

Designing a Data Management System for the Central California Air Quality Studies

Paper # 402

Gregory F. O'Brien, P.E.

Modeling Support, California Air Resources Board, 1001 I St., Sacramento, CA 95814

Vernon M. Hughes

Modeling Support, California Air Resources Board, 1001 I St., Sacramento, CA 95814

ABSTRACT

The Central California Air Quality Studies (CCAQS) comprise two studies, the California Regional Particulate Air Quality Study (CRPAQS) and the Central California Ozone Study (CCOS). CRPAQS is a multi-year effort of meteorological and PM10/PM2.5 air quality monitoring, emission inventory development, data analysis, and air quality simulation modeling. CRPAQS monitoring occurred during a 14-month study period, between December 1999 and February 2001. Monitoring for CCOS occurred during the summer of 2000. These studies utilized 500 instruments at over 100 monitoring locations to measure and analyze for 600 parameters. Additional data was obtained from supplemental data sources. It is anticipated that the database management system will manage 100 million hourly records.

The CCAQS Data Management System was developed to provide CCAQS sponsors and participants with quality assured data to support 3-dimensional meteorological and air quality modeling and data analysis. The CCAQS Database System operates within a Microsoft SQL Server 2000 Relational Database Management System (RDBMS) running the Windows 2000 Advanced Server. It incorporates a variety of features that make it well suited for handling and managing scientific data of this nature. To ease the task of populating data structures and performing quality assurance, a standard for data file transmittal was developed. Access to quality assured data is provided through a centralized data repository connected to the Internet. The database includes an interactive note-tracking feature with drill down capability where collaborating researchers can continuously add notes regarding either individual or groups of data observations. Other components include field and lab instrument tracking as well as tracking of filter and canister media. The system incorporates both data and project management components, web site and graphical user interface (GUI). It includes a data submittal tracking mechanism, automated quality assurance routines and flagging. GIS, data warehousing and online analytical processing (OLAP) will be added during a later phase of system development.

INTRODUCTION

The Central California Air Quality Studies (CCAQS) comprise two multi-year studies, the California Regional Particulate Air Quality Study (CRPAQS) and Central California

Ozone Study (CCOS). The CRPAQS is a multi-year effort of meteorology and PM10/PM2.5 air quality monitoring, emission inventory development, data analysis and air quality simulation modeling.¹ The data collection phase of CRPAQS took place during a 14-month study period beginning in December 1999 and ending in February 2001. The CCOS investigated ozone air quality and involved an intensive 4-month period of data collection during the summer/fall months of 2000.

One of the objectives of CRPAQS is to provide an improved understanding of emissions and the dynamic atmospheric processes that influence particle formation and distribution. Results from CRPAQS and CCOS will assist in identifying areas to control emissions when decision makers formulate candidate control strategies for attaining the Federal and State PM10/PM2.5 standards for Central California. It is also designed to provide data that can be used to model air quality, assisting in the development of a reliable means for estimating the impacts of control strategy options developed for PM10/PM2.5 on visibility and acidic aerosols and on attainment strategies for other regulated pollutants, notably ozone.

The goal of CCOS is to obtain a suitable database for grid-based, photochemical modeling.² The California Air Resources Board (CARB) and air quality control districts within Central and Northern California will use data from the CCOS to apply photochemical models. This will include examination of the effects of emissions on ozone formation concentrations and the preparation of the State Implementation Plan (SIP) for ozone for non-attainment areas in Central California.

The field programs are being followed by extensive data analysis and modeling. To support this work, the data collected during the studies must be stored, validated and made available to the Study participants. The CCAQS Database System provides the central data repository for the air quality data collected from both of the CCAQS studies, as well as for the Fresno USEPA Supersite air quality data collection effort. This document describes the database design and the application software development requirements. The system functions as the central data distribution point for quality assured data. Distribution is implemented using Internet web technology, ftp and accompanying software applications.

Database System Overview

The CCAQS Database System was developed to support the data collection and data analysis efforts of the CCOS and CRPAQS studies. The database is the central storage repository for archiving all study data and related *metadata*, which is the documentation of the data that is needed by air quality researchers. Data from these studies are accessible via an Internet server maintained by the California Air Resources Board (CARB). The CCAQS web site residing on this server has been used throughout these studies to maintain a high level of communication with Study sponsors and contractors, and to provide the conduit for meteorological and air quality data. The web site URL is <http://www.arb.ca.gov/airways/ccaq.htm>. The current design will enable study participants to access available quality assured data, in a form they request, on a 24x7

basis. This will be accomplished as users of the system submit data requests, which are then run on the database engine. The output data file is placed on the CCAQS ftp site for retrieval. As a data request is submitted and received by the system, an acknowledgement of receipt is displayed. This is followed by an email to the data requestor when the data requested are available for downloading. This email includes the necessary ftp and access instructions. Providing these sophisticated on-line tasks to participants is accomplished by utilizing state-of-the-art software.

The relational database management system (RDBMS) used for CCAQS is Microsoft SQL Server 2000. This RDBMS was selected for reasons of ease of administration, multiprocessor operation, scalability, and a rich feature set and capabilities like data warehousing with built-in online analytical processing (OLAP). The operating system is Windows 2000 Advanced Server. Microsoft Visual Basic 6.0 and InterDev 6.0 are the two primary development tools used to implement the CCAQS design and develop system and web applications. Visual Basic code is easily integrated with SQL Server 2000. During a later phase of system development, ESRI ArcSDE and ArcIMS applications will be incorporated to display data spatially and interactively with the database.

