

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

Les A. Hook¹, Sigurd W. Christensen¹, and William B. Sukloff²

Contact: Les A. Hook, 865-241-4846, hookla@ornl.gov

Technical Paper

The U.S. Environmental Protection Agency is sponsoring a major atmospheric particulate matter (PM) data collection effort in seven major U.S. cities, called the PM Supersites Research Program. The Supersites Program's objectives are to (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.

The need for consistency of metadata and data structures to facilitate information sharing among investigators, analysts, and ultimately secondary data users was recognized at the start of the EPA Supersites Research Program. Under the auspices of NARSTO³ a successful mechanism was created to develop and implement reporting standards. The development effort included working closely with Supersites data coordinators, investigators, and technical experts, and also leveraging from existing data standards and practices. Overall, the standards are getting good acceptance from the atmospheric research community.

Data reporting was addressed in the Cooperative Agreements that implement the Program. In addition to analyzing individual site PM conditions, the data from all the Supersites are to be capable of being integrated for cross-site analyses, and are to be archived in a timely manner and be readily available to the public. Data Coordinators support the data reporting process at each Supersite. The NARSTO Permanent Data Archive (PDA) at the Langley NASA DAAC was designated the final repository. The PDA has a required self-documenting data format, the NARSTO Data Exchange Standard (DES), with several metadata requirements. The archiving process is mediated by the NARSTO Quality Systems Science Center, Oak Ridge National Laboratory.

Until the initiation of the PM Supersites Program, the DES had been used primarily with gaseous atmospheric constituents and meteorological measurements (e.g., ozone, air temperature, and solar radiation). The sampling and measurement of these constituents is generally straightforward with well-defined methods and reporting conventions. It soon became clear that PM measurements are not so easily characterized. PM results (e.g., size-differentiated mass, number, and chemical composition) need more metadata than just the name, units, and analysis method. In many cases, results are operationally defined by the specific field sampling configurations, measurement devices, and conditions, and the laboratory sample preparation and analysis methods.

¹ Oak Ridge National Laboratory, Oak Ridge, Tennessee

² Environment Canada, Meteorological Service of Canada

³ A tri-national, public-private partnership for dealing with multiple features of tropospheric pollution, including ozone and suspended particulate matter.

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. The key characteristics are metadata fields that hold general descriptions of the field and laboratory conditions. Detailed information should be included in project documentation such as the Quality Assurance Project Plans. Key characteristics, metadata values, and organization of the DES were defined through invaluable interactions of data managers with PIs and with other field and laboratory technical experts.

To ensure that data can be integrated for successive analyses, consistently defined metadata elements and values are necessary. The key characteristics' values are maintained in lookup tables/picklists. Equally important has been the development of standards/guidance for naming and describing sampling sites, assigning data quality flags, data reporting units, and variable names for the measured environmental components. Variable names, for example, are highly structured, using Chemical Abstract Service (CAS) numbers for chemicals when available, or a consistently organized syntax for other measurements.

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, the CAS Registry Number and CAS Index Name for chemical identification from EPA Chemical Registry System (CRS), and the non-chemical variable naming syntax from the DOE Atmospheric Radiation Monitoring (ARM) Program.

This standards development process was successful in integrating the input of data management and research staff with existing applicable standards. The resulting robust set of Supersites data reporting standards will facilitate current and future PM data analyses and can be extended to additional data types and research projects.

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

ORNL research was sponsored by U.S. Environmental Protection Agency and 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.

The submitted manuscript has been authored by a contractor of the U.S. Government under contract DE-AC05-00OR22725. Accordingly, the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution, or allow others to do so, for U.S. Government purposes.