UCD CSN Technical Information #801A

CSN Data Ingest

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

September 28, 2017
Version 1.0

Prepared By: ______________________ Date: 10/16/2017

Reviewed By: ______________________ Date: 10/16/2017

Approved By: ______________________ Date: 10/17/2017

UCDAVIS
AIR QUALITY RESEARCH CENTER

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications

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 CSN network. This document is intended to guide users on the receiving and validating of CSN filter and laboratory records and ingestion to the CSN database. These include sample operational data and filter records from Amec, carbon and ion analysis results from DRI, and elemental analysis results from the UC Davis laboratory.

2. SUMMARY OF THE METHOD

Filter records are received from the filter shipping and handling laboratory (Amec) in delivery files. These files are ingested into the CSN database for subsequent calculation of concentrations and data validation. The UC Data Analyst will use the 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 AQRC data management staff assigned to this project all have advanced training in database programming and database management. All have direct experience through recent involvement in designing and managing a similar database for IMPROVE.

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 Amec.
2. Carbon and ion analysis results from DRI.
3. Elemental analysis results from UC Davis.

These three procedures are outlined below.

8.1 Filter records, sample operational data, and validity flags

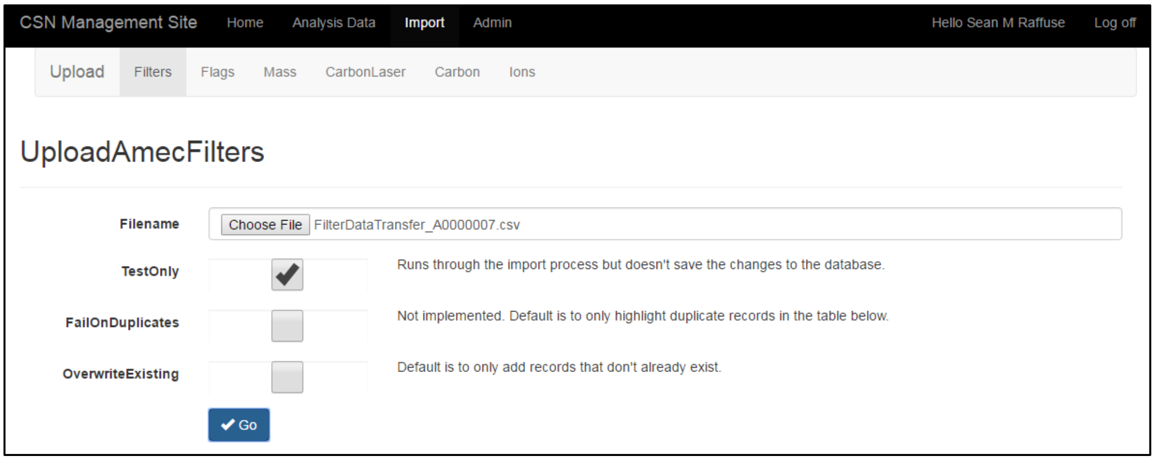
Filter records are sent from Amec to UC Davis via email to the Data Analyst and the UC Davis sample handling lab, typically on the same day as the shipment of corresponding physical filters. Filter records are delivered as three files:

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 CSN database through the CSN Data Management website. Figure 1 shows a screenshot of the upload page. The data uploader will first load 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 file server (U:\CSN\FromAmec).

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



UploadAmecFilters

Filename: Choose File FilterDataTransfer_A0000007.csv

TestOnly ☑ Runs through the import process but doesn't save the changes to the database.

FailOnDuplicates ☐ Not implemented. Default is to only highlight duplicate records in the table below.

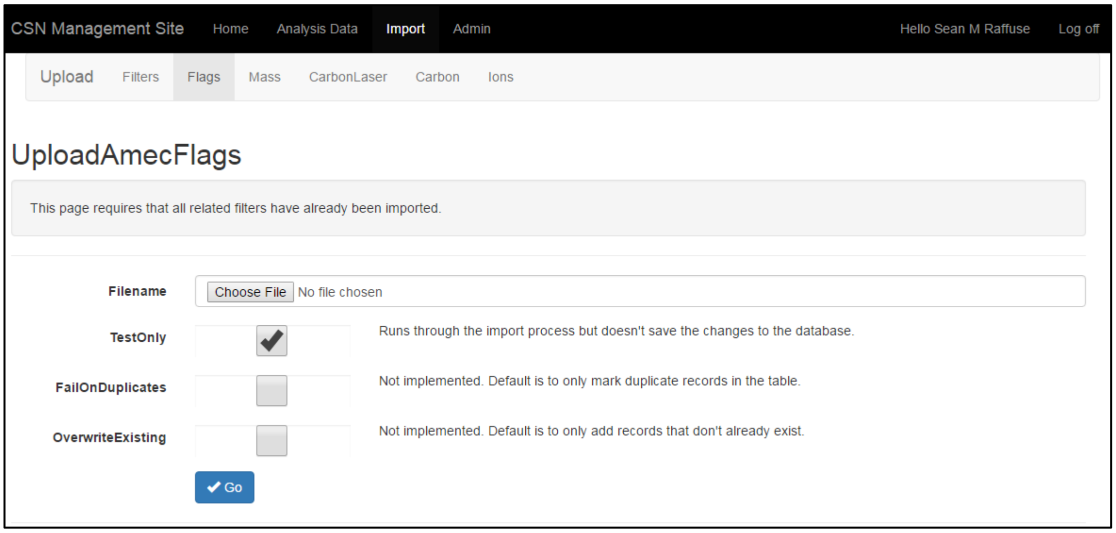
OverwriteExisting ☐ Default is to only add records that don't already exist.

