6 Data Management

6.1 Introduction

The number and variety of measurements in large collaborative efforts such as BRAVO generate volumes of data that must be stored in an organized, easily accessible format. DRI is responsible for assembling and maintaining the BRAVO database. This section outlines the protocol for management of BRAVO data.

6.2 BRAVO Data

There are several different types of data that are expected to result from BRAVO. They can be grouped roughly into one of four categories.

I. Automated pseudo-continuous samples (Analysis occurs at the time of sample procurement): This category encompasses data from instruments that are self-contained sample procurement and measurement devices. Typically, measurements are made at regular intervals that range from several minutes to one or two hours. Examples include surface meteorology, continuous measurement of airborne species (SO_2, SO_4^{2-}), nephelometers, and transmissometers.

II. Time-averaged samples (analysis occurs post-sample procurement): This category contains samplers that utilize a substrate such as a filter that requires chemical analysis in the lab. Generally the durations of the measurements are between one hour and one day. Examples include measurement of PM_{10} and $PM_{2.5}$ on filters, and speciated chemical analysis of aerosols.

III. Upper Air data: This category is different from the previous two because measurements can be at irregular intervals and because the same parameter(s) is measured at multiple altitudes at the same site.

IV. Size and Chemically Speciated Aerosol Data: This category includes analysis methods that break down particle measurements both by particle size and by chemical composition. SEM analysis of polycarbonate filters is an example of this type of measurement.

Table 6.1 lists the measurements obtained as part of BRAVO as well as the type of data they represent.

Table 6-1. Measurements obtained as part of BRAVO, the duration of the averaging times, and the type of data generated.

Measurement Type	Duration(s)	DataType
$PM_{2.5}$ elements (H, Na-Pb, mass, b_{abs}) (Teflon filter)	6, 12, 24 hr	II
$PM_{2.5}$ carbon (quartz filter)	12, 24 hr	II
$PM_{2.5}$ ions (nylon filter)	12, 24 hr	II
PM_{10} elements, ions, carbon	24 hr	II
SO2	6, 24 hr	II
Tracer	1, 6, 24 hr	II
$PM_{2.5}$ carbonaceous aerosol	24 hr	II
Carbon speciation by GC/MS for selected periods	24 hr	II
Gaseous nitric acid	24 hr	II
Gaseous ammonia	24 hr	II
MOUDI size resolved aerosol	24 hr	II
Scanning electron microscopy (SEM) analysis	24 hr	II
Gaseous hydro-peroxides	1 hr	I
High time resolution, high sensitivity SO_2	10 min	I
High time resolution particulate sulfate	10 min	I
High time resolution organic carbon	?	I
High time resolution particulate nitrate	?	I
Rawinsonde	Twice daily	III
SODAR Wind Profiler	1 hr	III
RADAR Wind Profiler	1 hr	III
Size resolved chemically speciated $PM_{2.5}$	12 hr	IV

6.3 Importing Data into the BRAVO Database

Data received by the data manager from the various groups that are collaborating in BRAVO has to be imported into a master database. The primary objective of the data management portion of the BRAVO study is to provide an efficient and simple way to extract desired data from a well-documented, accurate, and uncomplicated database. This requires that a thorough account be kept of all data that end up in the BRAVO database. The first step in this process is ensuring that data providers and the data manager are in agreement on a consistent, well-documented format for the raw data files. Important factors include measurement units, time reporting conventions, site mnemonics/codes, mnemonics and codes for the parameters that are measured, and data flagging conventions.

Once the conventions for reporting data are firmly in place, computer codes, written primarily in Microsoft Visual Basic and Visual C++ are used to import data into the database and convert measurement units, sampling times, measurement locations and so forth into the standard formats of the BRAVO database. In addition, during the data import process Level 1b validation is applied to each data set; it is expected that Level 1a validation is performed by the data provider (See section 6.4 for an explanation of data validation levels).

6.4 Data Validation

Mueller (1980), Mueller et al., (1983) and Watson et al. (1983, 1989, 1995) define a three-level data validation process that should be mandatory in any environmental measurement study. Data records are designated as having passed these levels by entries in the VAL column of each data file. Data providers are asked to report data only after Level 1A validation has been performed. These levels, and the validation codes that designate them, are defined as follows:

- **Level 0 (0)**: These data are obtained directly from the data loggers that acquire data in the field. Averaging times represent the minimum intervals recorded by the data logger, which do not necessarily correspond to the averaging periods specified for the data base files. Level 0 data have not been edited for instrument downtime, nor have procedural adjustments for baseline and span changes been applied. Level 0 data are not contained in the BRAVO data base, although they are consulted on a regular basis to ascertain instrument functionality and to identify potential episodes prior to receipt of Level 1A data.