Data Sources

The CCAQS field programs consisted of months of monitoring throughout the Central and Northern California. Air quality sampling within the CCAQS network lasted from December 1999 to February 2001. The monitoring locations consisted of a combination of full-scale "anchor" sites along with "satellite" and "backbone" sites. Additional, Research Sites (RS) of types 1,2 and 3 were used for the CCOS. The anchor stations measured gaseous and aerosol species. The satellite stations were set up specifically for CCAQS to use portable monitors for measuring aerosols. Data from the existing statewide "backbone" network of the California Air Resources Board (CARB) and local air pollution control district monitoring sites were also included in the Study. Together, this monitoring provided surface and aloft air quality and meteorological measurements on a daily basis. The network utilized a number of surface monitoring stations, radar profilers, and sodars.

The annual CCAQS monitoring program overlapped with episodic field programs. These episodic "intensive operation periods" (IOPs) of monitoring occurred during the summer, fall and winter when conditions for high ozone and PM10 and PM2.5 concentrations are typical. The fall episodic program took place during a period of eight weeks lasting from October through November of 2000. The focus was on both PM10 and PM2.5 in the central portion of the San Joaquin Valley. The winter episodic field study took place over a period of eight weeks on a forecast basis during mid-November through January of 2000/2001. The emphasis of the winter field program was on the collection of PM2.5 data. Special emphasis was placed on the collection of continuous and species-specific particulate measurements that would support both receptor and grid-based modeling. Additionally, an intense episodic ozone-monitoring program was carried out in the Central and Northern California during the summer months of 2000.

Study Data

The CCAQS was a large-scale program that generated considerable amounts of air quality and meteorological data during a 14-month study period. Data from this Study is still being received. To illustrate the final volume of data expected, estimates of data quantities are shown in Table 1. The record totals were derived based on monitoring duration, sampling frequency, and the number of parameters or species sampled. It is expected that the initial size of the CCAQS database will be at least 50 to 100 million records.

Table 1. Estimates of CCAQS Database Size

Monitoring	Number of species/parameters	Number of Sites	Number of Days	Sampling Frequency	Record Totals
Annual Particulate Matter	40 species	100	100	Daily	400,000
Annual Light Scatter	1 species	35	60	Hourly (24 hrs)	50,400
Annual continuous	15 species	5	400	Hourly (24 hrs)	720,00
Annual Surface Meteorology	4 parameters	40	400	Hourly (24 hrs)	1,536,000
Annual Upper Air	3 parameters	12	400	Hourly (24 hrs at 20 elev. Levels)	6,912,000
Annual HC	40 species	3	60		7,200
Winter HC	100 species	4	15	4 per day	24,000
Winter PM	40 species	10	15	8 per day	48,000
Winter/Summer Upper Air	3 parameters	12	90	Hourly (24 hrs at 20 elev. Levels)	1,555,200
Summer Continuous	10 species	10	90	24 hours	216,000
Summer HC	40 species	6	25	4 per day	24,000
Aircraft sampling	-	-	-	-	5,000,000
CCOS Data (total)					10,000,000
QA levels 1A, 2, 3, etc.	-	-	-	-	15,000,000
Emission Inventory*	-	-	-	-	Unknown (large)
Other Special Studies*	-	-	-	-	Unknown
Post-processed Modeling Outputs*	-	-	-	-	Unknown (very large)
Supplemental Data* (AIRS, CIMIS, etc.)	-	-	-	-	Unknown (large)

* Exact quantity of data is not known at the time of publication but is expected to be very large

The Design Process

Major data management problems were encountered during past air quality studies largely because data management was not provided sufficient resources during the planning phase. Field study design, data analysis and modeling received most of the attention and funding. As a result, proper data storage, data accessibility, thorough quality assurance processes, data tracking, and data flow documentation did not receive

adequate resources. This created considerable hardship for those asked to “manage” the data after it had been collected and received. By not considering data management and quality assurance processes upfront, during the study planning, too much of the available resources had to be focused later on data formatting, metadata collection, and rudimentary data management. This took resources away from quality assurance and data analysis efforts.

It was desirable to have more balance between data collection and data management during CCAQS.^{3,4} Fortunately for this study, funding for data management was allocated, recognizing that this needed to be an integral part of the Study. A data manager was employed early on with the responsibility to develop this system. This created the first opportunity to really have developed processes in place before data files arrived. It also meant that an integrated database system could be developed using new technologies that included Internet access. There was hope for a successful system for managing CCAQS data.

To address many of the previous issues, the system objectives of maintaining data quality assurance and reporting were given highest priority. A major goal of the system was to provide easy access to data by researchers. This would enable them to direct their efforts toward analyzing the air quality data and not having to hunt for and quality assure data.

Getting User Input

The first step in the system development was to gain understanding of the data needs of the CCAQS researchers and other data users. The type of information needed from previous users included:

- List of problems encountered in the past with database systems for air quality studies that should be avoided.
- Preferred formats for both importing data and the methods for receiving data.
- Types of summary information that should be available.
- Understanding of what the system “products” should include.
- System enhancements that users would find desirable.

