UCD IMPROVE Technical Information #351A

Data Ingest

Interagency Monitoring of Protected Visual Environments
Air Quality Research Center
University of California, Davis

December 17, 2024
Version 1.2

Prepared By: Signed by: Indu Thekkemeppilly Sivakumar 19F6B63B1B17443... Date: 12/18/2024

Reviewed By: DocuSigned by: Alex Murrain 2B7AC07297404D6... Date: 12/18/2024

Approved By: DocuSigned by: Marcus Langston 0A10CFCF79B0452... Date: 12/18/2024

UCDAVIS
AIR QUALITY RESEARCH CENTER

DOCUMENT HISTORY

Revision	Release Date	Initials	Section/s Modified	Brief Description of Modifications
1.0	10/03/2022	SRS	All	Previously anthologized version separated into individual TIs.
1.1	03/23/2023	DEY	9.5	Added section on analytical MDLs.
1.2	12/17/2024	ITS	3, 9	Updated definitions and information for Analytical MDL updates.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. PURPOSE AND APPLICABILITY

The purpose of this technical information (TI) is to provide information on handling electronic laboratory records from samples collected in the Interagency Monitoring of Protected Visual Environments (IMPROVE) network. This document is intended to guide users on the receiving and validating of IMPROVE laboratory records and ingestion to the University of California, Davis (UCD), IMPROVE database. These include ion analysis results from RTI International (RTI), carbon analysis results from Desert Research Institute (DRI), and gravimetric mass, elemental, and filter absorption analysis results from UCD.

2. SUMMARY OF THE METHOD

Ion analysis results from RTI and carbon analysis results from DRI are received in data files, typically delivered as .csv files and XML files, respectively. The files are ingested into the UCD IMPROVE database using the UCD IMPROVE Data Management website. Gravimetric mass, elemental, and filter absorption analysis results from UCD are automatically ingested.

3. DEFINITIONS

- **AQRC:** Air Quality Research Center.
- **CSV:** a comma-separated value file that is the common format for delivery files.
- **DRI:** Desert Research Institute.
- **Energy Dispersive X-Ray Fluorescence (EDXRF):** An analytical technique used to determine the concentration of elements.
- **Hybrid Integrating Plate/Sphere (HIPS):** An analytical technique for optical absorption.
- **Ion Chromatography (IC):** An analytical technique used to determine the concentration of ions.
- **Interagency Monitoring of Protected Visual Environments (IMPROVE):** Federal $PM_{2.5}$ and PM_{10} sampling network directed by the National Park Service, with sites located principally in remote rural areas.
- **IMPROVE database:** A SQL Server database that is the central warehouse of IMPROVE preliminary and final data at UCD.
- **Method Detection Limit (MDL):** A lower limit of detection specific to method of analysis and reported parameter.
- **NPS:** National Park Service.
- **PM:** Particulate Matter. $PM_{2.5}$ is particulate matter with diameters 2.5 micrometers (µm) and smaller. PM_{10} is particulate matter with diameters 10 µm or smaller.
- **RTI:** Research Triangle Institute, International.
- **SOP:** Standard Operating Procedure.

- **SQL:** database management system used by AQRC.
- **Thermal Optical Analysis (TOA):** An analytical technique used to determine the concentration of carbon. Also referred to as TOR (Thermal Optical Reflectance) and TOT (Thermal Optical Transmittance).
- **UCD:** University of CA—Davis.
- **Extensible Markup Language (XML):** a markup language defining a set of rules for encoding documents in a particular format; used for IMPROVE carbon files.
- **Sample Handling laboratory(SHL):** The filter handling laboratory for IMPROVE at UCD.
- **Field Group:** The group in charge of samplers and operation of IMPROVE network
- **UCD IMPROVE Data Management Website:** A web-based application developed with .NET framework to interact with the IMPROVE database. Also referred to as IMPROVE web app.

4. HEALTH AND SAFETY WARNINGS

Not applicable.

5. CAUTIONS

Not applicable.

6. INTERFERENCES

Not applicable.

7. PERSONNEL QUALIFICATIONS

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. EQUIPMENT AND SUPPLIES

The hardware and software used for IMPROVE data ingest are described in the associated *UCD IMPROVE SOP #351: Data Processing & Validation.*

9. PROCEDURAL STEPS

Prior to data processing and validation, data are ingested for each of the analysis pathways: (1) carbon results from DRI, (2) ions results from RTI, and (3) elemental and optical absorption results from UCD.

