UCD CSN Technical Information #801D

CSN Data Delivery

Chemical Speciation Network
Air Quality Research Center
University of California, Davis

July 31, 2019
Version 1.2

Prepared By: _____ Date: July 30, 2019

Reviewed By: _____ Date: July 26, 2019

Approved By: Nicole Hyslop Date: 7/31/19

UCDAVIS
AIR QUALITY RESEARCH CENTER

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications
11/30/18	NJS	1,2,3,7,8,9,10,11	Rewording for clarity and updating name changes.
7/31/19	DEY, KAG	8	Updated processes and added detail/clarity.

TABLE OF CONTENTS

1. PURPOSE AND APPLICABILITY 4
2. SUMMARY OF THE METHOD 4
3. DEFINITIONS 4
4. HEALTH AND SAFETY WARNINGS 4
5. CAUTIONS 4
6. INTERFERENCES 4
7. PERSONNEL QUALIFICATIONS, DUTIES, AND TRAINING 5
8. PROCEDURAL STEPS 5
8.1 Format data for DART Submission 5
8.2 Receive and Review Returned Data from DART 6
8.3 Prepare and Submit to AQS 9
9. DATA DELIVERY FLOW 12
10. EQUIPMENT AND SUPPLIES 13
11. QUALITY ASSURANCE AND QUALITY CONTROL 13
12. REFERENCES 13

LIST OF FIGURES

Figure 1. Returned DART datasets can be imported into the CSN database with this custom upload tool. 7
Figure 2. Login screen for the EPA's Exchange Network Services website. 10
Figure 3. Home screen of the Exchange Network Services website. 11
Figure 4. Enter "AQS" into the search bar. 11
Figure 5. Search results from "AQS". The analyst should choose the Service named "AQS Submit". 11
Figure 6. AQS data submission form. 12
Figure 7. Flow diagram of the data delivery process. 13

1. PURPOSE AND APPLICABILITY

The subject of this technical information document (TI) is delivery of the resultant data from the Chemical Speciation Network (CSN). This document describes the procedure for preparing and delivering data to the Data Analysis and Reporting Tool (DART) and the Air Quality System (AQS) database.

2. SUMMARY OF THE METHOD

The University of California, Davis (UCD) analyst prepares delivery files of the validated CSN data sets using custom tools in the *datvalCSN* R package. Data are formatted for delivery to DART, where they are reviewed by state, local, and tribal (SLT) agencies. After the 30 day review period, revised DART results are returned to UCD. The analyst ingests this file into the UCD CSN database, addresses any comments and/or requests from the DART reviewers, and reformats the dataset into AQS format. The final data file are checked for correctness and then submitted to the Environmental Protection Agency's (EPA) AQS database.

3. DEFINITIONS

- **AQS:** EPA's Air Quality System database.
- **Chemical Speciation Network (CSN):** EPA's $PM_{2.5}$ sampling network, with sites located principally in urban areas.
- **crocker:** A custom software package in the R language that contains the data processing code used to produce, check, and post the final results.
- **CSN database:** A SQL Server database that is the central warehouse of CSN preliminary and final data at UCD.
- **CSV:** a comma-separated value file that is the common format for delivery files.
- **datvalCSN:** A custom software package in the R language that contains the data validation code used to collect, compare, and flag the final results.
- **Data Analysis and Reporting Tool (DART):** A web application for environmental data visualization and validation procedures.
- **STI:** Sonoma Tech, Inc. Contractor developing and operating the DART interface.

4. HEALTH AND SAFETY WARNINGS

Not applicable.

5. CAUTIONS

Not applicable.

6. INTERFERENCES

Not applicable.

7. PERSONNEL QUALIFICATIONS, DUTIES, AND TRAINING

The UCD Air Quality Research Center (AQRC) Data & Reporting Group staff assigned to tasks described in this document have advanced training in data analysis and handling.

