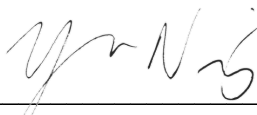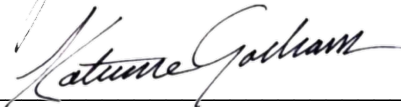# UCD CSN Technical Information #801A

CSN Data Ingest

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	1,2,7,8,9,10	Rewording for clarity and updating name changes.
7/30/19	YN, KAG	1,8	Changed process for carbon and ion analysis pathways, and wording changes for clarity.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. PURPOSE AND APPLICABILITY

The subject of this technical information document (TI) is handling electronic filter and laboratory records from samples collected in the Chemical Speciation Network (CSN). This document is intended to guide users on the receiving and validating of CSN filter and laboratory records and ingestion to the UCD CSN database. These include sample operational data and filter records from Wood PLC (Wood), ion analysis results from RTI International (RTI), and elemental carbon analysis results from the University of California, Davis (UCD).

2. SUMMARY OF THE METHOD

Filter records are received from the filter shipping and handling laboratory (Wood) in delivery files. These files are ingested into the UCD CSN database for subsequent calculation of concentrations and data validation. The UCD analyst uses the UCD CSN Data Management website to upload files and review the resulting output messages for errors.

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.
- **Database**: A normalized, relational data system designed to store unique information about each data point.

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 database programming and database management.

8.   PROCEDURAL STEPS

Three data ingest processes are required prior to data processing and validation.
1.   Filter records, including sample operational data and validity flags, from Wood.
2.   Ion analysis results from RTI.
3.   Elemental and carbon analysis results from UCD.

These three procedures are outlined below.

8.1   Filter records, sample operational data, and validity flags

Filter records are sent from Wood to UCD via email, typically on the same day as the shipment of corresponding physical filters. Filter records are delivered as three files:
1.   FilterDataTransfer_[xxx].csv,
2.   FilterDataNullFlags_[xxx].csv
3.   FilterDataValidFlags_[xxx].csv

Where [xxx] represents a number corresponding to the delivery batch. FilterDataTransfer contains a single record for each filter, including sample operational data such as flow rate and temperature. FilterDataNullFlags and FilterDataValidFlags include the null codes and validity codes, respectively. Null codes and validity codes are joined to corresponding filter data by the unique combination of SampleRequestID and ChannelID.

Filter records are ingested to the UCD CSN database through the UCD CSN Data Management website. Figure 1 shows a screenshot of the upload page. The data uploader first loads in "test only" mode, which will perform import validation, but will not save any changes to the database. Filter records are subjected to the automated validity checks as shown in Table 1. The data uploader will review the results of the validation and warn the analyst if any records fail to upload due to validation errors. Once the analyst has reviewed the output messages in the "test only" mode, the upload should be performed again with the "TestOnly" box unchecked to ingest the data into the database. After upload, the data uploader will store the source files on the UCD file server (U:\CSN\FromWood\Imported).

Figure 1. Filter data upload page from the UCD CSN Data Management website.

Table 1. Automated validity checks performed by the UCD CSN Data Management website during the filter data upload process.

Check	Action
Number of columns in header matches number of columns in row	Warning message
Any columns not found (or renamed)	Import aborted
Filter record matched more than one site or didn't match any sites	Warning message
More than one batch found in the import	Warning message
Number/date columns fail to parse into number/date	Warning message
Existing records in the database; if multiple matches generates message	Warning message
If matched existing record, checks for changed fields	Warning message
IntendedUseDate after SamplerStartDate	Warning message
SamplerStartDate more than a day after IntendedUseDate	Warning message
SamplerEndDate more than 25 hours after SamplerStartDate	Warning message
Calculated SamplerEventId doesn't match one in record	Warning message
SamplerEventId plus Channel position do not uniquely identify the record	Warning message
More than one LotNumber for teflon filters in the import	Warning message
Flow and ambient parameters are out of specification	Warning message
Calculated sample volume is not withing 10% of the reported sample volume	Warning message

