|  |  |
| --- | --- |
| To: | Tina Laidlaw (USEPA, Region 8) |
| Cc: | DwaneYoung and Laura Shumway (both USEPA HQ) |
| From: | Erik W. Leppo and Jon Harcum (both Tetra Tech) |
| Date: | September 27, 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 modify and deliver code for an updated Data Discovery Tool (DDT).

The following sections of this memorandum presents steps and tasks that 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. The final version of the DDT will be called v1.1.3.

As a separate attachment will be a Powerpoint (Data Discovery Tool QAQC Mods.pptx) describing changes from v1.1 of the DDT. Two html files with directions for setting up (DDT\_Packages.nb.html) and starting (DDT\_StartUp.nb.html) the modified DDT are included. The original DDT quick start guide will be included as well.

# Additional Libraries

The code used for the DDT is written in R. As a programming 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 (main branch version of DT package already in use in the DDT but this is the “editor” branch)

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). Zip files for each of these packages is included. The file “DDT\_Packages.nb.html” includes R code for installing both of these packages.

# Code Modifications

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

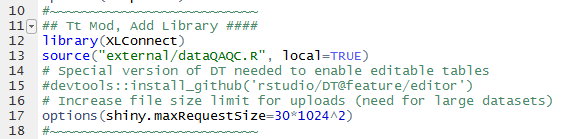
All new R code is marked with “# Tt Mod, TabName, Purpose ####” in the R files with sections set off with a string of tildes (Figure 1). The addition of this tag makes is easy to find the modification in the code and in the outline feature of RStudio (Figure 2).

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

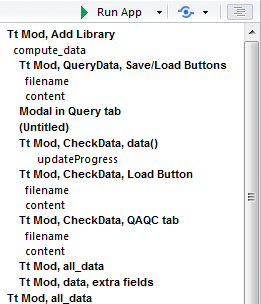
It was also necessary to increase the size of the maximum file size limit of Shiny for uploaded files (see line 17 in Figure 1). The file size limit was 5 MB (the Shiny default). While the DDT is restricted to downloads of a maximum of 200,000 records some files that were tested were larger than 5 MB 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 Tribal 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

Five new files (listed below) were added to implement the modifications. All files (except NEWS.rmd) are in the “external” folder in the file structure of the DDT. All other modifications were performed on existing files.

* NEWS.rmd
* DDT\_QAQC\_BLANK.xlsx
* DDT\_QAQC\_Default.xlsx
* dataQAQC.R
* Report\_QAQC\_Summary.rmd

In the root folder of the DDT the file “NEWS.rmd” was created to track the changes being made to the DDT and version numbers maintained on the GitHub repository.

The two Excel files contain the QAQC Decisions. One is the default file while the other is a blank version used when creating new versions.

dataQAQC.R contains some new code related to the importing the QAQC Decisions file and applying the decisions file to the data.

Report\_QAQC\_Summary.rmd is an R markdown file that is the template for the new data summary report with the plots and tables.

# 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 name based on the data being saved along with the date and time (see example below). This is done to keep the files organized on the user’s computer. With the date and time in the file name it is less likely that the user will overwrite and existing file. The user has the ability to change the location and name of the files but not the file type (RDS for queries and filters but RDA for data and url). The default folder is that of the web browser being used. Files are saved as RDS and RDA files to provide smaller files (as they are compressed) and avoid user manipulation. In R the benefit of these files types is that they are native and thus read much faster than other file types. These files also retain the column characteristics (e.g., data, text, or numeric) and other file structures. Files are saved according to the naming convention below.

DDT\_Abc\_YYYYMMDD\_HHMMSS.rds (rda or xlsx)

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

The only other new data output of the DDT is the all data table with QAQC decisions applied. This file is exported 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 staff (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 downloaded 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.

However, with a saved and reloaded dataset if you click “Submit!” on the View Data tab the map will change to the new dataset.

## 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 (Figure 3). After clicking “Submit!” for the default date range filter (2006-06-26 to 2007-10-30) the table results in only 24 records (Figure 4). 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).

