

UCD CSN Standard Operating Procedure #801

Processing & Validating Raw Data

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

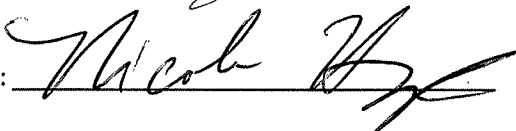
*September 28, 2017
Version 1.0*

Prepared By: 

Date: 10/16/2017

Reviewed By: 

Date: 10/16/2017

Approved By: 

Date: 10/16/17

Table of Contents

1. PURPOSE AND APPLICABILITY	4
2. SUMMARY OF THE METHOD	4
3. DEFINITIONS.....	5
4. HEALTH AND SAFETY WARNINGS	5
5. CAUTIONS	5
6. INTERFERENCES	5
7. PERSONNEL QUALIFICATIONS, DUTIES, AND TRAINING	5
8. PROCEDURAL STEPS	6
9. EQUIPMENT AND SUPPLIES	7
10. QUALITY ASSURANCE AND QUALITY CONTROL.....	8
10.1 Code Development	8
10.2 Bug Reporting	9
10.3 Data Validation.....	9
11. REFERENCES.....	9

1. PURPOSE AND APPLICABILITY

This Standard Operating Procedure (SOP) broadly outlines the procedures applied at Air Quality Research Center (AQRC) for processing and validating the sampling and analytical laboratory data from the U.S. EPA Chemical Speciation Network (CSN). Data processing and validation for CSN are the responsibility of the data management group at AQRC, under the supervision of the project Data Manager.

This SOP covers the steps involved in receiving the sampling and analytical laboratory data, processing the data into a format suitable for further review, conducting Level 0 and Level 1 validation, submitting the data to state and local agencies for their further validation and review, final processing and review of state changes, and submittal of the data to the EPA's Air Quality System (AQS) database.

This document is intended to give only the outline of how data are processed, validated, and delivered. Each of the required steps involved has a specific function and a set of procedures associated with that function. A detailed explanation of each of these steps is required. Thus, descriptions of the individual procedures are given in the Technical Information (TI) documents that are referenced within this SOP.

2. SUMMARY OF THE METHOD

Filter samples are collected routinely throughout the year in CSN, resulting in approximately 13,000 annual samples on each of three types of filters (PTFE, nylon, and quartz). Field sampling is conducted by representatives of state and local agencies. Filter packs are prepared and sent to the field, and then received after sampling, by a separate contractor, Amec-Foster Wheeler. Once the samples are received, Amec sends the exposed filters to AQRC and to our subcontracted laboratory, Desert Research Institute (DRI), along with associated sampling data such as flow volumes and sampling duration.

Samples are analyzed at AQRC for elements by x-ray fluorescence (XRF) on the PTFE filters and at DRI for ions by ion chromatography on the nylon filters and for carbon by a thermal optical method on the quartz filters. Following laboratory analysis all analytical results are assembled by AQRC for processing and initial validation.

Data processing involves calculating an ambient concentration, uncertainty, and MDL for each analyte using the laboratory result plus the sample volume and sampling duration determined from the field data. The calculated concentrations undergo two levels of validation at AQRC. Level 0 validation examines the fundamental information associated with each measured variable, such as chain of custody, shipping integrity, sample identification, and damaged samples. Level 1 data are reviewed more fully for technical acceptability and reasonableness based on information such as routine QC sample results, data quality indicator calculations, performance evaluation samples,

internal and external audits, statistical screening, internal consistency checks, and range checks.

Once the data have been processed and validated to Level 1 by AQRC they are submitted to the state and local agencies for further review and Level 2 and 3 validation.

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.
- **Ion Chromatography (IC):** An analytical technique used to determine the concentration of ions in a sample.
- **Interagency Monitoring of Protected Visual Environments (IMPROVE):** Federal PM_{2.5} and PM₁₀ sampling network directed by the National Park Service, with sites located principally in remote rural areas.
- **Thermal Optical Reflectance (TOR):** An analytical technique used to determine the concentration of carbon in a sample.
- **X-ray Fluorescence (XRF):** An analytical technique used to determine the concentration of elements in a sample.