8. PROCEDURAL STEPS

Data delivery is performed using the *datvalCSN* R package, which is developed and maintained by UCD specifically for data processing, monitoring, and validation of the CSN data. Data delivery is performed by the UCD Data & Reporting Group on monthly batches of data (a calendar month of sample start dates). Delivery occurs in three steps:

1. Format data and submit to DART (Section (8.1).
2. Receive and reformat returned data from DART (Section 8.2).
3. Prepare and submit to AQS (Section 8.3).

8.1 Format data for DART Submission

In the previous step, results for elements, ions, and carbon are validated using the tools of the *datvalCSN* package. This includes assigning qualifier and null flags to specific records. Once the analyst has completed the validation, the analyst will run the *csn_calculate_all* function from the *crocker* package to update data such as concentrations and MDLs where sample volumes or validity of field blanks changed during validation of the original dataset.

[month_data] <- crocker::csn_calculate_all([YYYY], [MM], 'production')

The analyst will post the data to the database using the following command, including the Results Set ID for the previous valid data set and the Analysis Quality Control (QC) Code that should be assigned to the previous Results Set.

[post] <- crocker::csn_post_results([month_data], 'production', AnalysisQcCode = 1, comment = ['Initial Posting'], replacingId = [previousResultsSetId], replacingQcCode = [PreviousResultSetNewQCcode])

The analyst will then run the *csn_validate* function with the *write.flags* option in the following command set to *TRUE*[1],

[monthData] <- datvalCSN::csn_validate(Month = ['MM'], Year = ['YYYY'], write.flags = [FALSE],)

The command form shown here is appropriate for this stage in the process; more options can be included by changing defaults. The default recipient is "DART", thus does not need specified in the command. To find out more information on this command, run *?csn_validate* in the R environment. Flags are assigned when this function is run (see

[1] Text in [brackets] indicates values that can be changed by the user. Other values should be typed as written.

UCD CSN SOP #801: Processing & Validating Raw Data). The analyst will then populate the *delivery.DartExport* table in the CSN database using,

[dart.post] <- post_output([allData]@Output.DF, label = ['YYYY Month'], comment = ['YYYY Month DART export'], recipient = 'dart').

This function can be used for posting data with different recipients (e.g. DART and AQS) to the UCD CSN database, as will be described later in the TI, although the default recipient is 'DART'. Other options can be included in this command such as whether data are to be overwritten in the database. To find out more information on this command, run *?post_output* in the R environment.

Next, the *generate_dart_file* function collects the necessary data from the CSN database and formats it for DART. To select specific data, the analyst can enter either one or more unique DartSet IDs or a date range to prepare the file. Additionally, the analyst can generate a local copy of the output file in their R session for final review before generating the delivery file by specifying the *make.file* parameter as *FALSE*. Using DartSet IDs and setting *make.file* to *TRUE* generate the final delivery file for DART in the directory specified for the *out.dir* parameter:

[dart.data] <- generate_dart_file([dartSets], make.file = TRUE, out.dir = [output directory]).