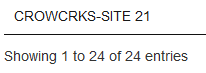


Figure 3. Example site screenshots of number of records and date range.

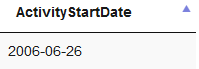
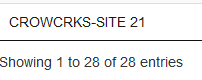
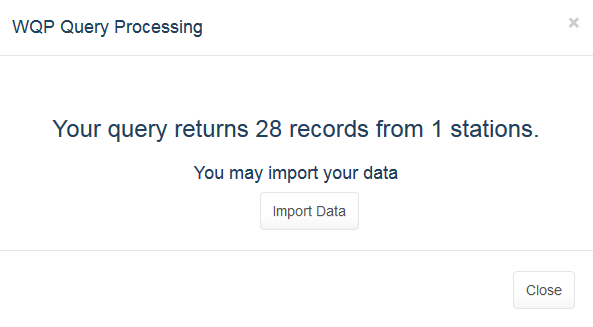


Figure 4. Example site number of records after applying default filter.

## 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 (Figure 5) and the clicking “Submit!”. Changes the record count from 28 to 0 (Figure 6). 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.

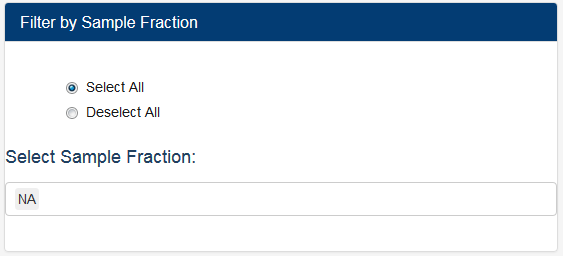


Figure 5. Sample Fraction filter on View Data tab showing only “NA”.

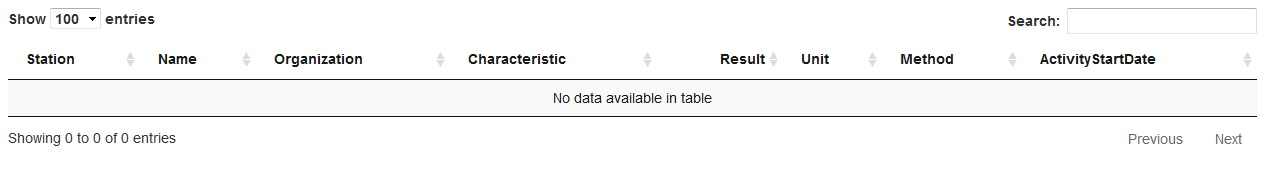


Figure 6. Table on View Data tab after Sample Fraction = NA filter turned off then on again; no records.

# Modifications by Section

Each major tab of the DDT is listed in the following sections with descriptions of the modifications.

## Query Data

Buttons for saving and loading of query parameters were added to the Query Data screen (Figure 7) to allow users to easily recreate previous data searches. This will allow repeat data searches across sessions or sharing between data users.

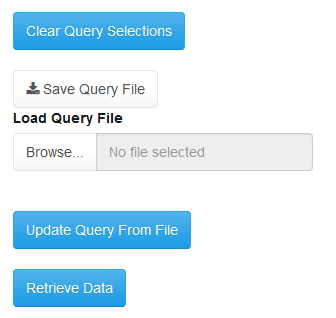


Figure 7. Query Data tab buttons.

The “Clear Query Selections” button allows the user the delete entries from all boxes on this screen. This is useful when a very detailed query has been constructed. Rather than have to remove entries from each box a single click removes all entries.

The “Save Query File” button allows the user to save the query for use later.

The “Browse” button allows the user to select a previously save query file. The “Update Query From File” button must be clicked after loading a query to update the boxes on this tab.

The “Retrieve Data” button fetches the data and was a pre-existing buttong that was not modified.

## Check Data

The Check Data screen is the location for accessing the QAQC Decisions, formerly named the Knowledge Base. This is also the location for dealing non-detectes, saving and loading of data, and a summary report.