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. The roles and responsibilities are as follows:

The Data and Reporting Group Manager oversees all aspects of data validation and reporting. Under their direction data validation analysts are responsible for data validation and submission, with specific responsibilities including:

- Receiving electronic data from Amec and DRI and ingesting records to the CSN database;
- Executing data processing code to calculate ambient concentrations;
- Reviewing the components of the measurements (flow rates, elemental concentration, etc.) in preparation for final data validation;
- Working with others in laboratory operations to resolve problems or discrepancies encountered during data review;
- Communicating with the filter handling lab and SLT validators to resolve issues;
- Validating the final data set, with input as needed from data analysts;
- Formatting the data to meet AQS standards; and
- Submitting the final data sets.

The Software and Analysis Group Manager oversees database and software development. Under their direction, software developers are responsible for:

- Maintaining and upgrading the data management system including the SQL Server database, data processing and visualization tools, and data reporting and data input forms;
- Working with staff to identify, map, design and implement improvements to the data management system;
- Testing, verifying, and documenting modifications to the system; and
- Designing and maintaining an archival system for all data and metadata records and source files.

8. PROCEDURAL STEPS

UCD CSN data processing and validation occurs in several steps, outlined below. The specifics of each step are detailed in the noted Technical Information document.

- 1) Data ingest (CSN TI 801A): Sample event information (including Filter ids, flow rates, flags, and comments) are retrieved from Amec via email and uploaded to the CSN database. XRF results are transferred into the database through an automated service. IC and TOR analysis results files are received via email from DRI. Results are ingested to the CSN database.
- 2) Level 0 Validation (CSN TI 801C): Data and metadata are reviewed through several visualizations to identify oddities such as inconsistent dates that appear to be typos. These are resolved through communication with Amec.
- 3) Data Processing (CSN TI 801B): Flow rates and analysis results are combined to calculate concentrations. Blank values are used to derive MDLs. MDLs and concentrations are used to estimate uncertainty.