To begin, a survey was developed and distributed broadly to the Study participants. A summary of the survey responses was prepared that would provide input during the design process. The responses provided an understanding of what specific features users desired most. Many of these became objectives for the system during the design phase. Respondents identified a need to store data in a manner that would enable data to be queried versus only being able to get data in flat files. They also wanted better

maintenance of data files, a simple inventory of what is in the database, better data quality reporting, easier and quicker access to data files, and a better tracking mechanism. Additionally, they would need data in various formats along with direct 24x7 access capability. Based on this input, the system would need to enhance the richness of detail that describes each data point stored in the system. A major design effort began to develop a database system that would deliver these capabilities.

The Design Team

Bringing together the right individuals for a database design is the most important factor in determining a successful outcome. This requires a group comprising computer specialists, database developers, scientists and engineers that can work quickly and cooperatively and possess a willingness to revisit the design and “hammer on it” until all the elements work together well. This is the one of the surest ways to guarantee that the system objectives and design goals will be met in the final system. The CCAQS database design team collectively possessed the expertise, experience and cooperative enthusiasm that enabled it to be a small, highly focused and productive collaboration. It comprised five members, including an air quality modeler with extensive data management experience; a former air quality data manager; the CCAQS project manager; a database consultant; and the CCAQS data manager. Together this rounded out the team very well. They had a very good idea of what they wanted but did not have all the know-how to design and develop it. Therefore, outside expertise to design and build this system was needed. It was important to select a database consultant that not only had the expertise but also approached his/her role as an “assistant” to the team.

The team met each week for a period of about 2-3 months defining the core design. The first task was to list all the known system “outputs” that would be needed. This is where air quality modeling experience and the input of the survey summary were invaluable. The main components that would enable the system to deliver these outputs were described and included in the design. The team produced the first draft database logical design in two months. This was sufficient time to reach a “90% completion”. Many additional refinements were included over the next few months. The last 10% can take most of the time, but we found that generally 1 or 2 people are adequate to smooth out the remaining rough spots in the design.

The data modeling software tools and related applications helped to significantly compact design time. Microsoft Visio Modeler was used for the data modeling. This is a component of Visio Enterprise 5.0. This tool was used to develop the initial design and document the logical design of the database. It produced CCAQS database schemas that were easily modified, regenerated and available for review the following week. Visio Modeler also produced a script file from the compiled logical design that was used to quickly build the database in SQL Server. The SQL Server RDBMS produced a complete diagram of the logical design showing all database tables and data relationships. This diagram was also used during the weekly design meetings.

The design developed by the team was presented to a large group of CCAQS participants and sponsors. Based on their comments, the system would be a significant improvement over past study databases. The positive outcome was clearly a result of a compact series of intense design reviews, reiterations of the design, and input and involvement of the highly motivated team.

Initial Design Elements

The CCAQS database design incorporates a number of evolutionary advancements over past air quality study data management efforts.⁵ A fundamental step in this advancement was giving CCAQS researchers the ability to get the data they need by simply “querying” the database remotely. This would free the data manager for other administrative and development work. Another evolutionary step was using a relational database that was connected to the Internet. A less obvious step was that the relational database allowed the integration of all of the major elements from the Study into one system. The advantage of this was that it permitted the development of a scientific database that could handle all of the details of a very select set of data elements. These are what give value to the data points from the Study. Even though the data values are of primary importance, they are greatly enhanced with accompanying metadata. There are many tables in the database system design to store this metadata.

The Methods table is probably the most complicated table in the system for conveying metadata. The information that is carried in this table is intended to “describe” how an instrument was used to make a measurement. That is how a “method” is applied and defined in CCAQS. Each method record in the database comprises a unique combination of Instrument, Parameter, Size, Sampling Frequency and Duration. This table was designed with researchers in mind. If a researcher wanted to know all of the methods used by a specific instrument that could be accomplished using a query of the Methods table. To make it easier to understand and identify each method, a concatenated string of codes was developed. These codes include all of the varying parameters that define a unique method for a measurement. Data providers can readily identify the method that was used to obtain data values. The Method ID and Code are submitted with each data record.

A means of tracking files in the database was considered an absolute necessity in the overall evolution. The Submittal Log table has this role and works in conjunction with the file screening routine, which is described later. This table gets updated each time that the system initiates file processing with a new Data_Source_ID, the File_Name, Status, etc. Regardless of whether the file is accepted or rejected, an event is recorded in the log. The same file may be submitted multiple times until it is accepted. Each run creates a new Submittal_Log_ID, which identifies the file run uniquely. Once a file is accepted, it cannot be processed again. Data records are checked individually using internal unique identifiers to ensure that data will not be reentered as part of another file using another file name. From a display of the log it can be determined where a data point originated by use of the Data_Source_ID. From this log an inventory report can be created for those wanting to know what data are available in the system for download.

A tracking facility was also needed to help identify the date, time and source of acquisition for *supplemental* data. This is data that is outside the purview of the CCAQS, but is useful for accomplishing a comprehensive data-modeling program. The data format document, which is described later, is used to prepare the supplemental data and participant data for submittal. The raw data source file name associated with each of these files is based on the file naming convention for CCAQS. This can be a tedious process and requires some translation of data elements; some fields remain “null” entries because information for that field is not available. Nevertheless, the system does support the input processing from these sources quite well.

The database system also provides the capability of tracking those instruments that provided data values for the Study data. This is implemented using an `Instrument_Tracking_ID`, which references the location of the instrument and the characteristic that makes the instrument unique, such as, a serial number or other unique identifier. If an instrument moved to a new location during the Study, a new `Instrument_Tracking_ID` is created in the Instrument Tracking table identifying the new location. A separate “Instruments” table includes a listing of all of the possible instruments used to make measurements, along with manufacturer and model number, etc. Instruments are referenced in the Instrument Tracking table using an `Instrument_ID`. Taken together, what is obtained by doing this is a report that describes what instruments were used, where and when, and when they were moved, and to where. This includes whether they were used in the field or the laboratory.

