Data and Metadata Reporting Standards for the U.S. Environmental Protection Agency's PM Supersites Research Program[1]

Les A. Hook and Sigurd W. Christensen

NARSTO Quality Systems Science Center

Oak Ridge National Laboratory[2]

William B. Sukloff

Environment Canada

Meteorological Service of Canada

[1]Office of Air Quality Planning and Standards, U.S. Environmental Protection Agency. Michael Jones, Project Officer.

[2]ORNL research was sponsored by U.S Environmental Protection Agency and U.S. Department of Energy and performed at Oak Ridge National Laboratory (ORNL). ORNL is managed by UT-Battelle, LLC, for the U.S. Department of Energy under contract DE-AC05-00OR22725.

U.S. EPA Supersites Research Program

The Environmental Protection Agency is sponsoring a major atmospheric particulate matter (PM) data collection effort in seven major U.S. cities.

Objectives:

1) Characterize PM and its constituents,
2) Collect data and samples to support health and exposure studies to reduce uncertainty in setting National Ambient Air Quality Standards, and
3) Compare emerging sampling and analysis methods with routine techniques to enable a smooth transition to advanced methods.

Data Management Considerations:

- Analyze individual site PM and atmospheric conditions,
- Integrate data from all the Supersites for cross-site analyses, and
- Archive data in a timely manner so they will be readily available to the public.

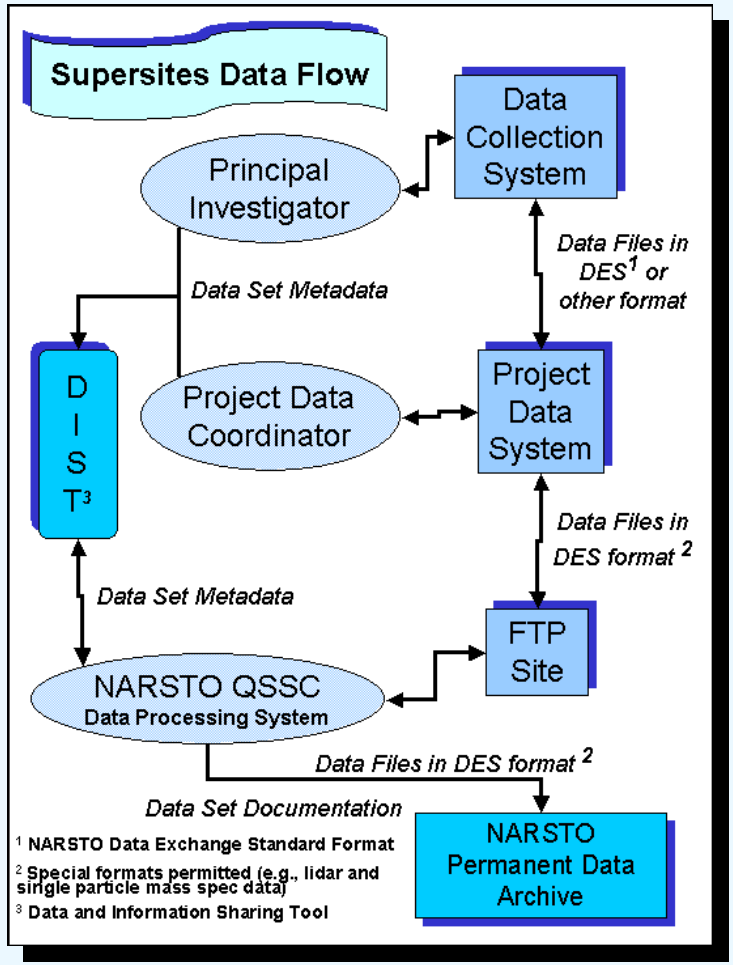
Right from the start……..

Data management was addressed in the Cooperative Agreements that implement the Program.

- Data Coordinators at each Supersite
- Designated data archive --The NARSTO Permanent Data Archive (PDA) at the Langley NASA DAAC
- PDA requires Data Exchange Standard (DES) file format with several metadata requirements
- Archiving process is mediated by the NARSTO Quality Systems Science Center (QSSC), Oak Ridge National Laboratory

Reality……

- Process of determining standards was going on concurrently with data collection
- Integration of data in future analyses demands consistently defined metadata elements and values.
- Benefit of standardization will be realized by those Supersites still planning to gather data and by the those analyzing integrated data.