9.1 Carbon Results

Carbon analysis results are sent from DRI to UCD via email in .xml format, including three files:

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

All three files are included in a zip folder which will be saved to U:\IMPROVE\RawDataReceived\Carbon DRI\To be Imported. The files will be extracted by right clicking on zipped folder, selecting 7-Zip and then 'Extract Here'. A new folder with the same name as the zipped folder will be created in the same location on the U Drive and contain all three files as named above.

All three files are ingested using the UCD IMPROVE Data Management Website. Figure 1 shows a screenshot of the carbon data upload page, which is accessed via the Analysis Data Section(https://improve.aqrc.ucdavis.edu/) as described in section 8 of this document, selecting the **Carbons** tab, and clicking the **Ingest Data** button. To ingest the files from the data upload page, select the relevant files, create a name for the import batch under *Batch Label*, and click **Submit**. The suggested batch label is the filename from DRI (First sample date - Last Sample date). *CarbonInformation*, *CarbonLaser*, and *CarbonData* are ingested simultaneously, and an automated validity check is performed (Table 1). Results from the validity check will indicate upload failures. The Data Validator will review the upload results and notify the Data & Reporting Group Supervisor if there are upload failures from validation errors. After ingesting, the source files are stored on the file server at U:\IMPROVE\RawDataReceived\Carbon DRI\Imported, within a folder which is named the same as the *Batch Label* created for ingest. After successfully ingesting the results, the Data Validator will save a copy of the Carbon Data Ingestion summary page by printing the page to a PDF and saving to U:\IMPROVE\RawDataReceived\Carbon DRI\IngestRecord\. Further details of the ingest file are recorded in a log file located at U:\IMPROVE\RawDataReceived\Carbon DRI\Carbon_Ingest_log.xlsx.

Figure 1. Carbon analysis results upload page.

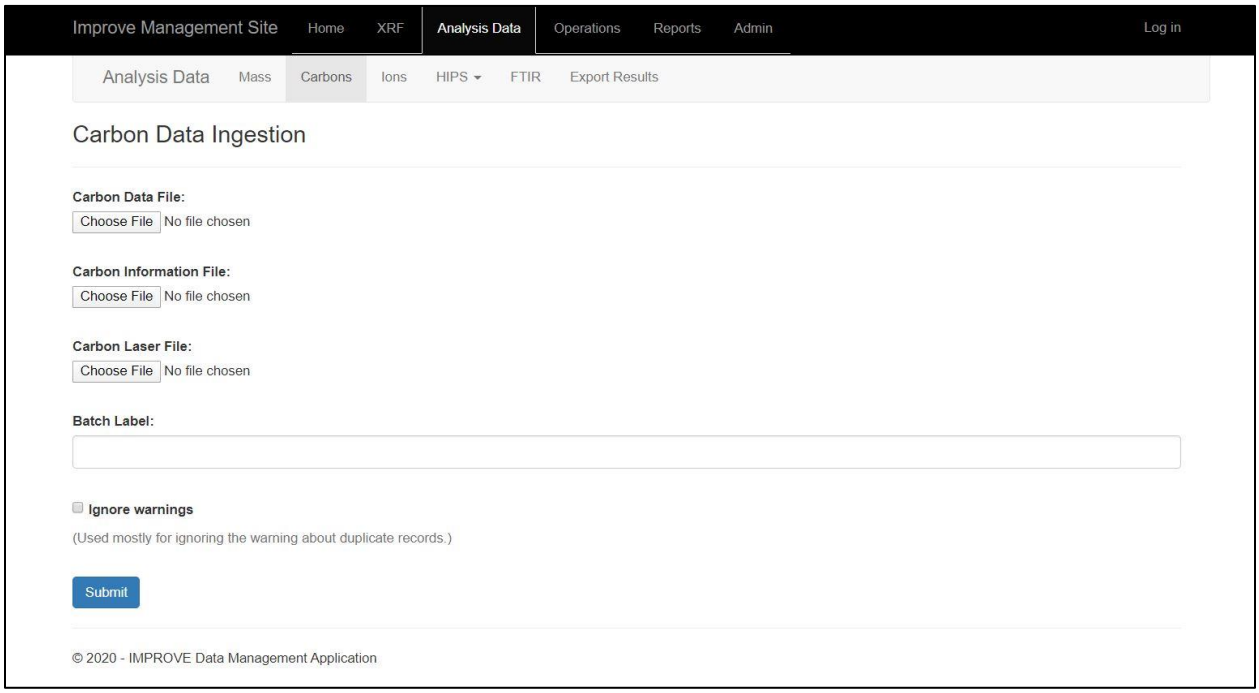


Table 1. Automated validity checks performed during carbon data upload.

