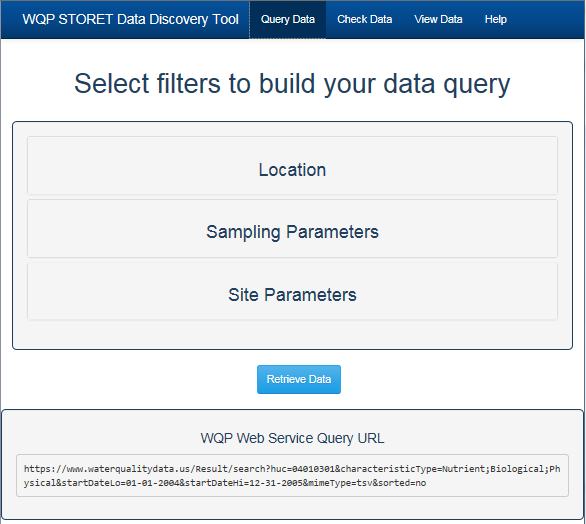
Table 1. Region 8 Tribal WQX users (n=19).

|  |  |
| --- | --- |
| Organization ID | Org Name |
| BLCKFEET | Blackfeet Nation (Montana) |
| CHEYRIVR | Cheyenne River Sioux Tribe (South Dakota) |
| CHIPCREE\_WQX | Chippewa Cree Tribe @ Rocy Boy Indian Reservation |
| CROWCRKS | Crow Creek Sioux Tribe (SD) |
| CSKTRIBE | Confederated Salish and Kootenai Tribes |
| FLASIOUX | Flandreau Santee Sioux Tribes (SD) |
| FORTPECK | Assiniboine & Sioux Tribes Fort Peck Indian Reservation (MT) |
| FTBLKNAP | Gros Ventre and Assiniboine Tribe (Fort Belknap Indian Res) |
| LWRBRULE | Lower Brule Sioux Tribe (South Dakota) |
| MHA\_NATN | Three Affiliated Tribes: Mandan, Hidatsa, and Arikara Nation |
| NRTHCHEY | Northern Cheyenne Tribe (Montana) |
| NTHRNUTE | Northern Ute Indian Tribe (UT) |
| OGLALAST | Oglala Sioux Tribe (South Dakota) |
| SOUTHUTE | Southern Ute Tribe |
| SPIRITLK | Spirit Lake Tribal EPA Programs |
| SRSTEPA | Standing Rock Sioux Tribe |
| SWO\_OEP | Sisseton-Wahpeton Sioux Tribe Lake Traverse Reservation (SD) |
| TURTLEMT | Turtle Mountain Environmental Office |
| UTEMTN | Ute Mountain Utes Tribe (Colorado) |

# Query DATA Screen

Buttons for saving and loading of query parameters will added to the Query Data screen (Figure 1) to allow users to easily recreate previous data searches. This will allow repeat data searches across sessions or sharing between data users.

Figure 6. Query Data screen proposed modifications.