NARSTO encourages scientists to document their data at a level sufficient to satisfy the well-known "20-year test". That is, someone 20 years from now, not familiar with the data or how they were obtained, should be able to find data of interest and then fully understand and use the data solely with the aid of the documentation archived with the data. National Research Council, Committee on Geophysical Data, *Solving the Global Change Puzzle, A U.S. Strategy for Managing Data and Information,* National Academy Press, Washington, D.C., 1991.

Standards Development Process

Formed a Data Management Working Group (DMWG) with the Data Management Coordinators for each Supersite. QSSC as the lead.

- The Working Group communicated through weekly teleconferences to deal with consistency of metadata content and data reporting format.
- Minutes of the teleconferences discussion and decisions were distributed to the DMWG, Site Principal Investigators, QA staff, and EPA program management.
- Developed a statement of degree of metadata standardization needed, by category (e.g., site names, parameters, units, methods, flags, etc), and principles of standardization.
- We incorporated metadata elements and values from other metadata standards when available to promote consistency within EPA and the atmospheric research community, and to anticipate integrating data from additional sources.
- For example, in addition to existing NARSTO standards, we used site descriptors and event flags from EPA AIRS, detailed flags from EPA Region 5, CAS Registry Numbers and CAS Index Names for chemical identification from EPA CRS, and the non-chemical variable naming syntax from the DOE ARM Program.
- Each metadata standard was completed within the DMWG and then sent to the Site Principal Investigators for approval, after which they were considered Consensus Metadata Standards. The DMWG updates these as needed.

Consensus Metadata Standards

- We worked for as much standardization as possible in the key areas.
- Site-specific implementation flexibility is always a consideration and was maintained when possible.
- While we worked for standardization, complete and immediate standardization in all relevant areas is not possible or desirable. Cannot anticipate all needs “up front.”
- Reference tables of metadata values are maintained for consistency.
- Configuration Control -- Changes and additions to the reference tables are controlled. A Data Coordinator recommends additions or changes to the DMWG and QSSC for discussion and consensus. Subject matter experts are consulted when appropriate.
- A Read and Verify Program checks metadata consistency in archive data files against the various reference tables, for example, that CAS numbers and variable names are in the reference tables.

Site Identification:

Identifies a 12-character standard syntax for naming fixed and mobile sites.

STUDY OR NETWORK ACRONYM + COUNTRY CODE + STATE OR PROVINCE CODE + SITE ABBREVIATION (MNEMONIC)

- A master list of site names from the Supersites program will be assembled by the QSSC.
- The master list will include additional information about each site, as available: latitude, longitude, elevation, EPA AIRS identifier, land use, and location type.

Identifying Chemical and Physical Variables and Descriptive Metadata:

Identifying Chemical Substances with a CAS Registry Number:

- Valid values are the CAS Number (with "C" prefix – to prevent spreadsheet date conversion).
- Chemical name may be either CAS-9CI, IUPAC, or other common name.
- Reference table of CAS numbers and preferred names is maintained.