There is a complement of tables that support the storage of information from the independent instrument audits. With the relationships of these tables, the audit information can be displayed for a specific `Instrument_Tracking_ID` associated with an instrument in the field. The identification numbers (IDs) are all system-generated unique numbers. This is useful because other fields in an instrument record can change, like the `Instrument_Desc` without affecting the relationships established within the database. This makes management of these records simpler and minimizes the impact of changes made to code names within the record.

For Study derived samples (e.g., filter or canister), the database stores and directly relates each air sample with the instrument that was used to obtain the sample. All air samples are individually tracked within the Sample Tracking table. This table was developed because during previous studies, samples were sometimes not analyzed and would sit in a laboratory unknown. Samples were sometimes discovered much later. But in the meantime researchers could not make use of this potential source of data that they did not know existed.

For data values derived from samples, the Methods table can be used to determine the instrument type, but not the specific instrument. This is obtained from the Instrument Tracking ID that is submitted with the data. Examining the Methods table in detail illustrates how storing extensive metadata can benefit the data user.

Storing Information in the CCAQS Database

The information used to define a “*method*”, which is how an instrument is used to measure a parameter, whether meteorological or air quality, is stored in the Methods table. Each method has a Method_ID and a Method_Code assigned to it. The measured parameter and instrument are also referenced within the Methods table. The Method_Code is the primary part of this table. It is a concatenation of all of the codes that make each unique Method_ID with the exclusion of the instrument used. For example the Method_Code: CR_ELE_PU0000002500_XFA_XRF_TEF_D1_H24_DUP is a concatenation of the following codes (descriptions are in parenthesis):

- | | |
|-----------------------------|--|
| 1. Parameter_Specie_Code: | CR (Chromium) |
| 2. Parameter_Property_Code: | ELE (Element) |
| 3. Size_Code: | PU0000002500 (Fine Particle Size Fraction) |
| 4. Device_Type_Code: | XRA (X-Ray Fluorescence Analyzer) |
| 5. Analysis_Method_Code: | XRF (X-Ray Fluorescence) |
| 6. Media_Code: | TEF (Teflon Membrane Filter) |
| 7. Sampling_Frequency_Code: | D1 (Every day 0-23) |
| 8. Sampling_Duration_Code: | H24 (24 hr) |
| 9. Type_of_Sample: | Dup (duplicate) |

Example: “Differential Mobility Particle Sizer (DPMS)”

The following is an illustrative case of a measurement system of medium complexity made up of multiple components. The case described is similar to a measurement method used during the Study. The intent of the CCAQS database design is to be able to finely differentiate the parameter measured, filter cut size (for particulate matter), device type, analysis method, media description and sampling frequency and duration associated with each data point. The Methods table has a reference to the Instrument_ID. The data providers can submit detailed method related metadata that “describes” their data well. Taken together researchers can retrieve information during data analysis and modeling that has the instrument identified and a detailed method. How will the CCAQS Database System store this information? The following example illustrates this.

The *Differential Mobility Particle Sizer (DMPS)* is one of the primary instruments for measurement of sub-micron size particle distributions. It consists of the following three components:

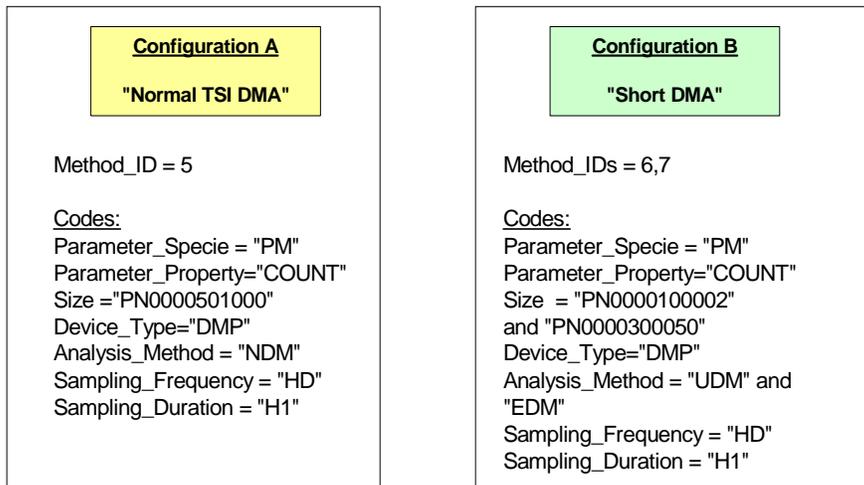
1. Differential Mobility Analyzer (DMA)
2. Condensation Nuclei Counter (CNC)
3. Computer for system control

There are several types of DMAs, depending on the particle size range being measured. In most cases, a ‘normal’ TSI DMA that sizes particles in the 5 to 1000 nanometer (nm) range is used. An alternative configuration is a “short” DMA that extends the range of particle sizing down to 1 to 2 nm. This range is possible with the use of an aerosol electrometer. An ultrafine CNC can also be used with the short DMA to extend the range from 3 nm up to 50 nm. Thus, the configurations for a DMPS include a choice of two possible DMAs and three possible detectors (an electrometer or two CNCs).