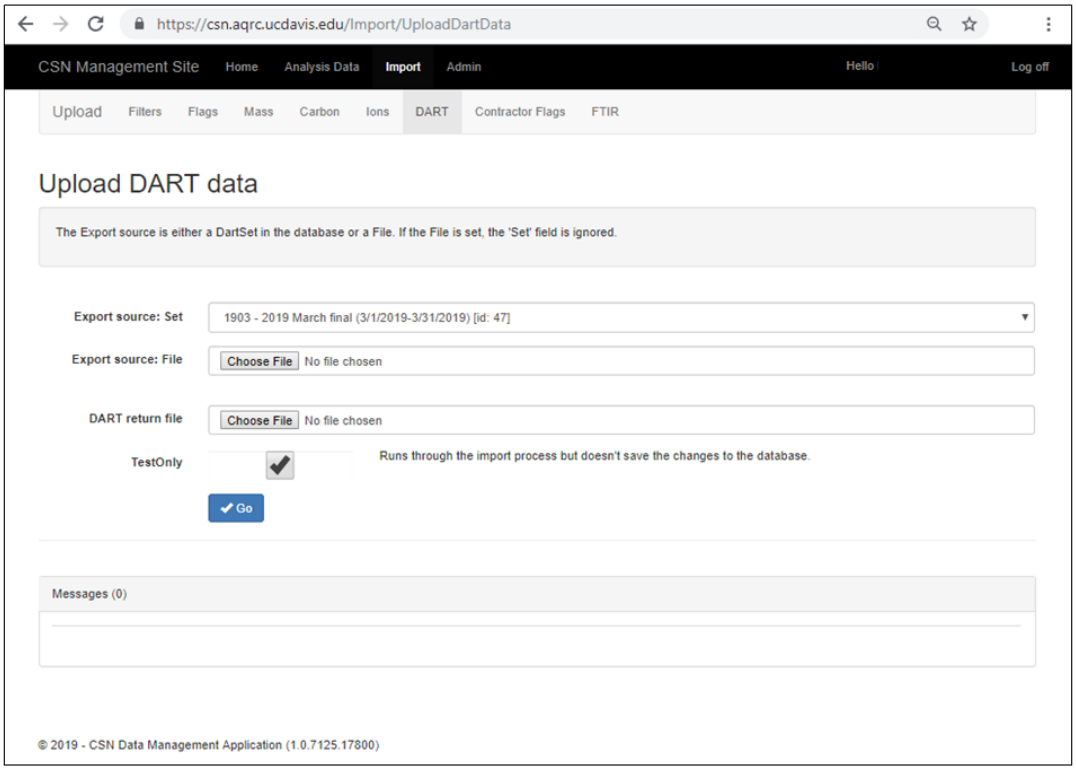
Files are delivered to STI by UCD via a File Transfer Protocol (FTP) service.

8.2 Receive and Review Returned Data from DART

The state, local, and tribal (SLT) agencies have 30 days to review their associated data and perform validation. Upon completion, the entire dataset is returned to UCD with a change log. The data are ingested using the DART import tool in the UCD CSN Management Site web application (csn.aqrc.ucdavis.edu). The analyst will navigate to the "Import" tools (top menu bar), then to the "DART" upload tool (second from right option near top of screen) (see Figure 1).

The DART return file is compared to the data that was delivered from UCD. The analyst can either specify the data set in the database using the top drop down menu labeled "Export source: Set" or by providing the path to the delivered file using the "Choose File" button next to the label "Export source: File". The "DART return file" is the comma-separated value (CSV) file containing the dataset that was validated in DART.

Figure 1. Returned DART datasets can be imported into the CSN database with this custom upload tool.



As with the other data upload processes, use the "Go" button to begin ingestion. It is advisable to first run the import with the "TestOnly" box checked and review any error or warning messages. The ingested data appear in the *analysis.DartReturn* table.

The analyst will review the change log from STI, including comments, and verify that all changes are consistent with the DART users' comments. To collect and review the changes made during DART validation, the analyst can run

[dartChanges] <- datvalCSN::get_changes(dartSetId = [SetId])

where the *dartSetId* is a numeric value for the DART data set of interest. The command form shown here is typical for obtaining a dataframe containing change details made between the DART export and DART return datasets; more options can be included by changing defaults. The analyst can also specify a date range instead of the *dartSetId* to obtain the change log. To find out more information on this command, run *?get_changes* in the R environment.

A comments log file, in CSV format, is provided from STI along with the data file. From the comments log file, the analyst can obtain information regarding changes made in DART as well as comments left by the SLT validators (all actions in DART require a comment to be added). The comments log can be filtered to review changes made at a given site and date as well as to view contact details for the SLT validator.