CHEMICALS WITH CAS REGISTRY NUMBERS

Archive_CAS	Name_9CI	Name_IUPAC
C120-12-7	Anthracene	Anthracene
C129-00-0	Pyrene	Pyrene
C135-48-8	Pentacene	Pentacene
C189-55-9	Benzo[rst]pentaphene	Benzo[rst]pentaphene
C189-64-0	Dibenzo[a,h]pyrene	Dibenzo[b,def]chrysene
C191-07-1	Coronene	Coronene
C191-24-2	Benzo[ghi]perylene	Benzo[ghi]perylene
C191-26-4		Dibenzo[def,mno]chrysene
C191-30-0	Dibenzo[a,l]pyrene	Dibenzo[def,p]chrysene
C192-65-4	Dibenzo[a,e]pyrene	Naphtho[1,2,3,4-def]chrysene
C192-97-2	Benzo[e]pyrene	Benzo[e]pyrene
C193-39-5	Indeno[1,2,3-cd]pyrene	Indeno[1,2,3-cd]pyrene
C198-55-0	Perylene	Perylene
C203-12-3		Benzo[ghi]fluoranthene
C203-33-8	Benzo[a]fluoranthene	Benz[a]aceanthrylene
C203-64-5	4H-Cyclopenta[def]phenanthrene	4H-Cyclopenta[def]phenanthrene
C205-82-3	Benzo[j]fluoranthene	Benzo[j]fluoranthene
C205-99-2	Benz[e]acephenanthrylene	Benz[e]acephenanthrylene
C206-44-0	Fluoranthene	Fluoranthene
C207-08-9	Benzo[k]fluoranthene	Benzo[k]fluoranthene
C208-96-8	Acenaphthylene	Acenaphthylene
C214-17-5		Benzo[b]chrysene
C215-58-7	Benzo[b]triphenylene	Benzo[b]triphenylene
C217-59-4	Triphenylene	Triphenylene
C218-01-9	Chrysene	Chrysene

Identifying Chemical Substances, Calculated Quantities, and Physical/Non-chemical Measurements that do not have a designated CAS Registry Number:

- Variable names are formed beginning with the root concept, and followed by a detailed modifier if needed, separated by a ":".

- For example,

  > Carbon: elemental (EC1),
  > PM10: mass,
  > Temperature: air,
  > Temperature: dew point.

- These variables can be method specific and may require special differentiation.

VARIABLE NAMES FOR CHEMICALS WITHOUT CAS REGISTRY NUMBERS

*TABLE COLUMN NAME	*TABLE COLUMN BASIS	*TABLE COLUMN USER NOTE	*TABLE COLUMN LABORATORY ANALYTICAL METHOD
Carbon: elemental (EC1)	550C He/O2	IMPROVE - EC1	Thermooptical reflection
Carbon: elemental (EC2)	550-700C He/O2	IMPROVE - EC2	Thermooptical reflection
Carbon: elemental (EC3)	700-800C He/O2	IMPROVE - EC3	Thermooptical reflection
Carbon: organic (OC1)	ambient-120C He	IMPROVE - OC1	Thermooptical reflection
Carbon: organic (OC2)	120-250C He	IMPROVE - OC2	Thermooptical reflection
Carbon: organic (OC3)	250-450C He	IMPROVE - OC3	Thermooptical reflection
Carbon: organic (OC4)	450-550C He	IMPROVE - OC4	Thermooptical reflection
Carbon: organic (OP)	550C He/O2	IMPROVE - OP	Thermooptical reflection
Carbon: total	organic + elemental	IMPROVE	Thermooptical reflection
Carbon: total elemental	EC1+EC2+EC 3-OP	IMPROVE	Thermooptical reflection

Identifying Calculated Quantities, and Physical / Non-chemical Measurements:

- For example, Temperature: air, Temperature: dew point.
- Definition of new variables is relatively straightforward when the format is known.
- Related variables sort together in a list

Identifying Metadata Elements:

- Valid variable names for metadata elements including site information, locations, dates, times, and sampling conditions are provided in a reference table.
- The correct format is the root concept, followed by a detailed modifier if needed, separated by a ":".
- For example, Date start: local time, Time start: local time, and Latitude: decimal degrees, Longitude: decimal degrees.

VARIABLE NAMES FOR NON-CHEMICAL VARIABLES

*TABLE COLUMN NAME	*TABLE COLUMN UNITS	*TABLE COLUMN BASIS
Wind direction: resultant mean vector	decimal degrees	Referenced to true North.
Wind direction: mean scalar	decimal degrees	Referenced to true North.
Wind direction: Sigma theta (std.dev. of azimuth angle)	decimal degrees	
Wind direction: unit vector average	decimal degrees	Referenced to true North.
Wind gust	m/s	
Wind gust: maximum	m/s	
Wind speed: scalar mean	m/s	
Wind speed: horizontal scalar minimum	m/s	
Wind speed: horizontal scalar maximum	m/s	
Wind speed: resultant mean vector	m/s	
Wind speed: Sigma phi (horizontal standard deviation)	m/s	
Wind speed: u (W-to-E component)	m/s	
Wind speed: v (S-to-N component)	m/s	
Wind speed: w (mean vertical component:+ up/- down)	m/s	
Wind speed: Sigma w (standard deviation of vertical winds)	m/s	