The Example 1 Diagram below for “Differential Mobility Particle Sizer” illustrates how the CCAQS database would handle the variations in instrument configuration, sampled parameters and analyses methods. Configuration A is the “Normal TSI DMA”. A unique Method_ID and a concatenated string of codes comprising the Method_Code are established within the Methods table for this instrument. The parameter being measured or counted is particulate matter (PM). The Parameter_Specie_Desc is ‘Particle Size Distribution Count’ which has a Parameter_Specie_Code of “COUNT”. The “COUNT” code becomes part of the Method_Code. The given particle size range captured would be a result of the detector properties and is uniquely identified by the Size_ID. A Size_Code equal to ‘PN0000501000’ is assigned for the TSI DMA size range specified above. This depends on the particle size fraction specifications of the CNC used. The DPMS has an associated ‘device type’, which has a Device_Type_Code of “DMP” assigned to it in the Device Types table. Similarly, codes for Analysis_Method, Sampling_Frequency and Sampling_Duration also need to be determined.

Example 1 Diagram: Differential Mobility Particle Sizer (DMPS)

The numbers and character codes assigned below are for illustration purposes only.



For illustrative purposes, the Method_ID is assigned identification “5” in the Methods table. The Analysis_Method_Code field found in the Analysis Method table identifies the analysis method associated with a Normal TSI DMA with CNC as “NDM”. This factor is important in differentiating the DPMS instruments in the Instruments table.

Configuration B will result in two new methods for the “Short DMA”. The Analysis_Method_Code is either “UDM” for the Short DMA with Ultra Fine CNC or “EDM” for the Short DMA with Aerosol Electrometer. The two Size_IDs play a role as well in differentiating between these methods associated with the Short DMA that identify how the instrument was used to obtain a measurement. The three Method_Code examples for the DPMS would be as follows assuming the frequency of sampling is every hour, “HD”, and that the duration of sampling of one hour, “H1”:

Normal TSI DMA

PM_COUNT_PN0000501000_DMP_NDM_HD_H1

Short DMA with Ultra Fine CNC

PM_COUNT_PN0000100002_DMP_UDM_HD_H1

Short DMA with Aerosol Electrometer

PM_COUNT_PN0000301050_DMP_EDM_HD_H1

With an understanding of the reference data codes used in the CCAQS database the concept of a Method_Code makes it easier to see how an instrument was used. It provides method related detail at a glance.

Data Formats

New data flow processes were introduced that would dramatically improve the level of performance end-to-end. One that generated major discussion surrounded data formatting and file transmittal. Contractors providing meteorological and air quality data typically utilize *denormalized* or “wide” formats. These files place many data values on each row within a file. (Air quality data management systems like the current EPA Aerometric Information Retrieval System – “AIRS” use this format. Much extra processing and code development is involved). These files have the benefit of being compact. Air quality researchers make use of them extensively. These formats work where “flat” files are sufficient and researchers are able to sort and quality assure the data. However, this is usually very time consuming. Air quality researchers like to work with data in Microsoft Excel spreadsheets, which are generally denormalized. The combination of denormalized formatting and the variety of air quality and meteorological files makes it difficult to efficiently transfer the data into a relational database like the CCAQS system efficiently. To manage and process large quantities of data efficiently in the denormalized format would be almost impossible. The concept instead was to take data in these denormalized files and normalize them for easy inputting, screening, quality control and storage in a relational database. Researchers then have freedom and flexibility to select specific data sets using whatever criteria they choose. It can be transferred back into a spreadsheet or an ASCII flat file directly. It has been quality controlled and ready for immediate use in a modeling or analytical application.

Prior to CCAQS, no resources were allocated towards trying to develop a specification to standardize the layout and structure of the incoming data files. Without such a

specification, it often was uncertain what “fields” in the data records were calculated fields and which were measured fields. Metadata type data elements were sometimes not included in the file itself. This included important items such as where data were collected. File notes arrived separately or in hard-copy form and had to be managed. Much of the data ‘history’ was maintain in hardcopy because it was not available in electronic form. Not setting any standards for the data providers made it essentially an open-ended process. It put a great burden on the data management staff. Specialized data processing codes had to be written for literally every incoming format and standards were “dynamic”. As a result, endless data management and handling issues were encountered. Generally, it was the data manager that was given the responsibility to determine the file contents and to place the data into a normalized database management system such as Microsoft FoxPro. There was much discussion between the people responsible for managing the data and the data contractors. This attempt to resolve issues with data required several months. As a result, it typically took several weeks or months for data to end up in a database system. In addition, answers to some questions regarding the collection of data could not be traced because the individual responsible for generating the file was no longer employed at the company.

The CCAQS design team wanted a more “efficient” means of handling and processing incoming data files that would avoid many of these problems. A more “universal” format specification for all the CCAQS data was required to minimize the problems of the denormalized format. The concept of using a normalized data file format was discussed and determined to be a technically superior way for processing and quality controlling incoming data. It is much easier to develop business rules and constraints on incoming data that make storing invalid data nearly impossible. A *normalized* format uses one data value per record in the file. Although a normalized file is much larger than the same file in a denormalized format, it provides a number of benefits for applications. It means that air quality and meteorological data files become more similar in structure and can be processed similarly. With a format specification, an automated file screening process was developed around it to check for errors and to ensure consistency in the incoming data. Otherwise, the data is not transferred into the database. This has enabled the database manager to eliminate the time consuming work of determining exactly what data were submitted, deciphering the format, and writing programs to format the data into a common database format. Most importantly, the turnaround time of a week or less between file submittal and data availability has become more typical.