Figure 8. Check Data screen proposed modifications.

### Non Detects

The Non-Detect option “ignore non-detects” was removed and the option “set to ½ times the limit of detection” was set as the default (Figure 9). This preserves all of the data records for the data analyst to determine the best course of action for dealing with non-detects that will be appropriate for the intended data analyses. Some times non-detects will be dealt with differently depending on the analysis to be performed.

### Saving and Loading Data

Figure 9. Non-Detect selection.

It is possible to save and load data out of and into the the DDT. This makes it easier as all work in the DDT does not need to be completed in a single session. This is especially important for large data downloads that take a long time to complete.

On the 2nd Subtab of the Check Data taba is the “Save/Load App Data” subtab (Figure 10). This is where the user is able to save and load data via two buttons. Each button prompts the user for file name and location.

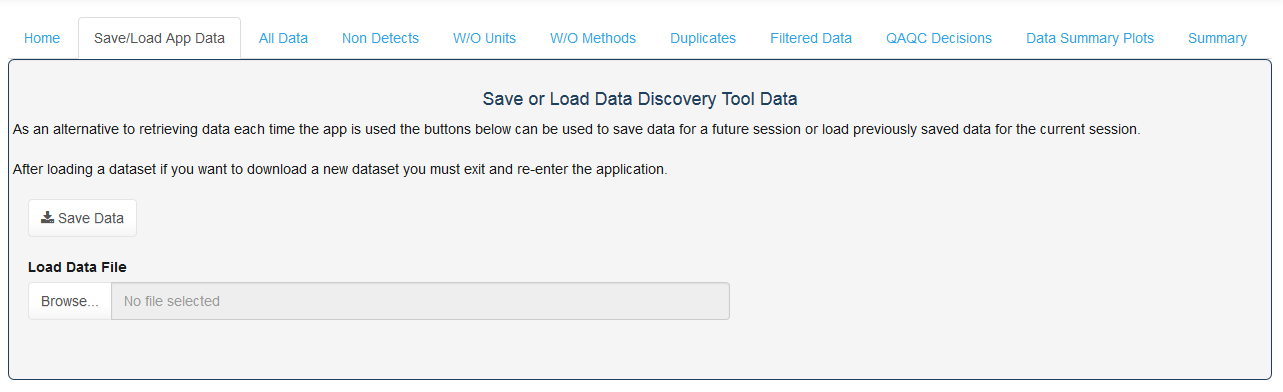


Figure 10. Save and Load App Data buttons.

### QAQC Decisions

The QAQC decisions file is a knowledge base with a series of reviewed QAQC manipulations of media, parameter names, units, and sample fractions so that matching data can be combined by the users for subsequent analyses. Updates to the QAQC decisions were 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 QAQC Decisions “out of box” with minimal effort.

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) |

The Check Data screen is the location for accessing the QAQC Decisions. This is in the form of an extra tab (Figure 11). The user will be presented with the results of the matching from the QAQC Decisions and will have the option of accepting or rejecting the matching with the use of TRUE / FALSE in the “Apply QAQC” field in a table displayed in the DDT.

Figure 11. QAQC Decisions subtab.

#### Add New QAQC Decisions

The first option the user is presented with is to add the new combinations of media, characteristic, sample fraction, and unit to the QAQC Decisions File (Figure 12). These are combinations present in the current data set but not in the QAQC Decisions file. If the user wishes to add these new combination then it happens when the “Add New QAQC Decision Combinations” button is clicked.

#### QAQC Decisions File

Figure 12. Button to add new QAQC Decisions present in the data to the QAQC decisions table.

The QAQC Decisions file is in Excel and can be edited offline by users. It can be downloaded from the QAQC Decisions tab on the Check Data tab using the “Save QAQC Decisions File” button (Figure 13). Or later uploaded viasn the “Browse” button.