Data Quality Flags:

- Values must be assigned at least one data quality flag by the data originator.
  - Valid without qualification,
  - Valid but qualified / suspect, or
  - Invalid due to serious sampling or analysis problems. (Missing)
- These flags may be the NARSTO flags or more detailed flags as defined by a Project. Project-defined flags must be mapped to NARSTO flags.
- Reference tables of NARSTO standard flags, detailed project flags, and EPA AIRS exceptional-event flags are provided for users.

NARSTO DATA QUALITY FLAGS

NARSTO Flag Code	NARSTO Flag Definition
V0	Valid value
V1	Valid value but comprised wholly or partially of below detection limit data
V2	Valid estimated value
V3	Valid interpolated value
V4	Valid value despite failing to meet some QC or statistical criteria
V5	Valid value but qualified because of possible contamination (e.g., pollution source, laboratory contamination source)
V6	Valid value but qualified due to non-standard sampling conditions (e.g., instrument malfunction, sample handling)
V7	Valid value but set equal to the detection limit (DL) because the measured value was below the DL
M1	Missing value because no value is available
M2	Missing value because invalidated by data originator
H1	Historical data that have not been assessed or validated

Data Exchange Standard Development

Background:

- Data files submitted for archiving are in the NARSTO Data Exchange Standard (DES) format.
- DES implements the consensus metadata standards for site names, data quality flags, and variable names, plus key characteristics (see below).
- DES format follows a spreadsheet-compatible layout and is stored as ASCII comma-separated value (.csv) files. People can easily view and understand the archived files.
- An Excel/97® template for inputting data and metadata has been developed. The template is annotated with comments, instructions, frequently asked questions, and examples of completed DES files. Within the template are picklists for selecting values for various metadata fields to ease metadata entry and to promote consistency in terminology.

Supersite Enhancements to the Data Exchange Standard:

- DES was used primarily with gaseous atmospheric constituents and meteorological measurements.
- PM measurements are not so easily characterized. Results are often operationally defined by the specific field sampling configurations, measurement devices, conditions, and laboratory sample preparation and analysis methods.
- A set of **key characteristics** was defined to capture enough of the measurement information to be meaningful and helpful in a data file, while avoiding excessive detail.

Key Characteristics provide general sampling and analysis information that describes the data.

Key characteristics, metadata values, and organization of the DES were defined through invaluable interactions of Data Coordinators with PIs and with other field and laboratory technical experts

Key Characteristics

> OBSERVATION TYPE U
> SAMPLING HEIGHT (M AGL) U
> FIELD SAMPLING OR MEASUREMENT PRINCIPLE U
> INLET TYPE U
> MEDIUM U
> COATING OR ABSORBING SOLUTION/MEDIA U
> SAMPLING HUMIDITY OR TEMPERATURE CONTROL U

> PARTICLE DIAMETER--LOWER BOUND (UM)
> PARTICLE DIAMETER--UPPER BOUND (UM)
> PARTICLE DIAMETER--MEDIAN (UM)

> WAVELENGTH (NM)
> WAVELENGTH--LOWER BOUND (NM)
> WAVELENGTH--UPPER BOUND (NM)

> SAMPLE PREPARATION U
> LABORATORY ANALYTICAL METHOD U

> VOLUME STANDARDIZATION U
> BLANK CORRECTION U

> INSTRUMENT NAME AND MODEL NUMBER (optional)
> MEASUREMENT PRINCIPAL INVESTIGATOR (optional)

Supplemental Information

> EXPLANATION OF ZERO OR NEGATIVE VALUES

> EXPLANATION OF REPORTED DETECTION LIMIT VALUES
> DETECTION LIMIT VALUES

> EXPLANATION OF REPORTED UNCERTAINTY VALUES
> UNCERTAINTY VALUES

U U U **Picklists for consistently selecting Key Characteristic values are included in the DES template**.

More detailed information about field, instrument, and laboratory conditions would be included in companion files, such as the Quality Assurance Project Plans.