The analyst will then convert the data set from DART to AQS format using the function,

[aqs] <- datvalCSN::dart2aqs([dartSetId])

with the appropriate DART data set ID. The analyst can instead provide start and end dates as in the *generate_dart_file* function. The *dart2aqs* function creates a local copy of the data that will be used to populate the *delivery.AqsStaging* and *delivery.AqsStagingQualifierCodes* tables in the UCD CSN database. The following command can be used to post the data created in the previous command to the relevant tables in the UCD CDN database:

[aqs.staging.post] <- datvalCSN::post_output([aqs], recipient = 'aqs.staging')

Note that this same function was used for posting the DART export data to the database; the recipient has been changed accordingly for this stage in the data pathway. A label and comment are not applicable when posting data to the *delivery.AqsStaging* and *delivery.AqsStagingQualifierCodes* tables as there are no sets for AQS staging data. The analyst will need to review the DART validator changes and comments and make any requested updates in the aforementioned tables. The *datvalCSN* package includes a number of functions to aid in this process, including:

- *update_conc:* corrects concentration values after a sample volume value was changed
- *update_values:* applies a scalar multiplication to analytical results for a specified filter ID
- *update_single_value:* replaces a result value for a specified parameter and filter ID
- *resurrect_flags:* reapplies qualifier flags for records that are no longer invalid
- *invalidate_aqs_filter:* applies a null code to a filter (may be parameter specific)
- *swap_aqs_records:* depending on the nature of the swap, this function swaps two sets of filter records by swapping dates between two samples, or by swapping the dates for the analytical species parameters and updating the concentrations as necessary
- *re_validate_filter:* remove null code and reapply qualifier flags for a specific filter (may be parameter specific)
- *remove_specific_quals*: removes a specific qualifier flag for a specific filter, parameter for a given filter, or group of parameters for a given filter
- *add_aqs_qualifiers*: adds a specific qualifier flag for a specific filter, parameter for a given filter, or group of parameters for a given filter
- *change_start_dates*: updates the sample start date for a given filter or filter parameters
- *update_fromDF*: updates all records for a filter based on the data frame provided when running the function in a command.

8.3 Prepare and Submit to AQS

Once all changes have been made in the *delivery.AqsStaging* and *delivery.AqsStagingQualifierCodes* tables, the analyst will create the final dataset in the *delivery.AqsExport* table using the *aqsStaging2export* function,

[aqs.export] <- aqsStaging2export([dartSets])

And post using another variation of the *post_output* function:

[aqs.export.post] <- datvalCSN::post_output([aqs.export], recipient = 'aqs.export')

The analyst will then prepare the final delivery files using the *generate_aqs_file* function, once for sample data and once for field blank data, as the formats of the two files are different. The analyst can review the final data before making the file by setting *make.file* to FALSE in the following commands:

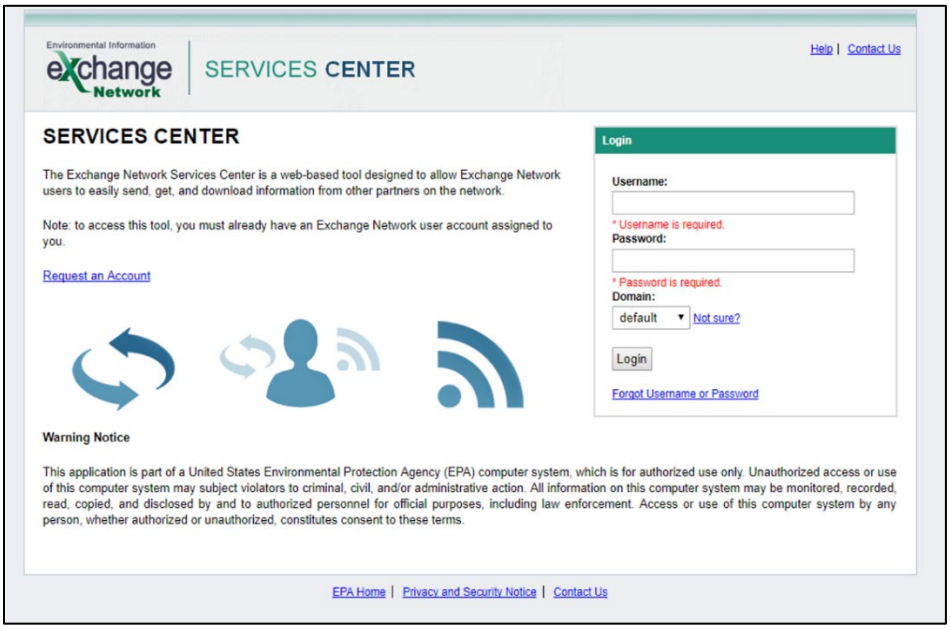
[aqs] <- generate_aqs_file([dartSets], out.dir = ['U:/CSN/Delivered/AQS/'], data.type = ['SAMPLE'], make.file = TRUE)