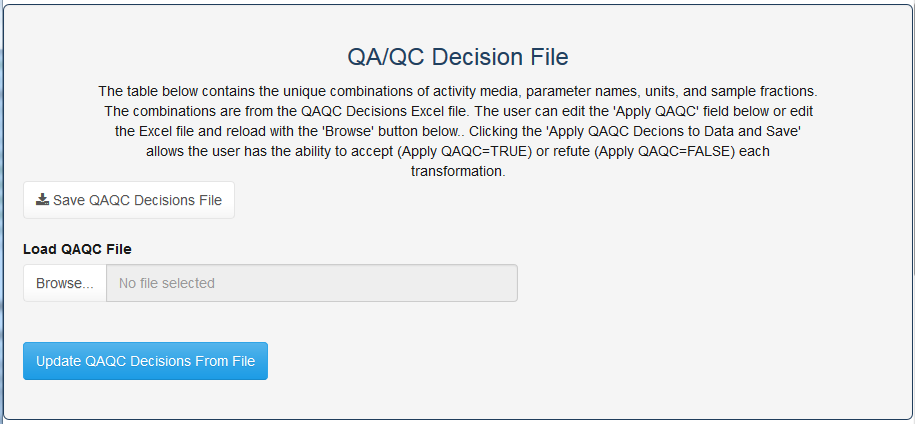


Figure 13. Buttons to save and load the QAQC Decisions Excel file.

#### Editing the QAQC Decisions Table

At the bottom on the page is the QAQC Decisions table (Figure 15). The user is allowed to edit the column “Apply QAQC” by double-clicking on the cell to edit. Only values of TRUE or FALSE are valid. After editing the user needs to save the file or the edits will be lost when the DDT is closed.