Data Exchange Standard

- The DES does not rely on row position to identify metadata information, but uses tags to describe the information contained in the row.
- The DES is a self-documenting format with three sections:
  - Header Section contains information about the contents of the file and the data originator;
  - Middle Section contains metadata tables that describe/define sites, flags, and other codified fields; and
  - Final Section is the main data table that contains key characteristics and data values.

Header Section

*DATA EXCHANGE STANDARD VERSION	NARSTO 2001/10/31 (2.213)
*COMMENT	Detailed instructions on how to fill in this template are provided at:
*COMMENT	http://cdiac.esd.ornl.gov/cdiac/programs/NARSTO/narsto.html#qsmp
*COMMENT	(Click on the link that says View the NARSTO Data Management Handbook.)
*QUALITY CONTROL LEVEL	
*DATE THIS FILE GENERATED/ARCHIVE VERSION NUMBER	2002/03/27 1
*ORGANIZATION ACRONYM	
*ORGANIZATION NAME	
*STUDY OR NETWORK ACRONYM	
*STUDY OR NETWORK NAME	
*FILE CONTENTS DESCRIPTION--SHORT/LONG	
*PRINCIPAL INVESTIGATOR NAME--LAST/FIRST	
*PRINCIPAL INVESTIGATOR AFFILIATION	
*CO-INVESTIGATOR NAME--LAST/FIRST	
*CO-INVESTIGATOR AFFILIATION	
*COUNTRY CODE	
*STATE OR PROVINCE CODE	
*SAMPLING INTERVAL AS REPORTED IN MAIN TABLE	
*SAMPLING FREQUENCY OF DATA IN MAIN TABLE	
*PRINCIPAL INVESTIGATOR CONTACT INFORMATION	
*DATA USAGE ACKNOWLEDGEMENT	
*NAME AND AFFILIATION OF PERSON WHO GENERATED THIS FILE	
*DATE OF LAST MODIFICATION TO DATA IN MAIN TABLE	
*NAME AND VERSION OF SOFTWARE USED TO CREATE THIS FILE	
*STANDARD CHARACTERS	!#$%&'()*,-./0123456789:;<>?@ABCDEFGHIJKLMNOPQRSTUVWXYZ\^_`abcdefghijklmnopqrstuvwxyz{}~
*COMPANION FILE NAME/FORMAT AND VERSION	

Please check with the QAC for the standard acronym used by each study.
Value restrictions:
1) <= 15 characters in length.
2) alphanumeric characters, plus, minus,dash or underscore only.
This is a **mandatory** key phrase.
e.g.
SCOS97,
NARSTO-CE96

The **length of time represented by one row of data** in the main data table under nominal sampling conditions. This may be different from the interval recorded by the measurement instrument (e.g., if **1 hour averages** are derived from **1 minute measurements**, the "**sampling interval as reported**... is **1 hour**.) Select from the **drop-down list**.
This is a **mandatory** key phrase.

Middle Section contains metadata tables that describe/define sites, flags, and other codified fields.

*TABLE NAME	Site information				
*TABLE FOCUS	Metadata				
*TABLE COLUMN NAME	Site ID: standard	Site abbreviation: stand	Description	Country code	State or province code
*TABLE COLUMN UNITS	None	None	None	None	None
*TABLE COLUMN FORMAT TYPE	Char	Char	Char	Char	Char
*TABLE COLUMN FORMAT FOR DISPLAY	12	4	50	50	20
*TABLE COLUMN MISSING CODE	None	None	None	None	None
*TABLE BEGINS					
*TABLE ENDS					

*TABLE NAME	NARSTO standard flags				
*TABLE FOCUS	Metadata				
*TABLE COLUMN NAME	Flag: NARSTO	Description			
*TABLE COLUMN UNITS	None	None			
*TABLE COLUMN FORMAT TYPE	Char	Char			
*TABLE COLUMN FORMAT FOR DISPLAY	2	120			
*TABLE BEGINS					
	V0	Valid value			
	V1	Valid value but comprised wholly or partially of below detection limit data			
	V2	Valid estimated value			
	V3	Valid interpolated value			
	V4	Valid value despite failing to meet some QC or statistical criteria			
	V5	Valid value but qualified because of possible contamination (e.g., pollution source, labo			
	V6	Valid value but qualified due to non-standard sampling conditions (e.g., instrument ma			
	V7	Valid value but set equal to the detection limit (DL) because the measured value was b			
	M1	Missing value because no value is available			
	M2	Missing value because invalidated by data originator			
	H1	Historical data that have not been assessed or validated			
*TABLE ENDS					

Final Section is the main data table that contains key characteristics and data values.

