

UCD CSN Standard Operating Procedure #801

Processing & Validating Raw Data

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

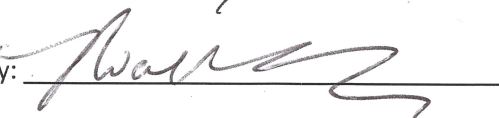
*November 30, 2018
Version 1.1*

Prepared By: 

Date: 11/28/2018

Reviewed By: 

Date: 11/28/2018

Approved By: 

Date: 11/28/18

DOCUMENT HISTORY

Date Modified	Initials	Section/s Modified	Brief Description of Modifications
11/30/18	NJS	1,2,3,7,8,9,10	Rewording for clarity and updating name changes. Included disaster recovery plan.

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
9.1 Disaster Recovery Plan	8
9.1.1 Facility Recovery	8
9.1.2 Hardware Recovery Plan	9
9.1.3 Software and Data Recovery Plan	9
9.1.4 Data Security.....	10
10. QUALITY ASSURANCE AND QUALITY CONTROL.....	10
10.1 Code Development.....	10
10.2 Bug Reporting	10
10.3 Data Validation	10
11. REFERENCES	10

1. PURPOSE AND APPLICABILITY

This Standard Operating Procedure (SOP) broadly outlines the procedures applied at the Air Quality Research Center (AQRC) for processing and validating the sampling and analytical laboratory data from the U.S. Environmental Protection Agency (EPA) Chemical Speciation Network (CSN). Data processing and validation for CSN are the responsibility of the Data & Reporting Group at AQRC, under the supervision of the project Data & Reporting 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, local, and tribal (SLT) 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 SLT agencies. Filter packs are prepared and sent to the field, and then received after sampling, by a separate contractor, Wood PLC (Wood). Once the samples are received, Wood 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 on the PTFE filters by x-ray fluorescence (XRF) and at DRI for ions on the nylon filters by ion chromatography on the nylon filters and for carbon on the quartz filters by a thermal optical method. Following laboratory analysis, all analytical results are assembled by AQRC for processing and initial validation.

Data processing involves calculating 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 SLT 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.
- **STI:** Sonoma Tech, Inc. Contractor developing and operating the DART interface.
- **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 UCD Air Quality Research Center (AQRC) Data & Reporting Group staff assigned to this project all have advanced training in database programming and database management. The roles and responsibilities are as follows:

The Data & 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 Wood 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 to the AQS database.

The Software & 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 Wood 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 UCD CSN database.
- 2) Level 0 Validation (CSN TI 801C): Data and metadata are reviewed through several visualizations to identify oddities such as inconsistent dates, transcription errors, and others that appear to be typographical errors. These are resolved through communication with Wood.
- 3) Data Processing (CSN TI 801B): Flow rates and analysis results are combined to calculate concentrations. Field blank values are used to derive MDLs and correct data for artifacts. MDLs and concentrations are used to estimate uncertainty.

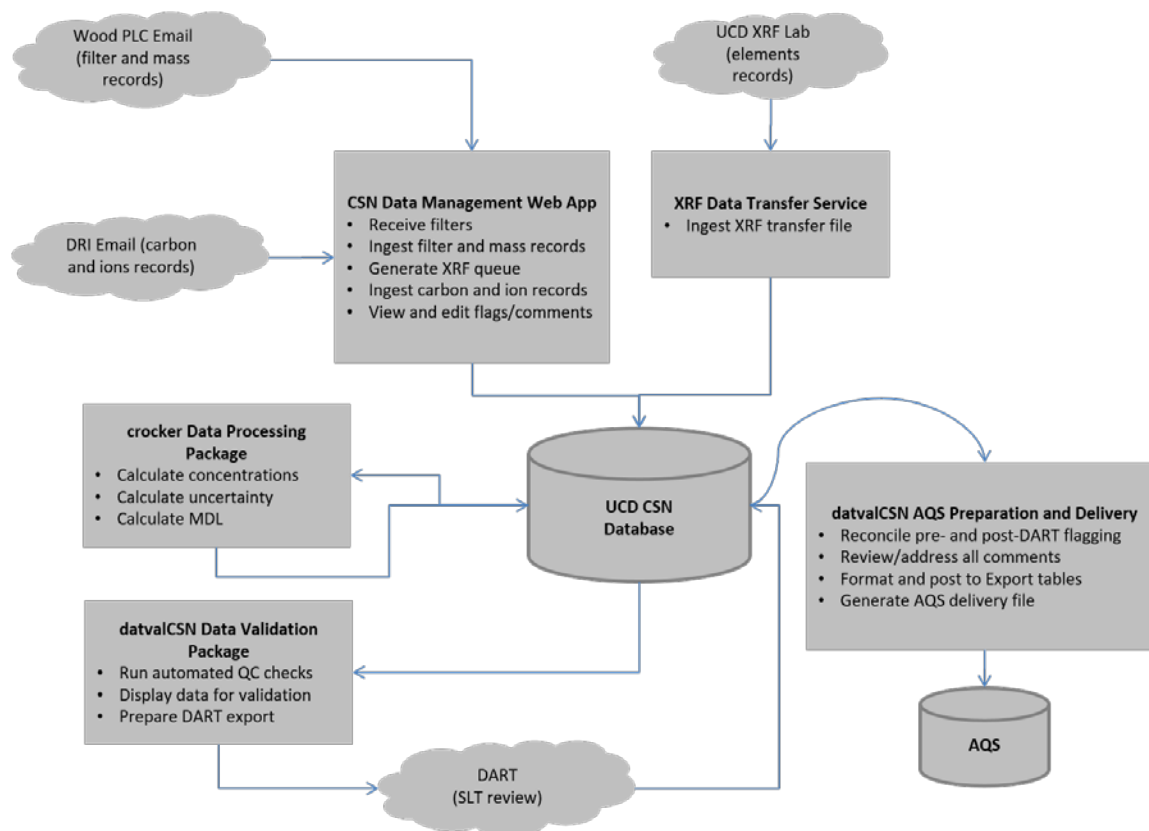
- 4) Level 1 Data Validation (CSN TI 801C): Several statistical and visual checks are applied and examined. Reanalyses are requested as needed. Data are flagged with qualifier or null codes.
- 5) Data Posting (CSN TI 801D): Initially validated concentration data and metadata are posted for state review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