- **Level 1A (1A)**: These data have passed several validation tests applied by the network operator that are specific to the network. These tests are applied prior to submission of data to the BRAVO data manager. The general features of Level 1A are: 1) removal of data values and replacement with -99 when monitoring instruments did not function within procedural tolerances; 2) flagging measurements when significant deviations from measurement assumptions have occurred; 3) verifying computer file entries against data sheets; 4) replacement of data from a backup data acquisition system in the event of failure of the primary system; 5) adjustment of measurement values for quantifiable baseline and span or interference biases; and 6) identification, investigation, and flagging of data that are beyond reasonable bounds or that are unrepresentative of the variable being measured (e.g. high light scattering associated with adverse weather).

- **Level 1B (1B)**: After data are received by the data manager, converted, and incorporated into the database, validation at level 1B is performed. This is accomplished by software which flags the following: 1) data which are less than a specified lower bound; 2) data which are greater than a specified upper bound; 3) data which change by greater than a specified amount from one measurement period to the next; and 4) data values which do not change over a specified period, i.e., flat data. The intent is that these tests will catch data which are obviously nonphysical, and such data will be invalidated and flagged. Data supplied by project participants which fail these tests may result in a request for data re-submittal.

- **Level 2 (2)**: Level 2 data validation takes place after data from various measurement methods have been assembled in the master database. Level 2 validation is the first step in data analysis. Level 2 tests involve the testing of measurement assumptions (e.g. internal nephelometer temperatures do not

significantly exceed ambient temperatures), comparisons of collocated measurements (e.g. filter and continuous sulfate and absorption), and internal consistency tests (e.g. the sum of measured aerosol species does not exceed measured mass concentrations).

- **Level 3 (3)**: Level 3 is applied during the reconciliation process, when the results from different modeling and data analysis approaches are compared with each other and with measurements. The first assumption upon finding a measurement which is inconsistent with physical expectations is that the unusual value is due to a measurement error. If, upon tracing the path of the measurement, nothing unusual is found, the value can be assumed to be a valid result of an environmental cause. The Level 3 designation is applied only to those variables that have undergone this re-examination after the completion of data analysis and modeling. Level 3 validation continues for as long as the data base is maintained.

A higher validation level assigned to a data record indicates that those data have gone through, and passed, a greater level of scrutiny than data at a lower level. The validation tests passed by Level 1B data are stringent by the standards of most air quality and meteorological networks, and few changes are made in elevating the status of a data record from Level 1B to Level 2. Since some analyses are applied to episodes rather than to all samples, some data records in a file will achieve Level 2 designation while the remaining records will remain at Level 1B. Only a few data records will be designated as Level 3 to identify that they have undergone additional investigation. Data designated as Levels 2 or 3 validations are not necessarily "better" than data designated at Level 1B. The level only signifies that they have undergone additional scrutiny as a result of the tests described above.

6.5 Database Architecture

There are two different designs for the BRAVO database, a master database, and a user database. The master database includes information that is superfluous for the day-to-day user, but important for the data manager. Examples of such information are: the line numbers in the original data files that are associated with each data point, the units used by the data provider before conversion to standard units, and the dates that data were imported into the database. While much of the information related to the data points that appear in the BRAVO database does not appear in the user version of the database, some fields such as data validity flags and sample analysis method descriptions are included for completeness.