Check	**Action**
Basic schema validation on xml files	Error
No filter found for record	Warning
Filter.Module doesn't match record Site field	Warning
Record is marked as re-analysis	Warning
Carbon Laser file has records missing wavelength	Warning
Found more parameter records than expected for an analysis	Warning
Parameter missing for an analysis	Warning
Comment from DRI on analysis	Note
Parameter/record already recorded in database	Warning
Incomplete analysis record (missing entries in either Carbon/Carbon Laser/Carbon Info file)	Warning

9.2 Ion Results

Ions analysis results are sent as one file from RTI to UCD via email in .csv format. The naming convention of the ions data includes the year followed by the ions data set

number (e.g. '2020 1 2 3 data export to UCD'). The file is saved to the file server at U:\IMPROVE\RawDataReceived\Ions RTI\To Be Ingested.

The ion analysis records are ingested using the UCD IMPROVE Data Management website(https://improve.aqrc.ucdavis.edu/). Figure 2 shows a screenshot of the ions data upload page, accessed via the Analysis Data Section as described in section 8 of this document, selecting the Ions tab and clicking the **Upload Data** button. To ingest a file from the data upload page, select the relevant file and click **Continue**. An automated validity check is performed, and the validity check results will indicate if there are upload failures (Table 2). The Data Validator will review the upload results and notify the Data & Reporting Group Supervisor if there are upload failures from validation errors. An ingest page after passing validation is shown in Figure 3. If the file is ready for ingest, click **Submit**. After ingesting, the source files are stored on the file server at U:\IMPROVE\RawDataReceived\Ions RTI\Ingested. After successfully ingesting the results, the Data Validator will save a copy of the Ions Data Ingestion Status summary page by printing the page to a PDF and saving to U:\IMPROVE\RawDataReceived\Ions RTI\Ingest_record\. Further details of the ingest file are recorded in a log file located at U:\IMPROVE\RawDataReceived\Ions RTI\Ions_DataIngest_Log.xlsx.

Figure 2. Ions analysis results upload page.

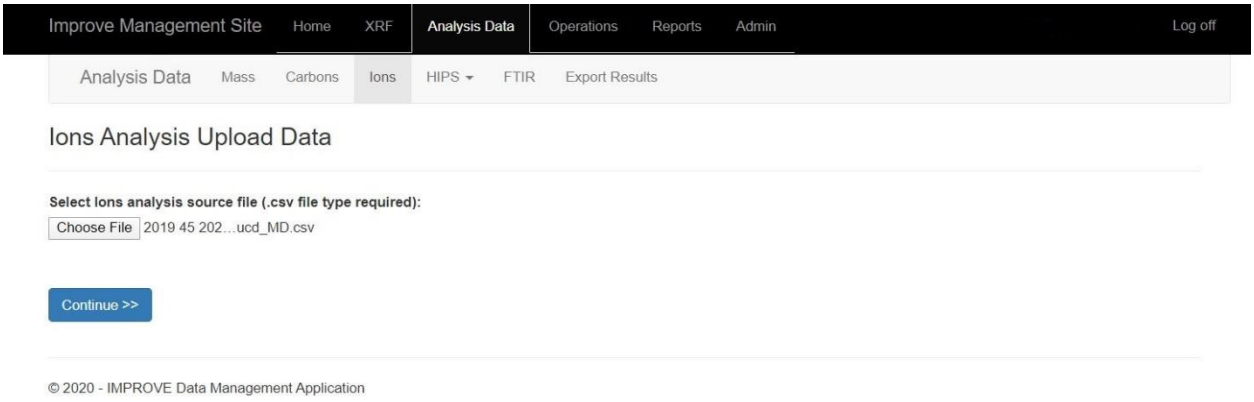
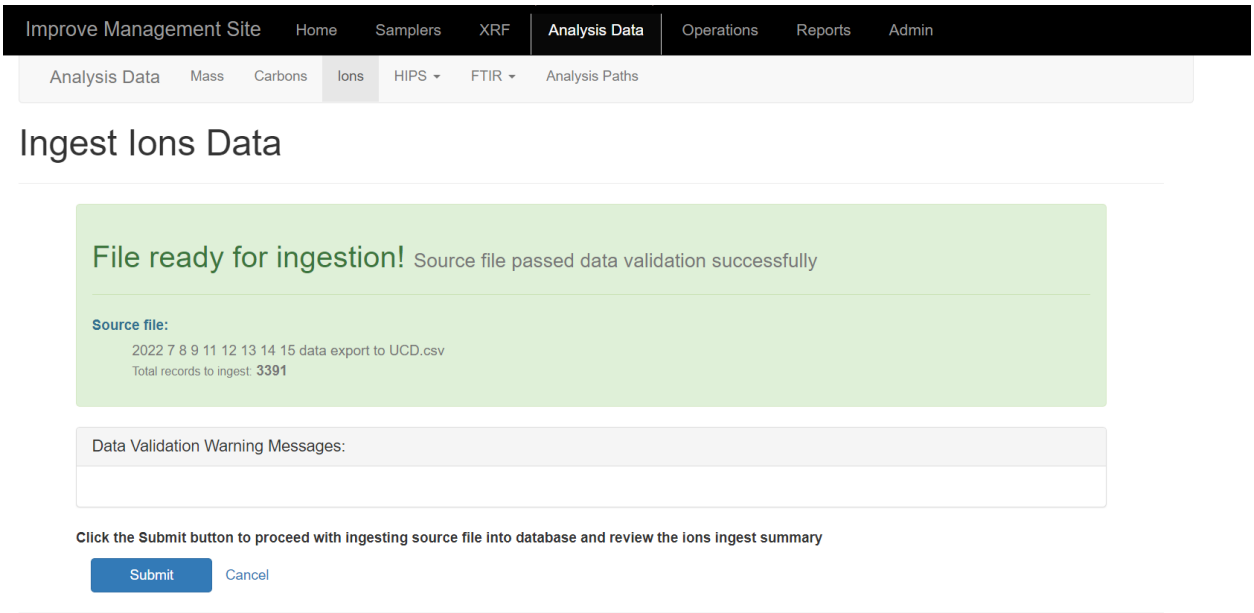