[aqs] <- generate_aqs_file([dartSets], out.dir = ['U:/CSN/Delivered/AQS/'], data.type = ['FIELD], make.file = TRUE)

with one or more appropriate DART data set IDs. For both the *aqsStaging2export* and *generate_aqs_file* functions, the analyst can instead provide start and end dates. The analyst should review the resulting output dataset for accuracy and consistency with AQS formatting rules. AQS reference documents are located on the shared network drive at U:/CSN/Documentation/Reference.

Once the AQS file is ready for delivery, the analyst will use a web browser of their choice and navigate to the EPA's Exchange Network Services website (Figure 2; https://enservices.epa.gov/login.aspx).

Figure 2. Login screen for the EPA's Exchange Network Services website.

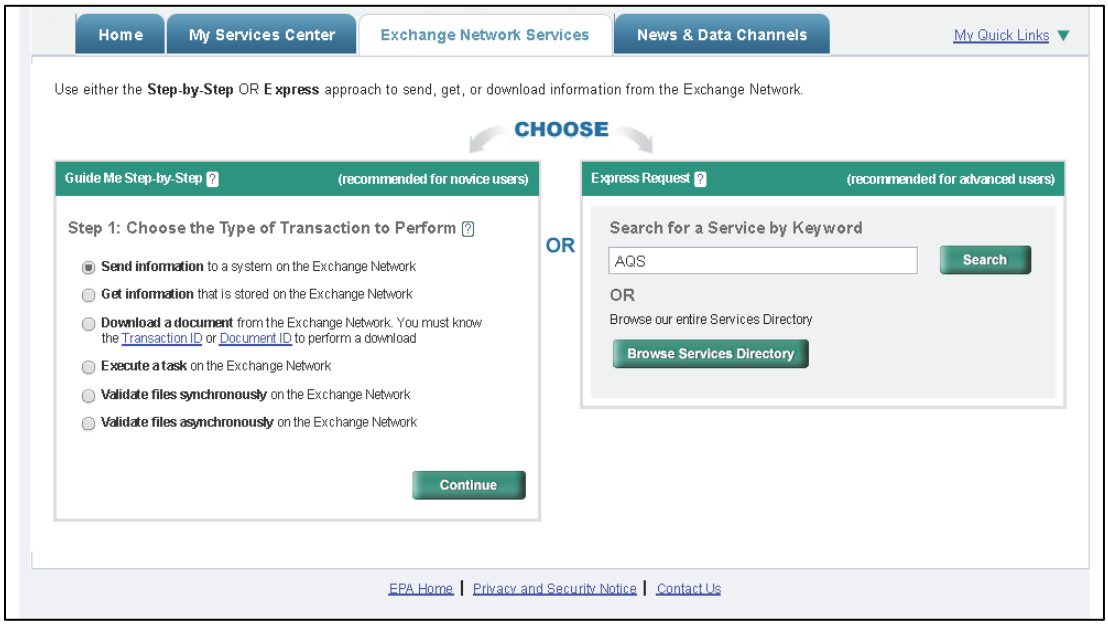


After logging in, the analyst will be presented with the home screen (Figure 3). To be efficient, the analyst should add the AQS service to the "My Quick Links" bar on their home screen. However, it is also possible for the analyst to search for the AQS submission form. To search, the analyst should click on the "Go" button of the Exchange Network Services bar.