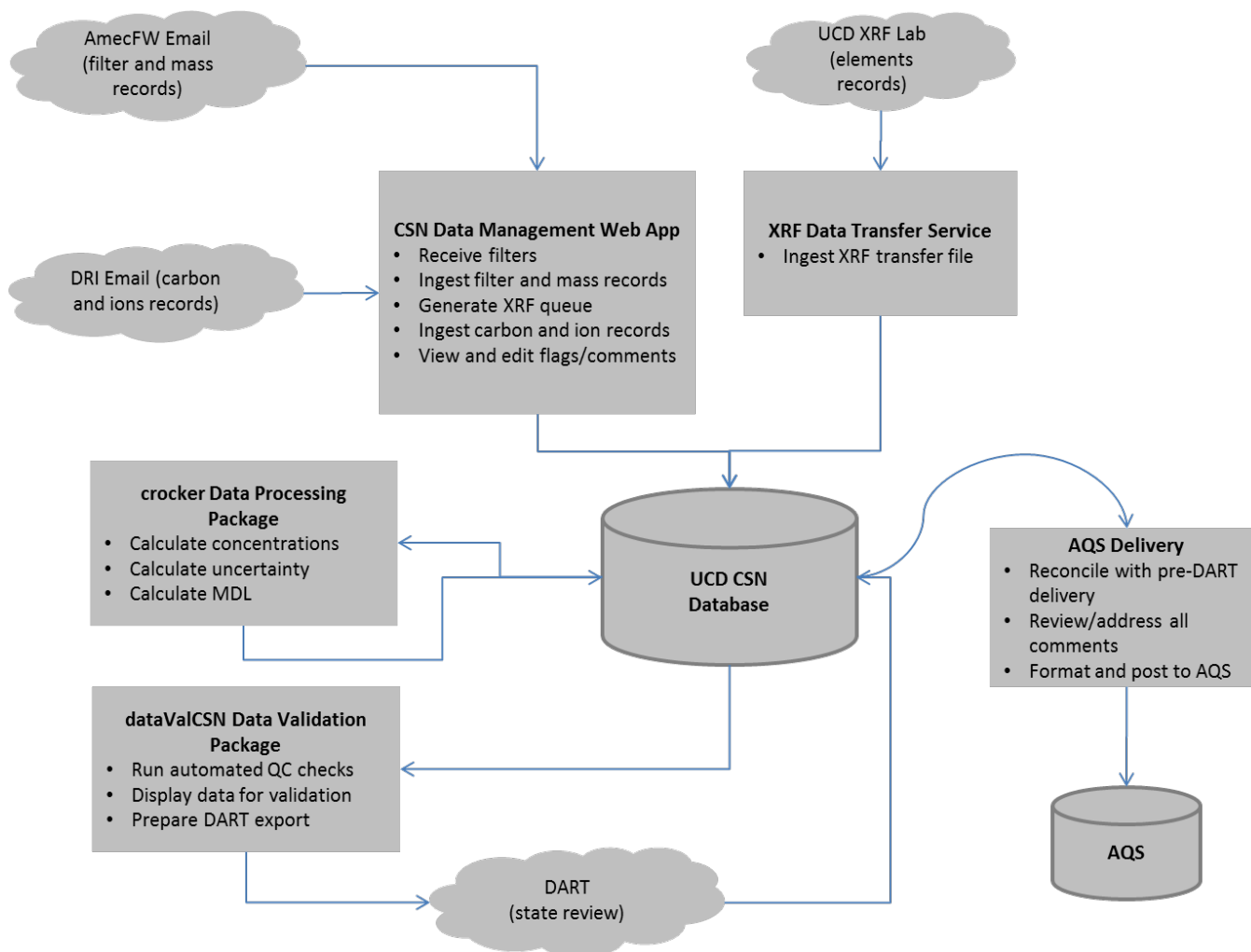
- 4) Level 1 Data Validation (CSN TI 801C): Several statistical and visual checks are applied and examined. Reanalyses are requested as needed. Data are flagged with qualifier or null codes.
- 5) Data Posting (CSN TI 801D): Initially validated concentration data and metadata are posted for state review to EPA's Data Analysis and Reporting Tool (DART). After the specified 30 day review period, changed or unchanged data are re-ingested to the CSN database.
- 6) AQS Delivery (CSN TI 801D): State initiated changes and comments are reviewed and resolved. Data are formatted for delivery to AQS and posted.

9. EQUIPMENT AND SUPPLIES

The CSN data are stored within a Microsoft SQL Server database. The database software is installed on a Rackform iServe R346.v4 hardware with RAID 10 data drives. Three virtual machines are installed on the server hardware for production, development, and testing.

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. Figure 1 illustrates the data flow and relationships between the data sources, software, and the CSN database. In addition, to support data validation and operational monitoring, several interactive visualizations have been developed using the R Shiny platform. These are discussed in their relevant Technical Information documents.

Figure 1. Diagram of CSN data management software and flow at UCD.



10. QUALITY ASSURANCE AND QUALITY CONTROL

10.1 Code Development

Software for data management, processing, and validation is developed in-house by professional software engineers. Source code is managed through a code repository. Development of code changes and new applications is conducted on a development environment that parallels the production environment. Prior to deployment in production, all code changes undergo testing within a separate test environment. The testing, which is conducted by developers, managers, and users, is targeted both at the identification of software bugs and the confirmation of valid data equivalent to the production system.

10.2 Bug Reporting

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.

10.3 Data Validation

Data integrity is enforced within the database via unique primary keys and non-nullable records. Data completeness and data quality are thoroughly checked through the data validation process, described in the TI documents.

11. REFERENCES

Not Applicable.