Table 2. Automated validity checks performed during the ions data upload.

Check	**Action**
Basic schema validation on csv files	Error message
No filter is found for record	Error message
Data already exists for filter record	Warning message
Parameter missing for a filter e.g., parameter has null value or parameter column isn't present at all	The ingest process will abort with an error message
Parameter already recorded in database	Warning

Figure 3. Ions ingest page after passing validation.



9.3 Element and Optical Absorption Results

Elemental analysis is performed at UCD. The PANalytical XRF software generates results files and automatically transmits them to a directory on an AQMT file server. A service on the server (internally named *XRF Data Transfer*) monitors the transmission directory, checking every five minutes for new files. The XRF results files are standard text files with the extension *.qan*. The file name includes XRF analysis dates and times in the format *YYYYMMDDHHMMSS.qan*. The results files and contents are automatically parsed and ingested into tables in the UCD IMPROVE database.

Optical absorption analysis is performed at UCD. The HIPS instrument generates results which are then verified by the operator to be complete and then written to the database. The data are then available on the UCD IMPROVE database.

9.4 Re-ingesting

If errors are identified in the source files from DRI or RTI that cause the import to fail, or if results are updated as part of the validation and reanalysis process, new files must be requested and provided for ingestion. Upload the new files using the process described in sections 9.1 and 9.2.

For carbon, whether the files contain new batches of data or reanalysis results, take care to ingest with the *ignore warnings* box unchecked. Scrutinize the messages and warnings to check for errors and take note of further actions that may be required after the data is

ingested (e.g., changing analysis QC codes). The import process indicates if there are matching existing records, if existing records are not updated, or if only new records are added (including cases with different analysis results from the sample filter). Once the messages have been reviewed and addressed, re-run the ingest process with the *ignore warnings* box checked. For carbon, if the reanalysis results are used, the analysis QC code can be updated using the tool available at https://improve.aqrc.ucdavis.edu/AnalysisData/Carbons/CarbonsQcReview.

For ions, the data are ingested without any changes to the original process; the QC code is updated using the tool available at https://improve.aqrc.ucdavis.edu/AnalysisData/Ions/IonsQcReview.

9.5 Analytical MDLs

Prior to a new sampling year of data being processed into concentrations, the Data Validator is to contact the laboratories to determine if any changes should be made to the analytical MDLs to be applied beginning with the next year of filters. The laboratories should be asked approximately two months prior to the anticipated processing date. Note that the laboratories may already have started analyzing the filters from the next sampling year so the analyst should be clear in the request emails about all the different dates being mentioned.