Null codes and validity flags are uploaded through the data management website as shown in Figure 2. Filter records must be loaded prior to the null and validity codes. Files should first be loaded in "test only" mode, which will perform import validation, but will not save any changes to the database. Null codes and validity flags are subjected to the automated validity checks as shown in Table 2. The data uploader will review the results of the validation and warn the analyst if any records fail to upload due to validation errors. Similar to the previous step, the ingest process should be performed again with the "TestOnly" box unchecked. After ingest, the data uploader will store the source files on the file server (U:\CSN\FromWood\Imported).

Figure 2. Null code and validity flag upload page.

Table 2. Automated validity checks performed during the null code and validity flag upload process.

Check	Action
Number of columns in header matches number of columns in row	Warning message
Any columns not found (or renamed)	Import aborted
Flag record matched more than one filter or didn't match any filters	Warning message
SetNumber or IntendedUseDate don't match the matched filter record	Warning message
Number/date columns fail to parse into number/date	Warning message
Flag doesn't match existing AQS Code	Warning message
Flags apply to more than one batch	Warning message
More than one Null flag applies to filter (also create FilterComment). (Also ranks according flags according to rank and marks extra as duplicates)	Warning message
The same code is applied to a filter more than once	Warning message
NullCode import tries to use any Non-terminal codes. Also if QualifierCode import tries to use any terminal codes	Warning message

8.2   Ion Analysis Results

Ion analysis results are sent from RTI to UCD via email in .csv format.

8.2.1   Ions

The ions data are delivered in one csv file named after the batch of data that is being delivered:

1.   AXXXXXXX.csv (where the X's represents the batch number; for example, A0000053 for batch 53)

The csv analysis records are ingested to the database through the UCD CSN Data Management website. Figure 3 shows a screenshot of the IonsData upload page. The data

uploader first loads in "test only" mode, which will perform import validation, but will not save any changes to the database. Records are subjected to the automated validity checks as shown in Table 3. The data uploader will review the results of the validation and warn the analyst if any records fail to upload due to validation errors. The ingest process should be performed again with the "TestOnly" box unchecked. After upload, the data uploader will store the source files on the file server (U:\CSN\FromRTI\Imported).

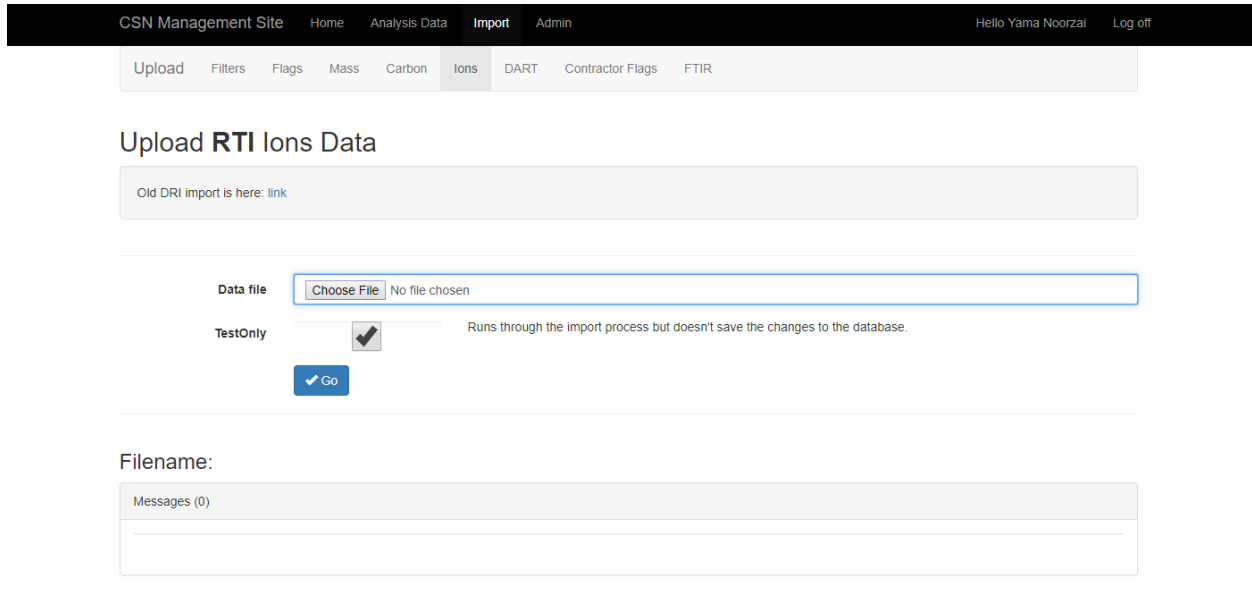Figure 3. Ion analysis results upload page.