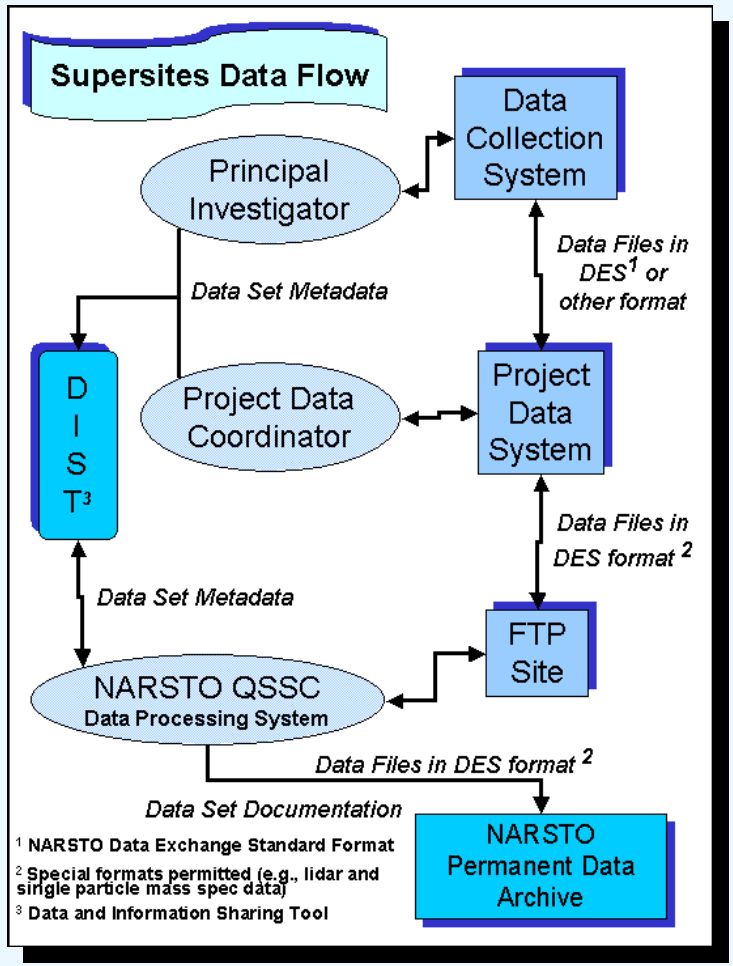
*TABLE ENDS				
*TABLE NAME	TVA PC-BOSS			
*TABLE USER NOTE	Particle Concentrator Speciation Sampler BYU Design-Ions and OC/EC data			
*TABLE FOCUS	Surface--fixed			
*TABLE EXPLANATION OF ZERO OR NEGATIVE VALUES	-9999. Indicates missing value			
*TABLE EXPLANATION OF REPORTED DETECTION LIMIT VALUES	The detection limit is 0.1 microns per cubic meter			
*TABLE KEY FIELD NAMES	Site ID: standard			
*TABLE COLUMN NAME	Site ID: standard	PM2.5: mass	PM2.5: mass	Sulfate
*TABLE COLUMN NAME TYPE	Variable	Variable	Flag--NARSTO	Variable
*TABLE COLUMN CAS IDENTIFIER	None	None	None	C14808-79-8
*TABLE COLUMN USER NOTE	None	Total mass	Total mass	SO4=
*TABLE COLUMN UNITS	None	micrograms per cubic meter	micrograms per cubic meter	micrograms per cubic meter
*TABLE COLUMN FORMAT TYPE	Char	Decimal	Char	Decimal
*TABLE COLUMN FORMAT FOR DISPLAY	12	8.2	2	8.2
*TABLE COLUMN MISSING CODE	None	-9999	none	-9999
*TABLE COLUMN LOOKUP TABLE NAME	Site information	None	NARSTO standard flags	None
*TABLE COLUMN OBSERVATION TYPE	Supplementary data	Particles	Particles	Particles
*TABLE COLUMN FIELD SAMPLING OR MEASUREMENT PRINCIPLE	Not applicable	Filter pack--multiple filters	Filter pack--multiple filters	Filter pack--multiple filters
*TABLE COLUMN PARTICLE DIAMETER--LOWER BOUND (UM)	Not applicable	0.1	0.1	0.1
*TABLE COLUMN PARTICLE DIAMETER--UPPER BOUND (UM)	Not applicable	2.5	2.5	2.5
*TABLE COLUMN MEDIUM	Not applicable	Filters--multiple--combination of impregnated and non-impregnated (see metadata)	Filters--multiple--combination of impregnated and non-impregnated (see metadata)	Filters--multiple--combination of impregnated and non-impregnated (see metadata)
*TABLE COLUMN COATING OR ABSORBING SOLUTION/MEDIA	Not applicable	None	None	None
*TABLE COLUMN INLET TYPE	Not applicable	Cyclone	Cyclone	Cyclone
*TABLE COLUMN SAMPLING HUMIDITY OR TEMPERATURE CONTROL	Not applicable	None	None	None
*TABLE COLUMN LABORATORY ANALYTICAL METHOD	Not applicable	Microbalance	Microbalance	IC (Ion chromatograph)
*TABLE COLUMN SAMPLE PREPARATION	Not applicable	Not applicable	Not applicable	Water extraction
*TABLE COLUMN BLANK CORRECTION	Not applicable	Blank corrected	Blank corrected	Blank corrected