Within the BRAVO master database, all data are stored in tables with consistent structures. Within the data tables there exists one record for every measurement that results in a datum. This record contains links to the information stored in the following fields (Actual field names are mnemonics of the field names shown below):

1. **Site_Parameter_ID:** This is a number that is unique for each combination of site, parameter measured, sample duration, particle size (if applicable), and source file (The name of the data file as provided by the data supplier). The Site_Parameter_ID number is linked to the "Site Information Table", The "Parameter List Table", the "Particle Size Definitions Table", and the "Source Data Files Table".
2. **Start_Date_Time:** Date and Time stamp indicating the beginning of the sample period.
3. **Duration:** Duration of sampling/averaging period in minutes
4. **Value:** Value of measurement.
5. **Uncertainty:** Uncertainty associated with the value.
6. **Value_Suspect:** A Flag field that contains either a "V" for valid data or an "S" for suspect data.
7. **Flag_Comment:** text field containing flags and comments as reported by the data provider.
8. **Source_File_Line_No:** The line number in the source file (data file from provider).
9. **Alt:** The altitude of the measurement (For Upper Air Data Only).
10. **Size_Bin_ID:** This field is linked to a table that contains lists of different particle size bins. This is different from the "Particle Size Definitions Table" which only contains standard particle size cuts e.g. D_p<10 μm, D_p <2.5 μm, etc. The Size_Bin_ID Field is only used when non-standard size cuts are reported from instruments like impactors, DMA, etc.

Note that traceability of data is built into the architecture of the master database. In other words, it is possible to take any record in a data table and trace the record entries back to the original source file. Likewise, using database queries, it is possible to modify/isolate a set of records by data provider, sample times, sample durations, source file, etc.

Note that the measurements obtained as part of BRAVO range in duration between 10 minutes and 24 hours (Table 5.1). Frequently, this can lead to difficulties in comparing data of different types. For example, comparing 10 minute-averaged SO_2 concentrations at one site with 24 hour-averaged SO_2 concentrations at another site requires averaging the 10 minute samples over the appropriate 24 hour period. In order to avoid cumbersome spreadsheet calculations, the BRAVO database (both master and user) contains time-averaged data tables. For example, in addition to a table that has all 10 minute-averaged SO_2 data, there are three more tables that contain those same data averaged over 1 hour, 6 hours, and 24 hours. While this design increases the amount of computer storage space required for the database, the presence of these additional tables considerably increases the speed with which different types of data can be extracted from the database and compared to one another.

6.6 End Product

In addition to being a means to safely and efficiently store data, the purpose of the BRAVO database is to provide quick and easy access to data that have been gathered as part of the study. The database will be made accessible to the different participants in BRAVO via internet. Figure 5.1 gives an example of a data request form that can be made available on the internet and can be used to retrieve BRAVO data. The user is asked to select the dates and times that are of interest, one or more sites where measurements were performed, one or more parameters that were measured, the averaging time of the measurement(s), and additional information regarding the desired format for the output file. A map of the BRAVO network on the form aids in the selection of sites that may be of interest to the user. Once the information is entered into the form, a program written in Microsoft Visual Basic for Applications will retrieve the relevant data and write a file (filename specified by user) to the DRI ftp server. The user may then download the file directly from the ftp server. The benefit of having a central database that is queried remotely is that updates to the database are available to the user as soon as they are implemented. Some users may require more complex data analysis tools or more flexible data retrieval options; in such cases, users can be provided with a copy of the BRAVO database either on CD-ROM or by specially arranged ftp.