The existing analytical MDLs are to be provided to the laboratories as part of the request. To do this, download the current analytical MDLs from the database using an SQL query like the following:

```
SELECT Parameter, AnalyticalMDL, Unit, EffectiveDate
  FROM [Improve_2.1].[analysis].[AnalyticalMDL] aMDL
  LEFT JOIN [Improve_2.1].[analysis].[AnalyticalMDLSets] aMDLsets on
aMDL.AnalyticalMDLSetId = aMDLsets.Id
  WHERE AnalysisQcCode = 1
  AND EffectiveDate >= '2022-01-01'
```

where the effective date is the previous year of data, which is typically the most recent set. Save the data into an Excel spreadsheet and make it clear that these data are currently being applied to samples. Add in columns for the laboratories to fill in for the next year of analytical MDLs. Split the spreadsheet into the different analyses and send to the appropriate laboratories to ask for any updates that should be applied to the next year of samples.

Save the returned completed spreadsheets in a sub-folder, with the name of the year of data the new analytical MDLs are to be applied to (e.g. ‘For_2023’), within the main folder on the U drive at U:\CSN\Collected Files\Analytical_MDLs\. Once all of the analytical MDLs have been received, compile the data into a single spreadsheet, with the fields ‘Parameter’, ‘MDL’, and ‘Unit’, and save.

For how the analytical MDLs are used during processing, the unit for elements is µg/cm^2, the unit for ions, carbon, and mass is µg/filter, and tau 633 (for HIPS) is unitless.

Before adding the analytical MDL (Method Detection Limit) to the database, it is crucial to conduct a thorough comparison of the current data with the data from previous years. The SQL query to download the current analytical MDLs mentioned above can be used for this purpose. This comparison aims to discern any substantial and abrupt fluctuations in MDLs. If this trend is observed, it would be advisable to contact the laboratories to ensure that the change aligns with analytical expectations and standards.

To insert the data into the database, the analyst should open an R project and load the four functions listed below, which can be found in the R script "Insert_IMPROVE_analytical_MDLs_db.R" located on the U drive at U:\IMPROVE\Collected_Files\Analytical_MDLs\.

- imp_analytMDL
- imp_analytMDL_get
- imp_insert_MDL
- imp_insert_MDL_set

The analyst should execute the following command in R:

imp_analytMDL(MDLdata = [path to Excel spreadsheet], effectiveDate = ['YYYY-MM-DD'], server = ['production'], AnalysisQcCode = 1, comment = ["Appropriate comment"], replacingId = NULL, replacingQcCode = NULL)

This command appends the provided analytical MDLs data to the `[Improve_2.1].[analysis].[AnalyticalMDL]` table in the UCD CSN production database as an analytical MDL set. It also inserts a records into the `[Improve_2.1].[analysis].[AnalyticalMDLSets]` table that provides summary information for this set, including the comment, the *effectiveDate*, and *AnalysisQcCode*. It may be necessary to modify a whole set of analytical MDLs in which case the analyst should prepare the revised Excel spreadsheet and run the *imp_analytMDL* command again with the relevant information updated, including changing the comment to record what the dataset relates to as well as change the analysis QC code of the previously posted set(s) by including the set ID of the previous posting (*replacingId*) and the analysis QC code (*replacingQcCode*) that should be associated with that data set.

An example of this command that was used to add the analytical MDLs to the database to be used with samples beginning January 2022 is as follows:

imp_analytMDL("U:/IMPROVE/Collected_Files/Analytical_MDLs/For_2022/MDLs_for_2022.xlsx", effectiveDate = '2022-01-01', server = 'production', comment = "Analytical MDLs updated in 2022, effective beginning with January 2022 samples")

The analyst should save the command and any associated notes in an R script and save in the relevant sub-folder on the U drive.

9.6 Issue Tracking

Software bugs and data management issues are tracked through JIRA tracking software. All users who have access to the internal UCD JIRA website can submit, track, and comment on issues. Users requesting new tools, modifications to existing tools, or to report bugs specific to the IMPROVE data should add JIRA tickets to the IMPROVE Data Management Software project at https://improve.atlassian.net/jira/software/c/projects/IMPSW/issues/

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 IMPROVE database via unique primary keys and non-nullable records. Data completeness and data quality are thoroughly checked through the data validation process, as described elsewhere in this SOP.

11. REFERENCES

Not applicable.