…and what comes after this? *TABLE BEGINS – and thousands of data values.

Completing the Data Archiving Process

•The QSSC is the source for information, assistance, and specific guidance for carrying out the final steps:

- Naming data sets and data files and configuration control,
- Submitting data to the FTP site,
- Using the Data and Information Sharing Tool (DIST) to input / export metadata,
- QSSC verifies submitted DES data file format compliance.



Data and Information Sharing Tool (DIST)

The NARSTO DIST was implemented for Supersites to support compiling data set metadata and generating archive documentation.

•Data Providers can enter metadata with the DIST metadata editor and output it in the formats needed for the NARSTO Permanent Data Archive.

•DIST also has a search capability for the entered metadata that displays links to the archived data sets.

Data Exchange Standard Read and Verify Program:

The Read and Verify Program checks numerous format and content elements of data files to verify DES format compliance.

Program Verifications:

- key characteristic values are in the picklist reference tables,
- key phrases are correctly formed,
- variable formats and format types are correct, and variable names are in reference table
- CAS numbers are in reference table,
- dates and times are properly formatted and the UTC time offset is correct,
- flags are in the flag look-up table, and
- sites appearing in the main data table have corresponding entries in the site information table.

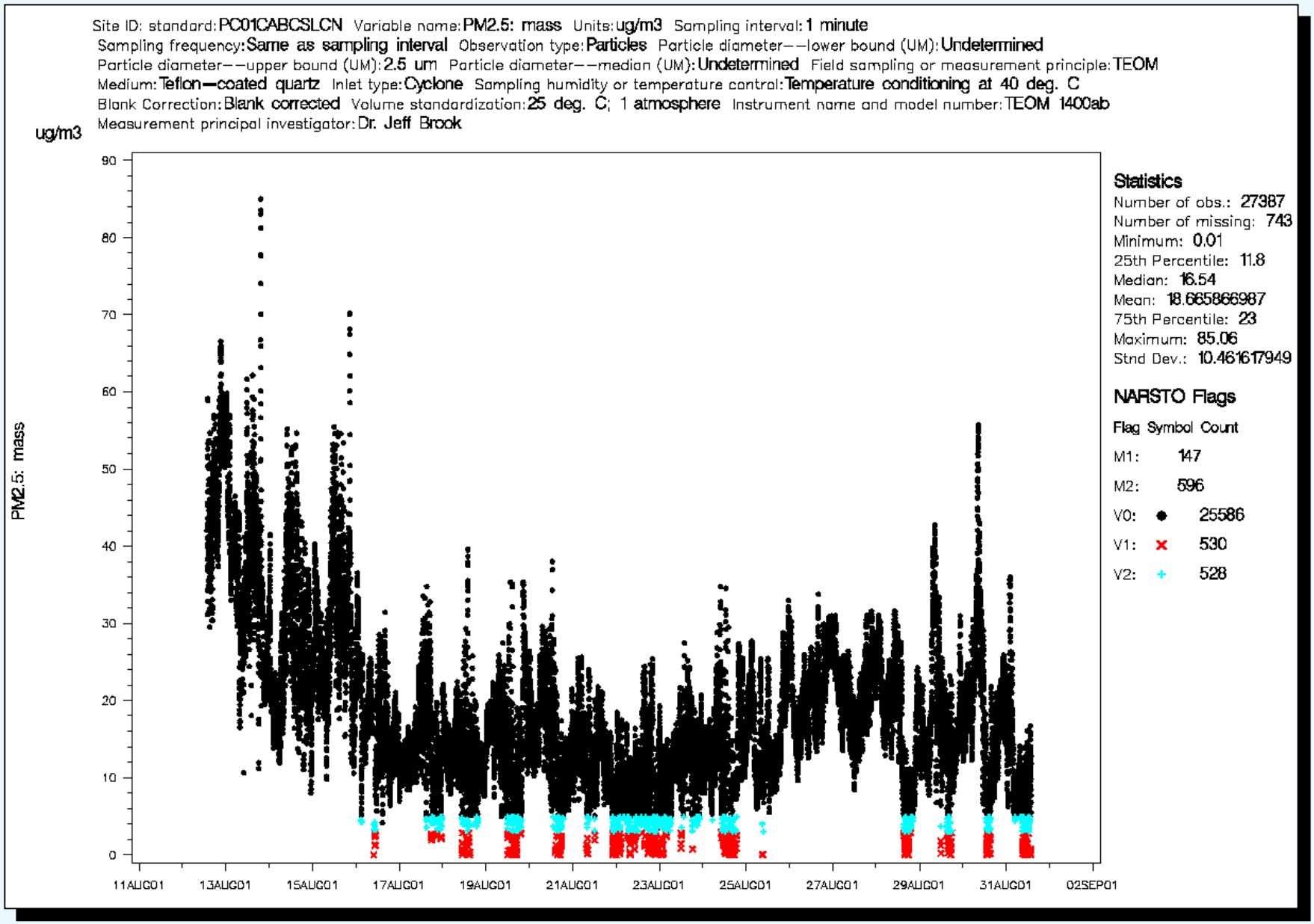
[Supersite Data Coordinators with SAS® software also run this program before submitting files.]

Program Outputs:

- Generates a QA problem report for data originator to use to correct the problems.