Describing the key elements of the new database system including the normalized record and file format and other specifications relevant to obtaining the critical metadata for the database are included in the *CCAQS Data Transmittal Format Document (DTFD)*.⁶ This is a detailed data-formatting definition was developed by the team and distributed via the CCAQS web site. After completing the work on the normalized format specification, it was later learned that the Re-engineered AIRS System (AIRS II)⁷ was also incorporating a normalized input file format, for many of the same reasons. This new approach has required additional programming by the data contractors to translate their files from a denormalized to a normalized format, which the contractors have agreed to do. Their cooperation was obtained and appreciated. The cooperation of contractors to comply

with the new specifications for data format and file transmittal has streamlined the implementation of improved data delivery performance and quality control.

Transmittal Files

Automated CCAQS data file processing is built around the CCAQS DTFD specification. This document defines a file layout that is much like electronic file transfers (EFTs) used in banking. A typical EFT file is comprised of several record types, such as a header record, multiple body records and a footer record. This file structure makes it suitable for use in an automated process that is programmed to accept specific items in a given sequence. The CCAQS data files are ASCII comma delimited text files utilizing a standard ASCII file definition. Each record type encountered in a data file has a special identification code shown in Table 2 below. A file header, observation records and file footer are required.

In addition to checking for the proper order of records within a file, the process checks that each record “layout” is in conformance with the CCAQS DTFD specification. At the same time, it checks to ensure that the data elements found within the file are listed in the CCAQS Reference Tables. These tables are available to data providers as a set of Microsoft Excel spreadsheets and as a web-based reference table application where data providers can submit reference data directly. They are used in conjunction with the CCAQS DTFD as “pick-list” in spreadsheet form of the data elements needed to correctly complete the file records.

Because the CCAQS database system has automated file format screening and quality assurance, data can be available to researchers much earlier than past databases. The screening routines were developed as Visual Basic applications that are used to review incoming data files. This assures compliance with the CCAQS data transmittal and formatting requirements. An essential part of this involves identifying file transmission problems. “Validity checkpoints” are incorporated into the file format for this purpose. The screening application includes code that uses these checkpoints so that each record within each incoming file is verified to be accurate and intact. Otherwise, error messages are sent by the system via email to the data manager and data provider. The *codes* that are associated with each record type defined in Table 2 below.

Table 2. CCAQS “EFT” Record Identification Codes

Code	Record Type	Required or Optional
1	File Header	Required
2	(This code is reserved)	(not used)
3	File Note	Optional (with data file transmittal)
4	(This code is reserved)	(not used)
5	Obs Note Header	Required (with Obs Note transmittal)
6	Obs Note	Optional
7	Obs Note Footer	Required (with Obs Note transmittal)
8	Observation	Required
9	File Footer	Required

To encourage data providers to submit data related notes, an ability to include file and observation notes (as separate records within the data file) was included in the specification. These notes were previously received as separate text files, accompanying the data file. This made it difficult to manage and relate notes to a corresponding data set. As an alternative, the CCAQS system incorporates both the File Note and Obs Note within the same file as the data. A File Note Record (Code 3 in Table 2) is used to communicate information relevant to an entire collection of observations found within a transmitted data file. These notes can be of an indefinite length, permitting very comprehensive file notes. For example, a file note is useful for describing the impact on data from an exceptional event such as a forest fire or nearby agricultural tilling activity. An Obs Note Record (Code 6 in Table 2) allows for any number of observation notes to be included in a file. Up to three of these observation notes can be associated with any given data record. This required some special handling in the database. The Obs_Notes associated with each record in a file are written to a “temporary” Obs_Notes table, actually a temporary array in memory, as shown in Step 1 below.

Temporary Obs_Notes Table

Step 1.

Note Number	Air Obs Note ID	Obs_Note
1		“Note 1 Text”
2		“Note 2 Text”
3		“Note 3 Text”

Each individual note causes a transaction to occur as they are written to the Air_Obs_Notes table. A unique Air_Obs_Note_ID can then be obtained (Step 2 below). These IDs are then copied to the temporary Obs Notes table and associated with the note numbers that were assigned to them in the incoming data files. Consequently, every note is given a unique ID within the database system and the original note numbers are discarded (Step 3 below).

Air_Obs_Notes Table

Step 2.

Air_Obs_Note_ID	Obs_Note s	...
...
123	“Note 1 Text”	...
124	“Note 2 Text”	...
125	“Note 3 Text”	...
126	“Note 4 Text”	...
...

Modified Temporary Obs_Notes Table

Step 3.

Note Number	Air Obs Note ID	Obs_Note
1	123	“Note 1 Text”
2	124	“Note 2 Text”
3	125	“Note 3 Text”

At Step 4 the observation data becomes associated with the corresponding observation notes in the Air_Obs_Matrix table. Air_Obs_ID and Air_Obs_Note_ID form a compound key. A one-to-many relationship is established between the unique Air_Obs_ID in the Air_Obs table in Step 5 and the Air_Obs_ID in the Air_Obs_Matrix table. All the observation notes can easily be displayed with each corresponding data record as part of the graphical user interface.

Air_Obs_Matrix Table

Step 4.

Air_Obs_ID	Air_Obs_Note_ID
1	126
1	123
2	125
2	126
2	124
...	...

Air_Obs Table

Step 5.

Air_Obs_ID	...	Obs_Value	...
1	...	0.1	...
2	...	1.4	...
...