Table 1. Automated validity checks performed by the 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

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\FromAmec).

Figure 2. Null code and validity flag upload page.



UploadAmecFlags

This page requires that all related filters have already been imported.

Filename: No file chosen

TestOnly ☑ Runs through the import process but doesn't save the changes to the database.

FailOnDuplicates □ Not implemented. Default is to only mark duplicate records in the table.

OverwriteExisting □ Not implemented. Default is to only add records that don't already exist.

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 Carbon and Ion Analysis Results

Carbon and ion analysis results are provided by DRI via email to the Data Manager and Data Analyst in .xml format.

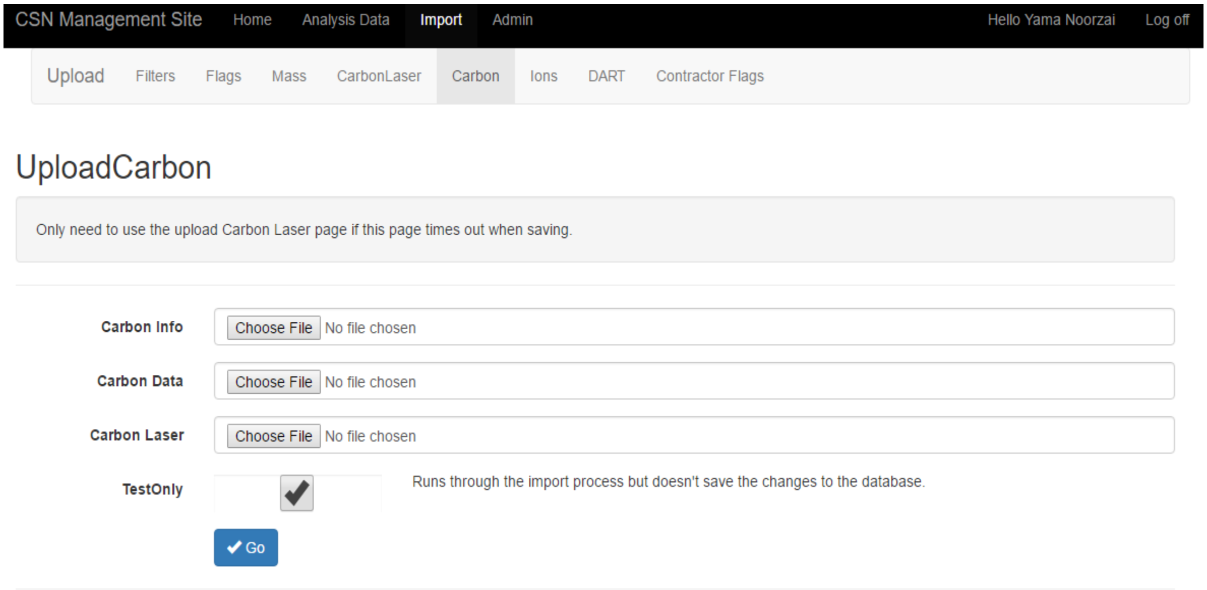
8.2.1 Carbon

The carbon data are delivered in three files:

1. CarbonData.xml
2. CarbonInformation.xml
3. CarbonLaser.xml

All three files are ingested to the database through the CSN Data Management website. Figure 3 shows a screenshot of the CarbonData upload page. The data uploader will first load in “test only” mode, which will perform import validation, but will not save any changes to the database. CarbonInformation, CarbonLaser and CarbonData are ingested simultaneously. 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\FromDRI).

Figure 3. Carbon analysis results upload page.



UploadCarbon

Only need to use the upload Carbon Laser page if this page times out when saving.

Carbon Info: Choose File No file chosen

Carbon Data: Choose File No file chosen

Carbon Laser: Choose File No file chosen

TestOnly ☑ Runs through the import process but doesn't save the changes to the database.

✔ Go

Table 3. Automated validity checks performed during the CarbonLaser and CarbonData upload.