- Calculates and inserts into the file a set of summary statistics records for each numeric variable.
- Time series plots can be created. The plots include summary statistics and key characteristic values.

When the QSSC data file verification checks are complete and documentation is complete, the properly formatted data files can be moved to the NARSTO Permanent Data Archive.

Time Series Plot with Summary Statistics and Key Characteristic Values



All of the standards, the DES template, DIST, and guidance documents referenced in this paper can be accessed through the QSSC web site at URL: http://cdiac.ornl.gov/programs/NARSTO/.

Summary

The result of this development effort is a robust set of data and metadata reporting standards that will facilitate current Supersite PM data reporting, analyses, and archiving activities, can be extended to additional data types, and will support integrated analyses and future research projects.

By taking advantage of existing resources, the technical expertise and developing data management resources of the Supersites, and NARSTO, we were able to develop and implement a useful set of reference materials and supporting systems.

Credit for the accomplishments of this development effort are due to:

- The early recognition of the need for data management planning and implementation and the inclusion of requirements in the Supersites Cooperative Agreements,
- Participation of the Site Data Management Coordinators in the Data Management Working Group (DMWG),
- Coordination and integration of the input of Supersites data management and research staff with existing NARSTO and other applicable standards, and
- The continued support of the EPA Supersites Program Mangers.