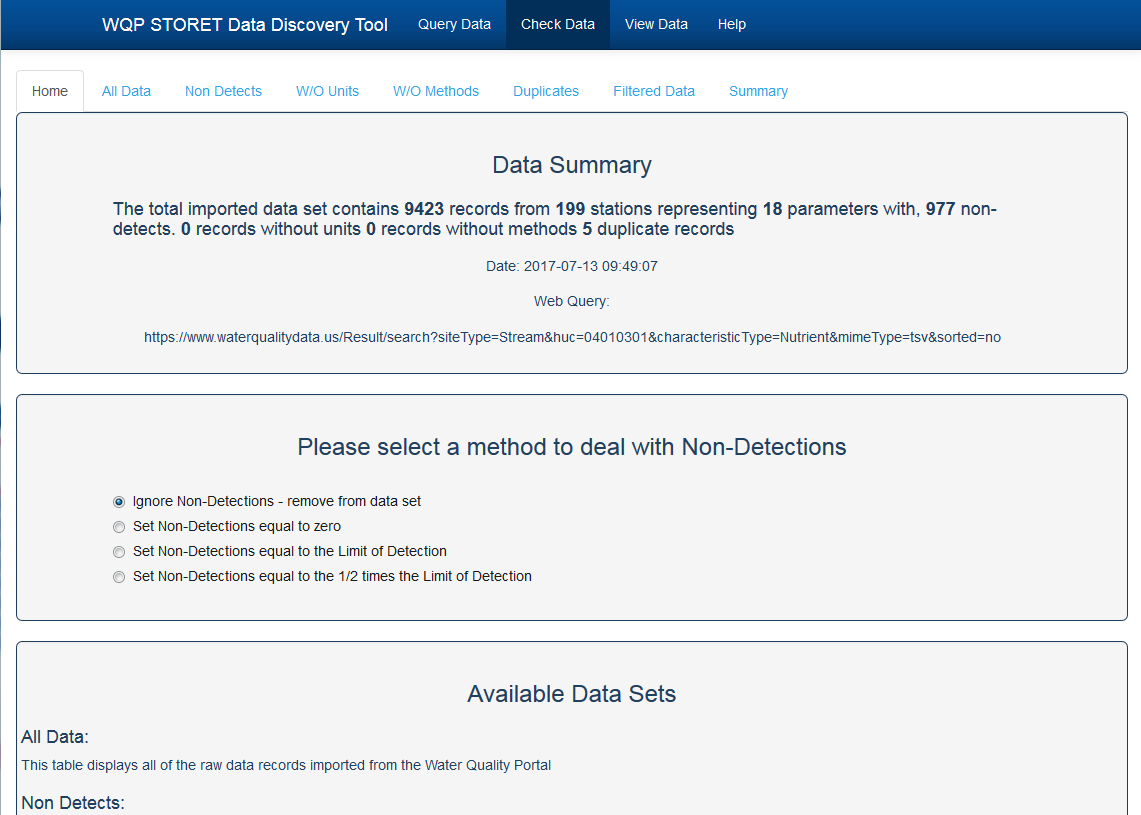
Load Query …

Save Query …

# Check Data Screen

The Check Data screen will be the location for accessing the Knowledge Base. This will be in the form of an extra tab (Figure 2). The user will be presented with the results of the matching from the Knowledge Base and will have the option of accepting or rejecting the matching with the use of TRUE / FALSE field in a table displayed in the DDT. The ‘ignore non-detects’ options will be removed. Figure 3 is an example of a similar table in Excel that will be used as the template (although analytical methods will be removed).

Figure 7. Check Data screen proposed modifications.