Figure 3. Home screen of the Exchange Network Services website.

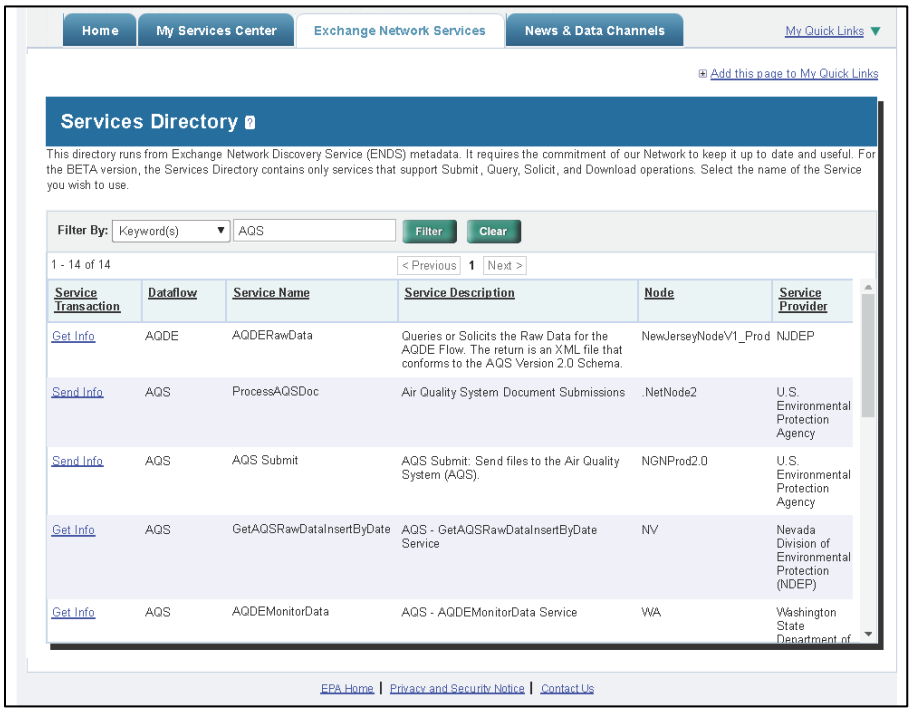
The analyst will then be presented with a choice between a Step-by-Step guide and a search bar (Figure 4). The analyst should type "AQS" into the search bar.

Figure 4. Enter "AQS" into the search bar.



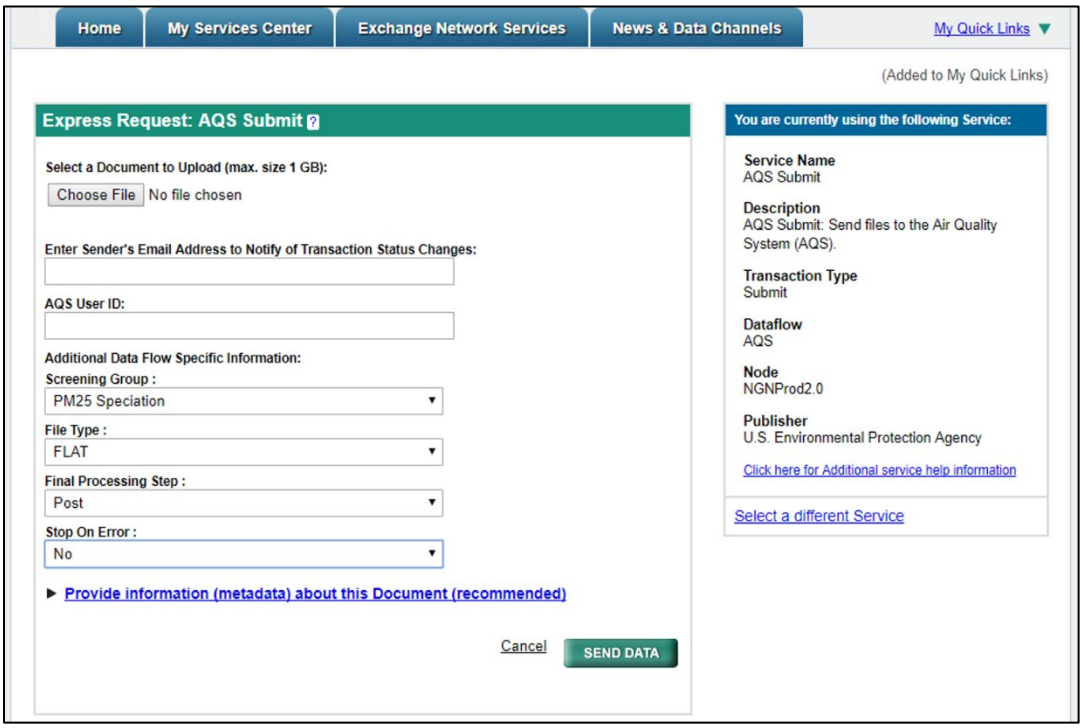
The search results will show all available processes associated with the AQS system (Figure 5). To access the AQS submission form, the analyst should choose the service that has “AQS Submit” specified in the “Service Name” field (usually the third option listed).

