

Section 508 Compliant  Yes  No

Processing & Validating Raw Data

UCD SOP #801, Version 1.5

October 31, 2022

Page 1 of 10

# UCD CSN Standard Operating Procedure #801

## Processing & Validating Raw Data

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

*October 31, 2022*

*Version 1.5*

Prepared By: _____	DocuSigned by: <i>Alex Murrain</i> 4D6ECE5415824A8...	Date: _____ 11/15/2022
Reviewed By: _____	DocuSigned by: <i>Dominique Young</i> BB55DBA34BAB407...	Date: _____ 11/15/2022
Approved By: _____	DocuSigned by: <i>Marcus Langston</i> 0A10CFCF79B0452...	Date: _____ 11/15/2022



**DOCUMENT HISTORY**

<b>Revision</b>	<b>Release Date</b>	<b>Initials</b>	<b>Section/s Modified</b>	<b>Brief Description of Modifications</b>
1.1	11/30/18	NJS	1,2,3,7,8,9,10	Rewording for clarity and updating name changes. Included disaster recovery plan.
1.2	7/31/19	KAG	1,2,3,7,8,9	Changed process for carbon and ion analysis pathways, and wording changes for clarity.
1.3	3/5/20	KAG	4,5,6,7,9	Removed laboratory hardware from Section 9.1.2. Wording changes for clarity.
1.4	3/31/21	DEY, AMM	1,8	Added information on additional data deliveries. Updated TI document references.
1.5	10/31/22	DEY, AMM	2,3,8,9	HIPS added, Figure 1 updated.

## 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.....	6
7. Personnel Qualifications .....	6
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 .....	8
9.1.3 Software and Data Recovery Plan .....	9
9.1.4 Data Security.....	9
10. Quality Assurance and Quality Control.....	9
10.1 Code Development .....	9
10.2 Bug Reporting.....	10
10.3 Data Validation.....	10
11. References.....	10

## LIST OF FIGURES

Figure 1. Diagram of CSN data management software and flow at UCD.....	8
--	---

## 1. PURPOSE AND APPLICABILITY

This Standard Operating Procedure (SOP) broadly outlines the procedures applied 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 within the Air Quality Research Center (AQRC) at University of California, Davis (UCD); the AQRC Data & Reporting Manager supervises the project.

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 SLT changes, and submittal of the data to the EPA's Air Quality System (AQS) database, UCD's CSN/IMPROVE Archive (CIA) database, and the U.S. National Park Service.

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 – per the EPA defined schedule – throughout the year in the CSN, resulting in approximately 13,000 annual samples on each of three types of filters (PTFE, nylon, and quartz). Field sampling is conducted by the SLT agencies. Filter packs are prepared and sent to the field, and then received after sampling by the Sample Handling Laboratory contractor, Wood PLC (Wood). Once the samples are received, Wood sends the exposed filters to UCD laboratory and to the UCD subcontract laboratory, RTI International (RTI), along with associated operational sampling data such as flow volumes and sampling duration.

Samples are analyzed at Wood for PM<sub>2.5</sub> mass on the PTFE filters by gravimetric analysis before sending to UCD laboratory. Samples are analyzed at UCD for elements on the PTFE filters by energy dispersive X-ray fluorescence (EDXRF), for filter optical absorption on the PTFE filters by Hybrid Integrating Plate and Sphere system (HIPS) and for carbon on the quartz filters by thermal optical analysis (TOA). As a subcontractor to UCD, samples are analyzed at RTI for ions on the nylon filters by ion chromatography (IC). Following laboratory analysis, all analytical results are assembled by UCD for processing and initial validation.

Data processing involves calculating ambient concentration, uncertainty, and method detection limit (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 UCD, Level 0 and Level 1. 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 via the Data Analysis and Reporting Tool (DART), hosted by Sonoma Technology, Inc. (STI), 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<sub>2.5</sub> sampling network, with sites located principally in urban areas.
- **CSN database:** A SQL Server database that is the central warehouse of CSN preliminary and final data at UCD.
- **Data Analysis and Reporting Tool (DART):** An interface for data review hosted by STI.
- **Energy Dispersive X-Ray Fluorescence (EDXRF):** An analytical technique used to determine the concentration of elements.
- **Hybrid Integrating Plate and Sphere system (HIPS):** An analytical technique for filter optical absorption.
- **Interagency Monitoring of Protected Visual Environments (IMPROVE):** Federal PM<sub>2.5</sub> and PM<sub>10</sub> sampling network directed by the National Park Service, with sites located principally in remote rural areas.
- **Ion Chromatography (IC):** An analytical technique used to determine the concentration of ions.
- **Method Detection Limit (MDL):** A lower limit of detection specific to method of analysis and reported parameter.
- **STI:** Sonoma Technology, Inc. Contractor developing and operating the DART interface.
- **Thermal Optical Analysis (TOA):** An analytical technique used to determine the concentration of carbon.

### 4. HEALTH AND SAFETY WARNINGS

Not applicable.

### 5. CAUTIONS

Not applicable.

## 6. INTERFERENCES

Not applicable.

## 7. PERSONNEL QUALIFICATIONS

The UCD AQRC Data & Reporting Group staff assigned to this project have the following roles and responsibilities:

- 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, RTI, and STI 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;
  - Communicating with laboratory staff at both UCD and RTI to resolve problems or discrepancies encountered during data review;
  - Communicating with the Sample Handling Laboratory (Wood), STI, and SLT validators to resolve issues;
  - Validating the final data set;
  - Formatting the data to meet AQS standards; and
  - Submitting the final data sets to the AQS database, CIA database, and NPS.
- 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 documents.