Knowledge Base

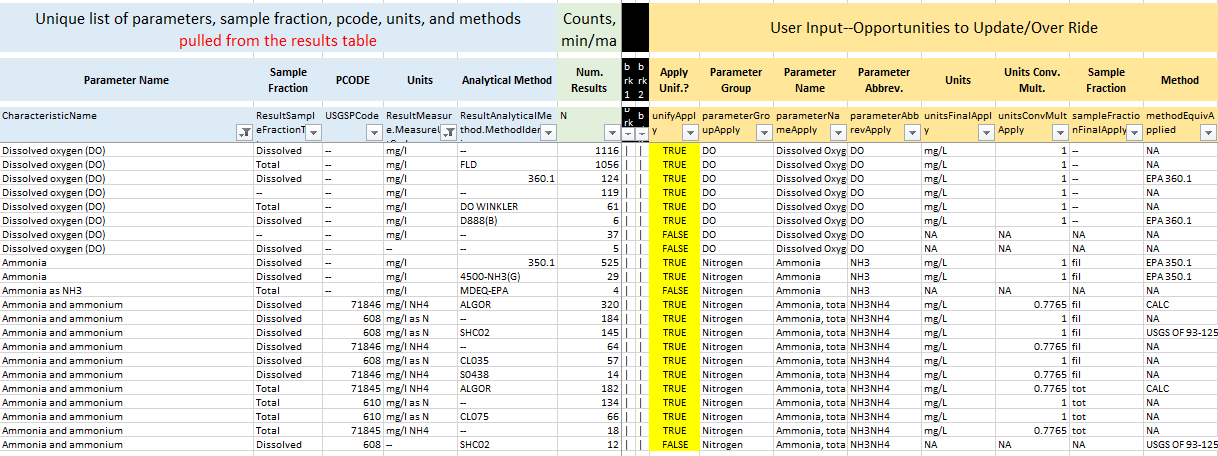


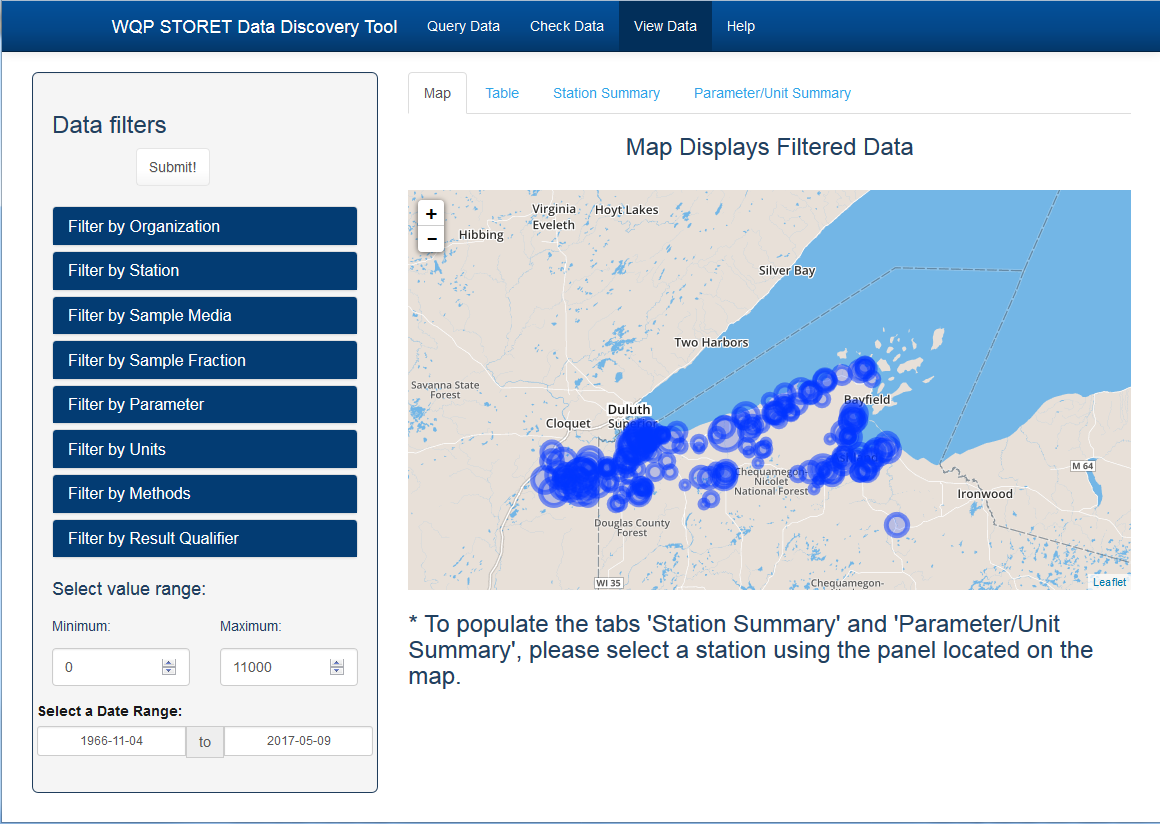
Figure 8. Example table of TRUE/FALSE matching.

# View Data Screen

The “View Data” screen will receive two sets up updates. The first update will be to add buttons to save and load filters. This will allow users to repeat previous filters of the data and/or share those filters with other users. Example buttons are shown in Figure 4.

The second update will be the addition of more filters. Filters for Activity Type, Equipment, and Status ID will be added in the format of the existing filters.

Figure 9. View Data screen proposed modifications.



Load Filter…

Save Filter…

Additional edits will added to the View Data screen (in a new tab) to include summaries on the filtered data. The current summaries are by station and parameter (Figure 4). A new tab will be added to generate summaries for the data set. These summaries will be both graphical (Figure 5) and in table format (Table 2 and Table 3) and can be created with a save button.

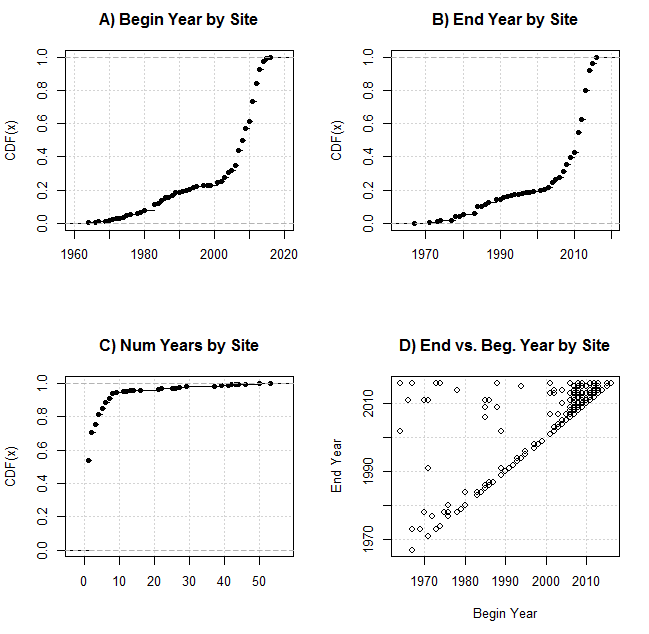


Figure 10. Comparison of beginning and ending year of record.

Table 2. Count of sites by county.

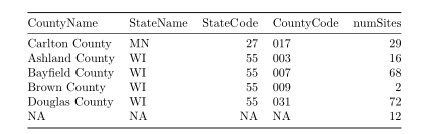
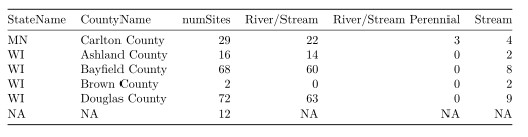


Table 3. Count of sites by county and station type.