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 UCD 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 UCD 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.



9.1 Disaster Recovery Plan

The scope of recovery activities will depend on the nature of the disaster. Response to an actual disaster may require implementing multiple sections of this SOP.

9.1.1 Facility Recovery

The UC Davis police department patrols buildings on a regular basis (including nights, weekends, and holidays). In addition, campus facilities and maintenance staff are on call at all times.

In the event of damage to the Jungerman Hall data facilities, the UC Davis police will notify the Information Technology (IT) Administrator. The IT Administrator will assess the damage to determine the scope of recovery operations. If the building can be safely entered, surviving equipment will be relocated to another building. All buildings on the UC Davis campus are connected to internal Ethernet, and a relocated server could be immediately operable.

If equipment is substantially damaged, arrangements will be made to relocate activities on other UC Davis servers and/or acquire new hardware.

9.1.2 Hardware Recovery Plan

Database and file servers: The campus network of IT Administrator staff allow for rapid response to server failure and recovery issues.

Bar-code scanners: Bar-code scanners are used to record sample information. In an emergency, a keyboard could be used for data entry rather than a bar-code scanner. Bar-code scanner replacements are available on short notice.

XRF system computers: Each XRF instrument has an associated computer. Instrument service contracts with PANalytical for each instrument guarantee service within 48 hours, enabling quick replacement of XRF computers with little disruption to the flow of samples.

9.1.3 Software and Data Recovery Plan

9.1.3.1 UCD XRF

Raw and processed spectra are saved and available for use at any time on the Epsilon 5 computers. Data safety and security are ensured by frequent transfer of computerized raw data from the Epsilon 5 PCs in the CNL XRF Laboratory (Jungerman Hall) to two different servers located in the CNL and LAWR buildings on campus. Differential backups are performed daily and full backups are performed weekly.

9.1.3.2 DRI Ions and Carbon

Raw data files are automatically backed up to a virtual file and database server, which is run on a physical clustered RAID 1 (Mirror) server, once a day. Once data is on the server it is stored in an instantly accessible, un-modifiable directory for 35 days and an instantly accessible, modifiable directory for 10 days. All data in these locations begin as exact copies of data that was on each individual laboratory computer. After data is safely in those locations, the raw data is extracted from the files and imported to the database server for possible modification. After data has been on the server for 35 days, it is automatically written to tape and stored indefinitely. Daily e-mails are automatically generated to confirm backups and notify computer personnel of data processing and data management issues.

All hard drives and tape, once filled, are stored in a special media storage room. The room has no windows, no drop ceilings, and is buried in a side of a hill in the lower section of the DRI building. It also contains UV filters on the lights to prevent damage to media.

Newer analytical instruments typically have frequent software modifications to provide enhanced data processing and review capabilities. The DRI EAF archives major software modifications for analytical instruments and maintains computers to run them in order have

the ability to reprocess or review older data. Similar archiving applies to legacy systems and software for analytical systems no longer being used.

9.1.4 Data Security

UCD and DRI access policies: Access to database and computers associated with this project is limited to authorized project personnel by use of access control lists for files, programs, and database access. Access to laboratory and office space is controlled by keycards.

Password policies: Unique passwords are issued to each employee by the UC Davis campus system administrator. Password integrity is monitored by the UC Davis campus system administrator.

Termination policies: System access is revoked for terminated personnel. The IT Administrator disables domain accounts and passwords upon termination of employment.

Virus protection: Microsoft Endpoint Protection is used for virus scanning and protection. All staff are required to complete annual cyber security awareness training.

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 tracking software. All UCD users have access to an internal JIRA website and can submit, track, and comment on bug reports.

10.3 Data Validation

Data integrity is enforced within the UCD CSN 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.