A goal of the database system was to have the capability to view each observation data value along with other relevant information, including notes. The system provides this capability. The database system stores all of these notes and associates them with the corresponding data values and can easily display them together. By incorporating notes into the format specification, there is a channel for communication and information flow from the data provider to researcher. There is more continuity because the notes can report a data history from the field, the laboratory, and the data analyst.

User Interface Requirements

Providing a means of general access to the CCAQS Study data was the heart of all the design effort. An Internet web site was developed that would include a Graphical User Interface (GUI) for remote database management, metadata entry and data access via a browser. The CCAQS GUI development effort can be categorized into three different areas: viewing observation data, database querying and reference data maintenance. In

the current system there is no direct observation data entry. Reference metadata elements can be entered for such things as new instruments and chemical and meteorological parameters, monitoring station updates or corrections. The system has a GUI for viewing individual data points and associated file and observation notes. Researchers and data modelers can submit comments regarding individual data points. For making data requests, there is a GUI that builds a query from user input. This provides query selection dialog boxes for specific parameters, station, date range, method, instrument, etc. The output file from the query can be reviewed directly within a display built into the GUI. The full file from the query can be directed to an ftp site for downloading later.

For data management, a separate interface exists to maintain and update individual tables in the database. There are displays provided for updating data elements within the main reference tables such as Methods, Parameters, Supports, etc. Tables can be searched, added to, or modified. The database administrator does the deletes at the database level. This makes it easy for the data manager to add new elements to the tables from a web browser. The data providers can also submit additional data elements to these tables. This makes it easy for them to package and submit new data files that include data elements that were not previously in the database. These submittal updates are applied to a copy of the 'production' database system. Differences between the two databases can be obtained readily. New submittals are reviewed before updating the production database. Once the reference data is in the database the file screening routines can accept new data file submittals without problems.

Revisiting the Design

The need for a different approach to the original CCAQS "Sites" table became apparent when contractors began preparing their data for transmittal. What do they do with airplane data or data obtained at a given height on a tower? What if someone wants to query the data for a given tower at a given height? As it was, the design could not accommodate this easily. More and more air quality measurements are coming from sources other than "sites" or "stations". This was true for CCAQS where there were airplanes, blimps, portable vans and towers used to collect samples or make air quality and meteorological measurements. Therefore, there needed to be a better way to accommodate these within the database. It was determined that the conventional Sites table needed to be redesigned. It was not desirable to maintain a Site_ID and Site_Code as well as additional codes that would group sites into a 'collection' with airplanes, blimps, etc. This would require separating sites and adding two new fields for these other means of supporting instruments that make measurements.

The scope of the Sites table was broadened and renamed to "Supports". The "Supports" table replaced the Sites table in the database. The word "Supports" was considered more inclusive of all the classes of things that support instruments in the field from mobile vans to space satellites. To make the collection of supports more consistent, the concept of a 'Site' was explored. These were simply *stations*, which are fixed ground level buildings, trailers or monitoring equipment, with associated location identifiers (lat/long) and elevation. A station is therefore like a temporary tower or mobile van, which can

also have locations. A Support_ID and Support_Name replaced the Site_ID and the Site_Name for all the existing sites in the table. The appropriate changes were made in the CCAQS DTFD. Data providers now had the Support_ID and Support_Code to define a specific height on a tower that would have ground location information or an airplane that did not. The “views” that the system displays to the user and which rely on the Supports table are developed to refer to ‘station’, ‘airplane’, ‘tower’, etc. The researcher or other data user does not need to know that the underlying table that stores and provides this information is the Supports table. Views are used this way to avoid creating confusion.

The reference ground based station (Ref_GBS) field was added to the new Supports table. A researcher can pull out all the information for one monitoring site. This includes data from mobile vans and temporary towers that were monitoring at that location. In the case of a mobile van, if it heads off somewhere else to monitor, it is given a new identification and code for the new location. The Support_Name remains the same to confirm that the same van was in both places.

Data Quality Assurance & Validation

The basic framework for all CCAQS data processing uses four-levels of data quality assurance (QA) designation: Level 0; and Levels 1, 2, and 3 as defined and applied by Mueller^{8,9} and Watson¹⁰. Associated with these data quality levels is a comprehensive set of data quality flags. CCAQS flags provide data value specific quality indicators and are invaluable to data analysts and air quality modelers. They are a source of “documentation” for supporting data validity and maintaining the credibility of the data. This information is critical for air quality modelers.

The CCAQS system has been developed to encourage greater use and dependence on flags to determine data quality. In the past, a -99 or a “NULL” value replaced missing measurements and included no other quality indicators. Instead of relying on a flag to specify data being invalid, data values were discarded entirely and replaced with -99 or 9999 (a value that would not normally occur in the data set). The concern among air quality researchers was that invalid data could be used incorrectly if it remained in the data file. With proper data management there no longer is any reason to continue either of these practices. For maintaining, “data of record” and data history, invalidated data retains the original value and are flagged using the appropriate invalid flag code (INV). The CCQAS centralized database simply does not provide access to invalid data so this prevents dissemination and use of this data.