# Saving and Loading Data

Both the “Check Data’ and “View Data” screens have the ability to save data (see example in Figure 5). Three modifications will be made to the functionality on these screens. The first is to allow the user the ability to “load” previously saved data. This will allow the user to modify existing data with more filters. This will make it so that the user can filter data across sessions and not have to complete all filtering in a single session using the DDT.

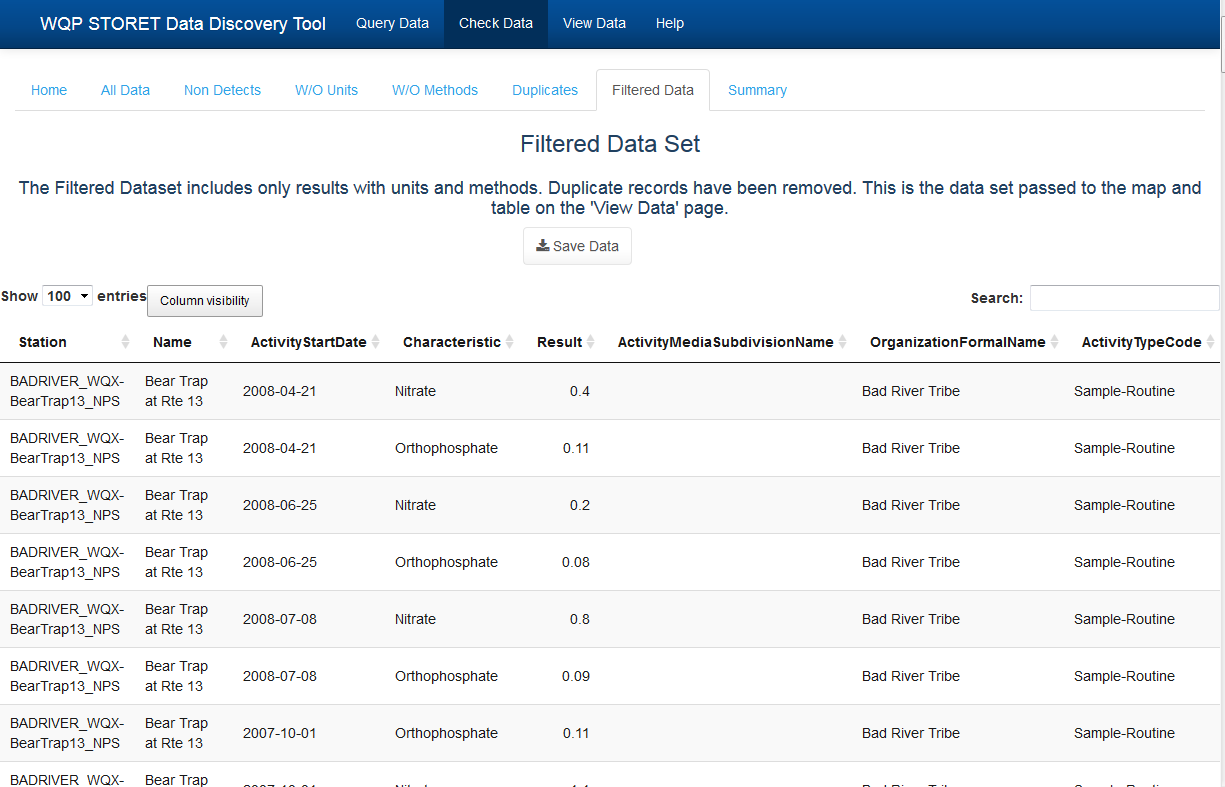
The second modification will be to change what data is saved. Currently filtered data is only what is shown to the user and all other data is removed (data without method, without units, and duplicates). It is proposed to save the data in two files; one to be called “keep” and the other “excluded”. The names may change but the idea is to allow the user to save both the filtered data and the data that was removed. This allows the user, if desired, to recreate the entire dataset by appending the two data files.

Figure 11. Example screen with a "save data" button.

The final modification will be to allow the user to save (and later load) their data as a native R file. That is, to use a data format (.Rdata) that will be smaller and load faster than the tsv format currently used in the tool.

# Additional Comments

Test data from Region 8 Tribes was used to ensure functionality of the tool and to address any potential issues with the modifications. All issues were addressed. As stated above there were three bugs in the base version of the DDT that were beyond the scope of this project and were not addressed.

The modifications to the DDT were made to a version outside of the portable R framework of the DDT. That is, all necessary packages were installed and used in the latest version of R and RStudio. Code versioning and issue tracking was performed on a private Tetra Tech GitHub site. The attachment is a zip file of that code and represents the deliverable of the modified DDT code.

In most cases, the modified code includes comments as to the nature of each code piece. Experimental code was at times used in the code and then commented out. These comments were removed but some may have been overlooked and remain in the code. But as they are comments they do not affect the proper functioning of the DDT.