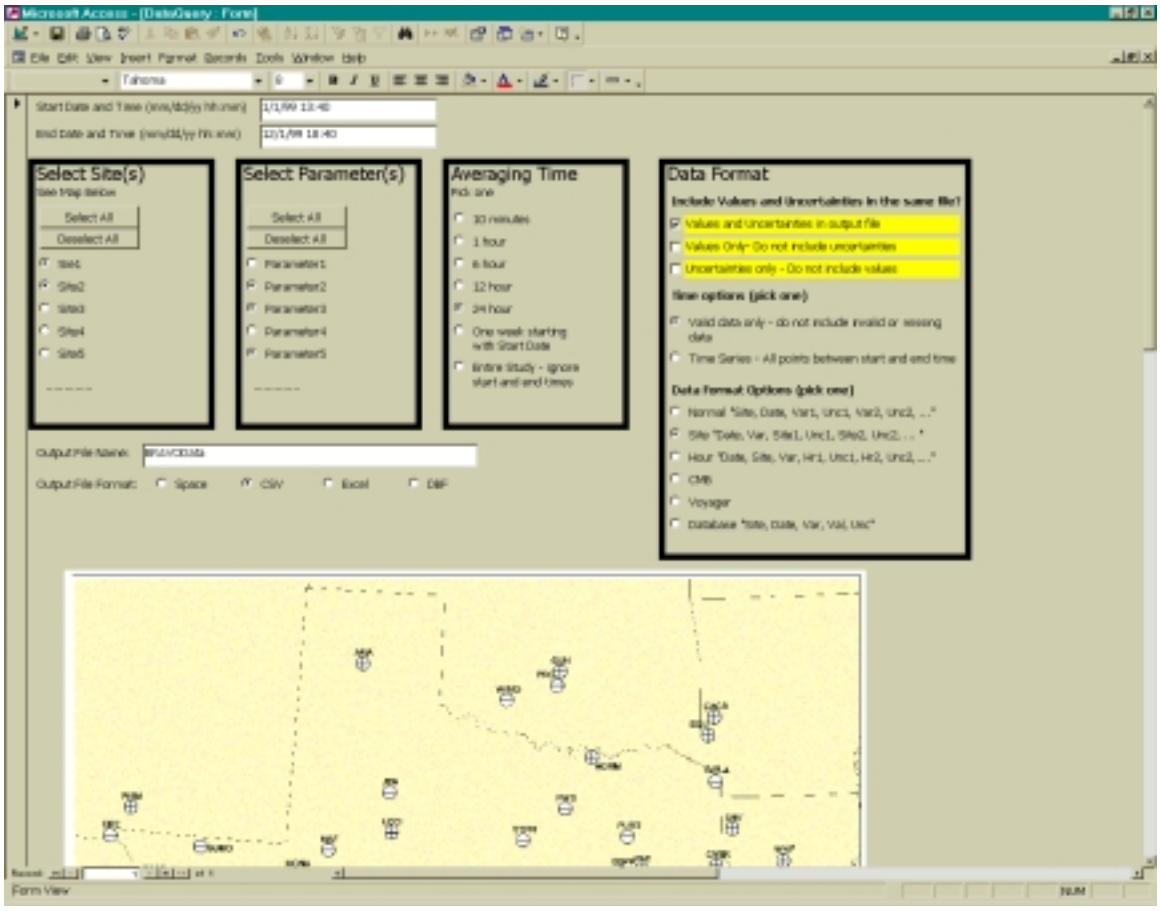

Figure 6-1. Example of form available on the internet used to retrieve BRAVO data.

A "BRAVO Database Information" web page will also be placed on the internet for user access. This latter page contains information about the BRAVO database, specifically, the date and time of the last database update, the nature of the update (i.e. what was changed from the previous version), the current status of the database, and a general description of the database.

7. Quality Assurance

A well-defined program to assure the quality of data collected in a monitoring program is essential to the credibility of its results. Each of the monitoring components (e.g. aerosol sampling, laboratory analysis, & upper air meteorology) has written protocols that describe how the method is done. These protocols also identify the quality control procedures used to avoid problems with the data and to document their quality. An independent quality assurance audit program is used to check how well the protocols, especially the quality control procedures, are being followed.

The major emphasis of independent quality assurance in BRAVO is upon verifying the adequacy of the participants' measurement procedures and quality control procedures, and upon identifying problems and making them known to project management. Although routine audits play a major role, emphasis is also placed upon the efforts of senior scientists in examining methods and procedures in depth. This approach has been adopted because fatal flaws in experiments often emerge not from incorrect application of procedures by operators at individual sites or laboratories, but rather from incomplete procedures, inadequately tested methods, deficient quality control tests, or insufficient follow-up of problems.

At the beginning of the study, senior auditors will review study design documents to ensure that all measurements are being planned to produce data with known precision and accuracy. The auditors will focus on verifying that adequate communications exist between measurement and data analysis groups to ensure that measurements will meet data analysis requirements for precision, accuracy, detection limits, and temporal resolution. Quality control components of the measurements include: determination of baseline or background concentrations and their variability; tests for sampler contamination; adequate measurements of aerosol and tracer sampler volume and time; blank, replicate, and collocated samples; assessment of lower quantifiable limits (LQL), and determination of measurement uncertainty at or near the LQL; regular calibrations traceable to standard reference materials; procedures for collecting QC test data and for calculating and reporting precision and accuracy; periodic QC summary reports by each participant; documented data validation procedures; and verification of comparability among groups performing similar measurements.

Field performance and system audits will be conducted at each of the BRAVO monitoring sites in Texas and adjoining states. Measurement systems to be audited at the majority of sites included aerosol sampling using the IMPROVE sampler and tracer sampling using the Brookhaven BATS sampler. Performance audits will include flow rate