Check	**Action**
CarbonLaser and Carbon	
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
Carbon only	
No CarbonLaser entry for filter	Warning message

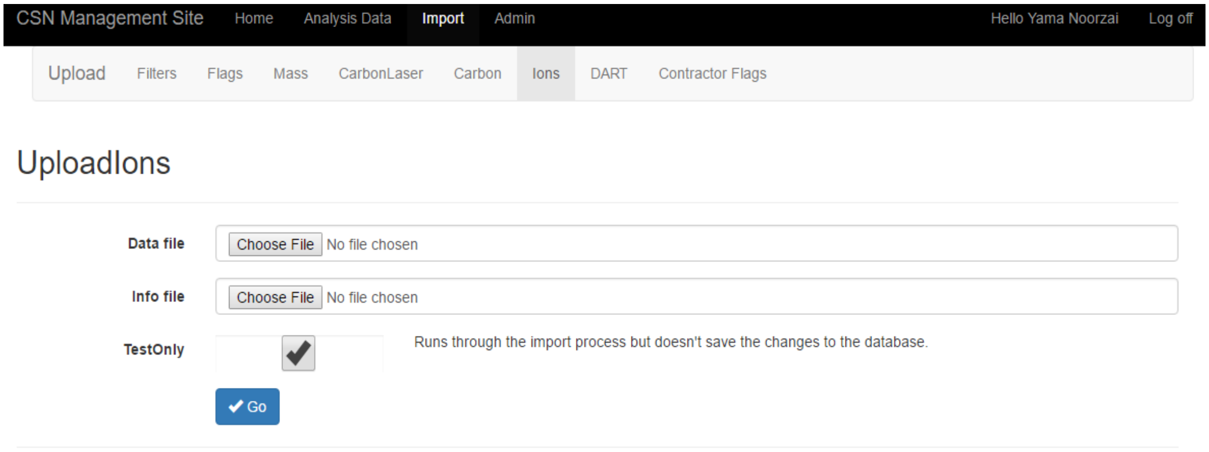
8.2.2 Ions

The ions data are delivered in two files:

1. IonData.xml
2. IonInformation.xml

Both the IonData and Ion information analysis records are ingested to the database through the CSN Data Management website. Figure 4 shows a screenshot of the IonsData upload page. The data uploader will first load 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 4. 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\FromDRI).

Figure 4. Ion analysis results upload page.



UploadIons

Data file: Choose File No file chosen

Info file: Choose File No file chosen

TestOnly ☑ Runs through the import process but doesn't save the changes to the database.

Table 4. 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 Analysis Results

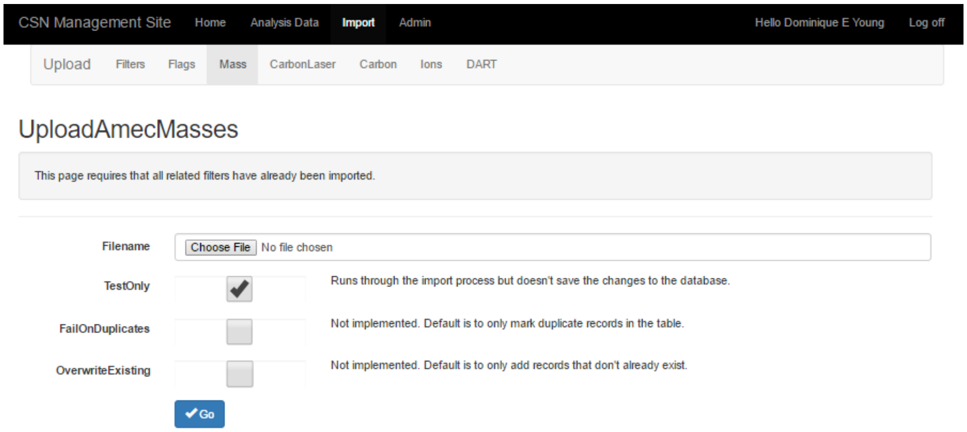
Elemental analysis is performed at the AQRC XRF lab. Results files created by the PANalytical XRF software are automatically ingested on a schedule by a software service.The Results files are transmitted to a directory on the PC connected to the

PANalytical XRF analyzer (C:\PANalytical\Transmission).A Windows Service (that we have 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 CSN database.

8.4 Mass Data

Filter masses for specific sites are determined at AMEC and the results are sent to UC Davis via email to the Data Analyst as the MassTransfter_[xxx]_[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 CSN database through the CSN Data Management website. Figure 5 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/FromAmec/Imported/Mass).

Figure 5. Ion analysis results upload page.



UploadAmecMasses

This page requires that all related filters have already been imported.

Filename: No file chosen

TestOnly ☑ Runs through the import process but doesn't save the changes to the database.

FailOnDuplicates □ Not implemented. Default is to only mark duplicate records in the table.

OverwriteExisting □ Not implemented. Default is to only add records that don't already exist.

8.5 Reingesting

In the event that corrections must be made by Amec or DRI, 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. Briefly, CSN data are stored within a Microsoft SQL Server database. Data management is handled through custom software that interfaces with the CSN database. The primary applications for data ingest and management were developed on the .NET platform. In addition, to support data validation and operational monitoring, several interactive visualizations have been developed using the R Shiny platform.

10. QUALITY ASSURANCE AND QUALITY CONTROL

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

11. REFERENCES

Not Applicable.