Table 3. Automated validity checks performed during the IonData and IonInformation upload.

Check	Action
Basic schema validation on xml files	Warning message
No filter is found for record	Warning message
Multiple records for a parameter filter pair	Warning message
Parameter missing for a filter	Warning message
Parameter already recorded in database	Warning message
Import file does not use the same units for each parameter	Warning message
Filters belong to more than one batch	Warning message

8.3 Elemental and Carbon Analysis Results

Elemental and carbon analyses are performed at the AQRC.

8.3.1 Elemental Analysis

Results files created by the PANalytical XRF software are automatically ingested on a schedule by a software service. The results files are transmitted to a directory on the PC connected to the PANalytical XRF analyzer (C:\PANalytical\Transmission). A Windows Service (internally named XRF Data Transfer) is installed on each individual PC

connected to a PANalytical XRF analyzer and monitors the transmission directory checking it every hour for any files created. The results files are standard text files with the extension *.qan*. The file names are the XRF analysis dates and times in the format YYYYMMDDHHMMSS.qan. The results files and contents are parsed by the service and ingested into tables in the UCD CSN database.

8.3.2 Carbon Analysis

The software application used to run the carbon analyzer automatically stores data acquired during an analysis in a comma-delimited text format for later computation, display, and printing. The text file containing raw carbon data is automatically ingested into the UCD CSN database by a software service. Upon ingestion, the areal densities of OC (transmittance and/or reflectance), EC (transmittance and/or reflectance), and TC, as well as OC1, OC2, OC3, OC4, EC1, EC2, EC3 and OP (Pyrol C) (in μg C/cm^2) are automatically calculated and stored.

8.4 Mass Data

Filter masses for specific sites are determined at Wood and the results are sent to UCD via email as MassTransfter_[xxx].csv files, where [xxx] represents a number corresponding to the delivery batch. These files typically include the mass data for multiple analysis batches. Mass analysis data is ingested to the UCD CSN database through the UCD CSN Data Management website. Figure 4 shows a screenshot of the upload page. The data uploader will first load in "test only" mode. The data are subjected to the automated validity checks, which the data uploader will review and warn the analyst if any records fail to upload due to validation errors or there are any other issues with the data. After upload, the data uploader will store the source files on the file server (U:/CSN/FromWood/Imported/Mass).

Figure 4. Mass analysis results upload page.

8.5   **Re-ingesting**

In the event that data corrections are made by Wood or RTI, they will supply new files for ingestion. The new files will be uploaded using the same systems described above. The ingest processing will identify any changed records. The data validation analyst will first run the ingest process in test only mode and scrutinize the changed records to ensure that they are correct before re-running the process in overwrite mode. Only changed records will be overwritten.

9.   EQUIPMENT AND SUPPLIES

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

10.   QUALITY ASSURANCE AND QUALITY CONTROL

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

11.   REFERENCES

Not Applicable.