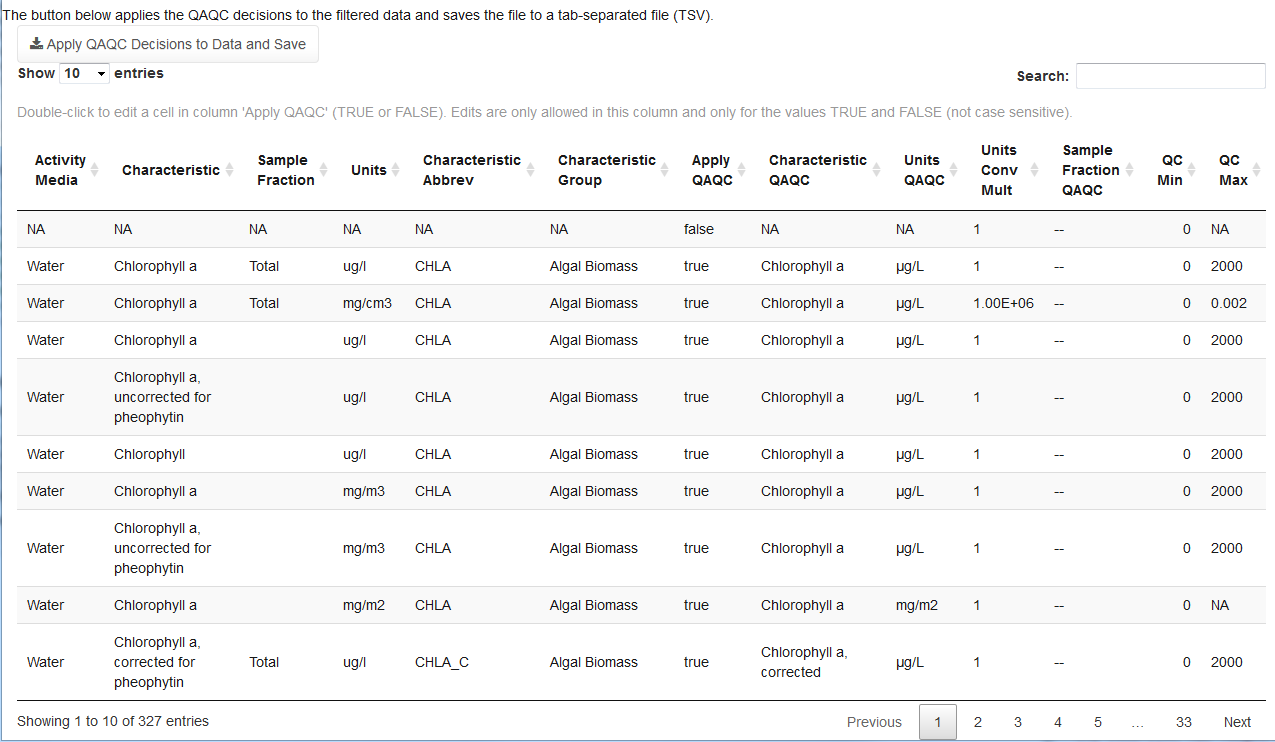


Figure 14. QAQC Decisions table.

#### Applying the QAQC Decisions

The final operation available to the user is to apply the QAQC Decisions to the data. Clicking the “Apply QAQC Decisions to Data and Save” button (Figure 15) will make modifications to the users data and prompt for a file name and location. Any combinations of media, characteristic, sample fraction, and unit that have a value of TRUE in teh column “Apply QAQC” will have modified data in the final output. The output will be a TSV file similar to other TSV outputs in the DDT.

### Data Summary Plots and Tables Report

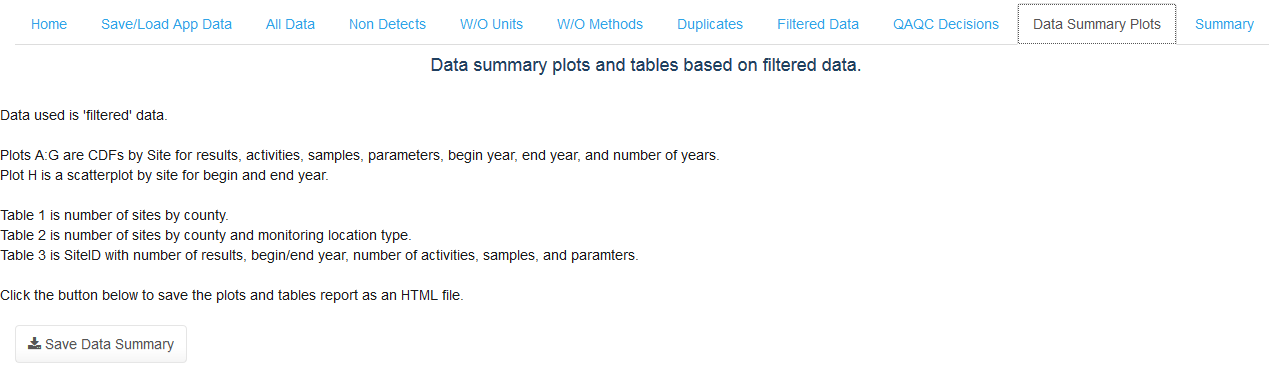
The ability to save a report with summary plots and tables for the “filtered” data was added to the “Data Summary Plots” subtabs on the Check Data tab (Figure 16).

Figure 15. Save Data Summary button.

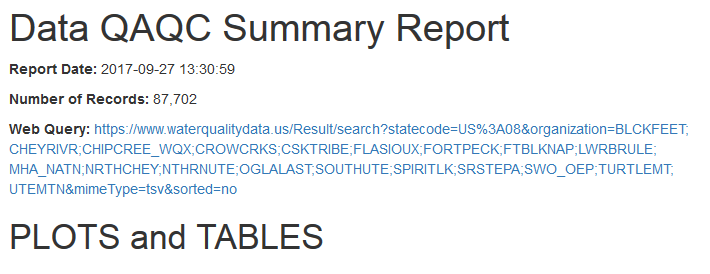
The summary report saves as html a file with a header (Figure 17) with the date of the report, number of records, web query, several CDF plots (Figure 18 and Figure 19) and tables (Figure 16).

Figure 16. Summary report header.