1. Data ingest (CSN TI 801A): Sample event information (including Filter IDs, flow rates, flags, and comments) and gravimetric mass results are retrieved from Wood via email and uploaded to the UCD CSN database. EDXRF, HIPS, and TOA results are transferred into the database through an automated service. IC analysis results files are received via email from RTI, and results are ingested to the UCD CSN database.

Electronic documents are official. Paper copies are for reference only.

2. Level 0 Validation (CSN TI 801C): Data and metadata are reviewed through several statistical checks and 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 and SLT validators.
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, concentrations, and collocated results 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<sup>1</sup>.
5. Data Posting (CSN TI 801D): Initially validated concentration data and metadata are posted for SLT 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 801E): SLT initiated changes and comments are reviewed and resolved. Data are formatted for delivery and posted to the AQS database, CIA database, and NPS.

## 9. EQUIPMENT AND SUPPLIES

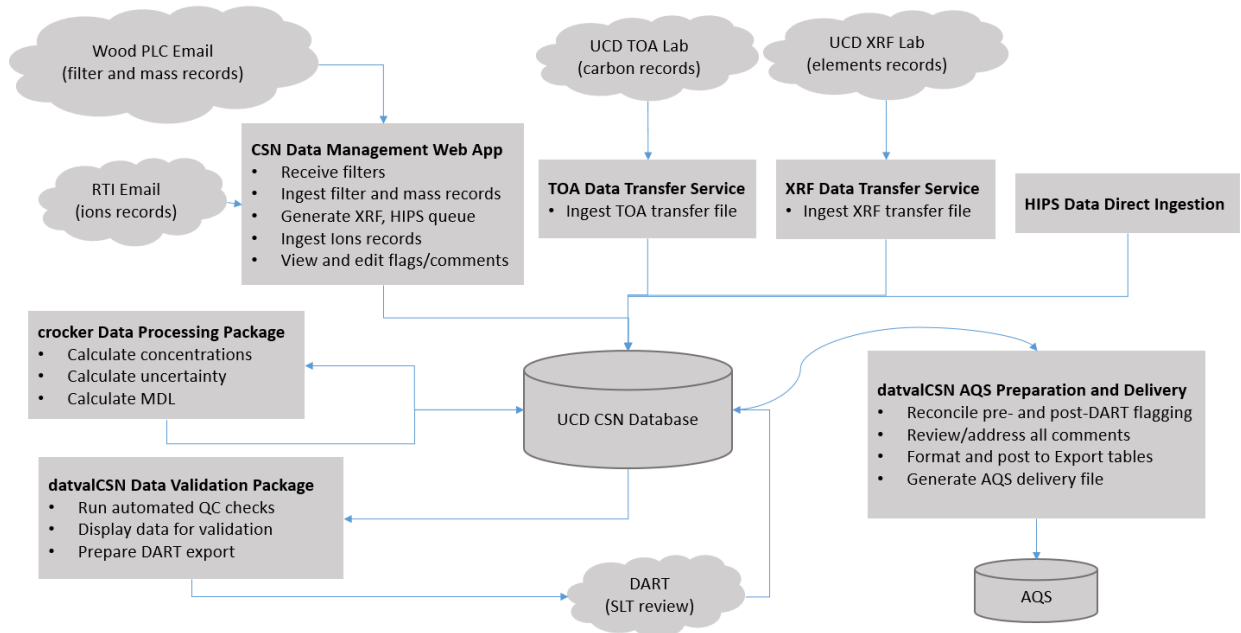
The CSN data are stored in Microsoft SQL Server Databases at UC Davis. The production database is run on a dedicated Windows Server with a RAID array for storage and with offsite backups. The development and test database environments are virtual machines. To test back up recovery, the development and testing environments are regularly restored from the production backups.

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.

---

<sup>1</sup> The terms qualifier flags/codes and validity flags/codes may be used interchangeably throughout this SOP and TI documents. Separately, null codes and null flags may also be used interchangeably.

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 Drew Ave facility is secured with a security alarm that notifies UCD police when activated.

Databases, file servers, and web server virtual and dedicated machines operate primarily out of the Metro IT data center in Hoagland Hall on the UCD campus. Metro IT has a highly-available, disaster recoverable virtualization environment. Weekly backups of the virtual hard drives are taken offsite and stored in the Campus Data Center. In the event of a disaster in Hoagland, critical machines will be mounted at the Campus Data Center. The Drew Avenue laboratory is directly connected to the main campus internet. In the event that connection is disrupted (such as through a construction accident) and the outage would be greater than a day, the physical servers would be relocated to the Drew Ave. location.

### 9.1.2 Hardware Recovery Plan

The campus network of IT Administrator staff allows for rapid response to server failure and recovery issues.



### 9.1.3 Software and Data Recovery Plan

#### 9.1.3.1 UCD EDXRF, HIPS, and Quartz Carbon Laboratories

Raw and processed spectra from the EDXRF analyses are saved and available for use at any time on the PANalytical Epsilon 5 computers in the EDXRF Laboratory.

Thermograms from the TOA analyses are saved and available for use at any time on Sunset analyzer computers in the Quartz Carbon Laboratory.

Reflectance and Transmittance from the HIPS analyses are saved and available for use at any time in the CSN Database.

Data safety and security are ensured by frequent transfer of computerized raw data from the UCD EDXRF Laboratory, UCD HIPS Laboratory, and UCD Quartz Carbon Laboratory (all located in the AQRC building on Drew Ave) computers to two servers located in different buildings on campus. Differential backups are performed daily and full backups are performed weekly.

#### 9.1.4 Data Security

**UCD and RTI access policies:** Access to databases 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 UCD campus system administrator. Password integrity is monitored by the UCD 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.