Figure 5. Search results from "AQS". The analyst should choose the Service named "AQS Submit".



Finally, the analyst will fill out the submission form with their email address, AQS user ID, screening group (PM2.5 Speciation), the file type (FLAT), the final processing step (LOAD), and whether or not to stop on errors (NO). See Figure 6 for an example. Use the "Choose File" button to select the sample data file generated from the previous step. Press the "SEND DATA" button to submit the form. The progress of the data submission can be monitored through the same web portal. The analyst should repeat this process to submit the field blank data file generated in the previous step.

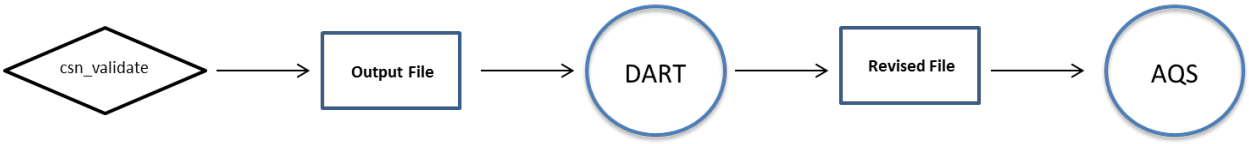
Figure 6. AQS data submission form.



9. DATA DELIVERY FLOW

This section describes the data flow through the data delivery process used to execute all CSN validation checks. Figure 7 outlines the flow of data from the UCD validation results to final delivery to AQS. The function *generate_dart_file* is executed by the analyst to generate the initial data set for DART review (see 8.1). Then upon receipt of the validated data from DART, the data is re-ingested in the UCD CSN database into the *DARTreturn* table. The analyst then executes the *dart2aqs* function to reformat the data into AQS format and populate the *delivery.AqsStaging* and *delivery.AqsStagingQualifierCodes* tables. Any requested changes are made by the analyst in these two tables. The final dataset is prepared by executing the *aqsStaging2export* followed by the *generate_aqs_file* function. The data is reviewed again and submitted to AQS through the Exchange Network Services web portal (see 8.3). Source code for the functions shown in Figure 7 is stored in the *crocker* source repository.

Figure 7. Flow diagram of the data delivery process. Rectangles represent data files, diamonds represent R functions, circles represent databases, and lines represent inputs and outputs.



10. EQUIPMENT AND SUPPLIES

The associated hardware and software used for CSN data delivery are described in the associated UCD CSN SOP #801.

11. QUALITY ASSURANCE AND QUALITY CONTROL

Software bugs and data management issues are tracked through JIRA tracking software. All users have access to our internal JIRA website and can submit, track, and comment on bug reports.

12. REFERENCES

UCD CSN SOP #801: Processing & Validating Raw Data