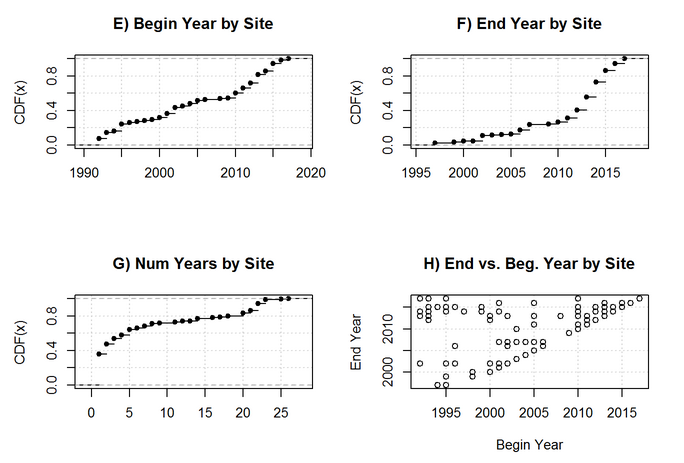
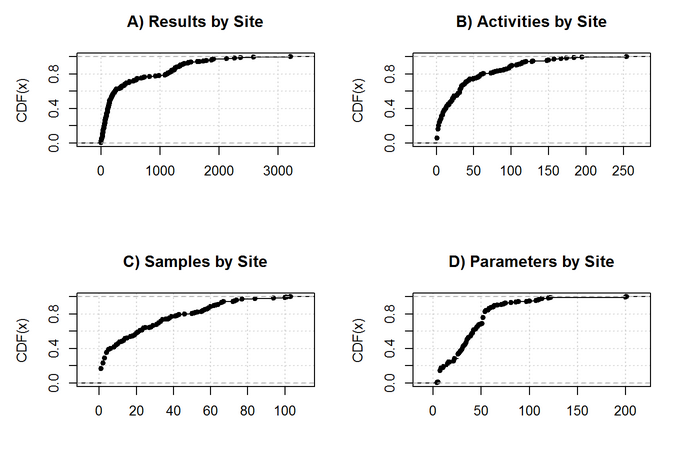


Figure 17. Summary report CDF plots by site and results, activities, samples, and paramters.

Figure 18. Summary report charts comparing beginning and ending years by site.

## View Data

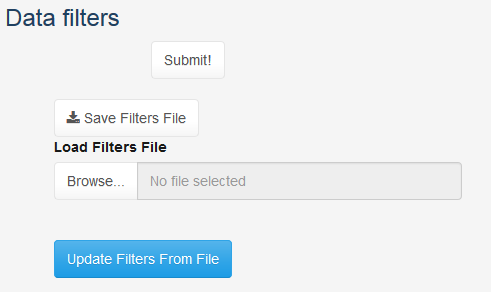
The “View Data” screen received two minor updates. The first update was to add buttons to save and load filters (Figure 19). This will allow users to repeat previous filters of the data and/or share those filters with other users. The “Submit!” button is the same and was not modified. After changing the filters or loading a filter the “Submit!” button must be clicked to apply the filters to the data.

Figure 19. Additional Data Filter buttons.

The second update was the addition of more filters. Filters for Activity Type, Equipment, and Status ID were be added in the format of the existing filters (Figure 20).

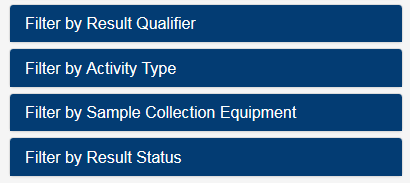


Figure 20. Addtional data filters.

# DDT Created FIles

During the testing of the modifications of the DDT queries and data files were created for each of the states in Region 08 for the 19 Tribes that were the focus of this work. These saved query and data files are included with the deliverable so they can be uploaded into the DDT. The QAQC Decisions file that was updated for Region 08 Tribal data is the default QAQC Decisions file in the DDT. It is also included as a separate file with the deliverable. Finally, the data summary plots and table reports for each state are included with the deliverable.

# 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 and appeared to work properly in the final version of the modified DDT. 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. Code that has been commented out can be ignored as they are comments they do not affect the proper functioning of the DDT.