The CCAQS Data Transmittal Format Document (CCAQS DTFD) was developed to accommodate Primary, Secondary and Activity Flags. The *Primary Flag* indicates the “status” of the data. There are three variations of primary valid flags: V0, V1, V2. The other three are S (suspect), M (missing), or I (invalid) as shown in Table 3 below. This is very similar to the flagging scheme used by NARSTO to archive air quality data.¹¹ The use of a primary flag can result from conditions encountered during field sampling, field or lab measurement or the initial data analysis phase. They carry the highest-level of

quality designation in the hierarchy of quality assurance. The *Secondary Flag* is used to reflect the most important sampling, measurement or analysis consideration affecting the data. They provide more detailed information pertinent to the data. The data manager can add other flags to these lists as needed. The *Activity Flag* is used to identify exceptional events or certain environmental conditions that could have influenced the data monitoring in the field.

The data validation process will be incorporated into the CCAQS database system.¹² This will require implementing another QA flagging process and a mechanism to enable researchers to collaborate during the data validation process.

TABLE 3. CCAQS Study and Data Validation Flags

Process Point	Primary Flag	Secondary Flag	Activity Flag	QA Level	Internal CCAQS DB Flags
Field/Lab/Data Analysis	V0, V1, V2, S, M or I. (See Diagram 2 below)	Detailed flag. Dependent on the Selection of the Primary Flag. (See Diagram 2 below)	Exceptional Event or Environmental Condition Flag (See Diagram 1 below)	0 - 1A	NA
Data Validation (post-sampling & measurement analysis)	NA	NA	NA	1A, 1B, 2, 3	TBD

CONCLUSION

The CCAQS relational database was developed to store air observation data and other relevant information from the Central California Ozone Study (CCOS) and the California Regional Particulate Air Quality Study (CRPAQS). The system design applied a relational database model that was found to be an efficient and effective way to collect, qualify, store and deliver large quantities of quality controlled data. With the proper design, one that enforces adherence to business rules, storing invalid data is nearly impossible. The benefits of requiring a standard format and record layout were demonstrated. Use of a common data format with file screening greatly decreased the time between file transmittal and data availability on the Internet. It also ensured that the important information could be obtained at the front of the process, thereby reducing the detective work that often results when data file contents are not described adequately. Keeping the system output products in mind determined the detail of data record content and storage mechanisms that were necessary to provide the level of information for CCAQS researchers. Once air quality data is organized within a relational database

structure, other applications can be readily integrated with it. These include GIS applications and OLAP technologies which can greatly increase understanding and broaden the audience of users. As technology advances, data from some instrumentation will require new approaches for storage and access. For CCAQS a relational database was a significant advancement in air quality study data management. Putting together interconnected well-designed processes make all the difference in meeting the needs of data analysts and modelers alike.

ACKNOWLEDGEMENTS

NARSTO, Les Hook and Sig Christensen for their invaluable input

T & B Systems, Liz Niccum for all her efforts in developing the database system

Capital DataWorks, Sacramento, CA, for developing just what we asked for

Microsoft Corp., Sacramento Offices, for software and technical support

REFERENCES

1. Watson, J.G., et al, *Aerometric Monitoring Program Plan for the California Regional PM2.5/PM10 Air Quality Study*, Desert Research Institute, Doc. 9801.1D5, Dec. 20, 1998.
2. Fujita, E., et al, *California Ozone Study – Vol. 1 Field Study Plan*, Desert Research Institute, Nov. 11, 1999.
3. Hackney, R., Hughes, V., Niccum, E. *Collecting, Managing and Displaying Air Quality Data for Large Scale Studies*, California Air Resources Board (CARB), Nov. 1993.
4. Hughes, V., Hackney, R., *Planning and Designing a Comprehensive Data Management System for a Large-Scale Air Quality Study*, AWMA Proceedings Regional Photochemical Measurement and Modeling Studies, Vol. III, Nov. 1993
5. O'Brien, G., Hughes, V., Holmes, M., Niccum, L. *CCAQS Database System Design*, CARB, <http://www.arb.ca.gov/airways/ccaqs.htm>
6. O'Brien, G., Holmes, M. *CCAQS Data Transmittal Format Document*, CARB, <http://www.arb.ca.gov/airways/ccaqs.htm>
7. Re-Engineered Aerometric Information Resource System (AIRS), US EPA, <http://www.epa.gov/ttn/airs/aqs/reeng/index.html>
8. Mueller, P.K., "Comments on the Advances in the Analysis of Air Contaminants". JAPCA 30:988, 1980.

9. Mueller, P.K., Hidy, G.M., Baskett, R.L., Fung, K.K., Henry, R.C., Lavery, T.F., Nordi, N.J., Lloyd, A.C., Trasher, J.W., Warren, K.K., and Watson, J.G. *Sulfate Regional Experiment (SURE): Report of Findings*. EPRI Report EA-1901, Electric Power Research Institute, Palo Alto, CA, 1983.
10. Watson, J.G., Roth, P.M., Ziman, S.D., Neff, W.D., Magliano, K.L., Pederson, J.R., Solomon, P.A., and Thuiller, R.H. *Data Analysis Plan for the San Joaquin Valley Air Quality Study/Atmospheric Utilities Signatures, Predictions and Experiments (SJVAQS/AUSPEX) Program*. Report No. DRI Document 8932-001.1F1, Prepared for Pacific Gas & Electric Co., San Ramon, CA; San Joaquin Valleywide Air Pollution Study Agency, Fresno, CA; California Air Resources Board, Sacramento, CA, Desert Research Institute, Reno, NV, 1993.
11. Christensen, S., Boden, T., Hook, L., Cheng, M-D. *NARSTO Data Management Handbook*, ORNL, Feb. 4, 2000.
12. O'Brien, G., *Data Flow, Application of Data Quality Flags, and Data Validation Processes for CCAQS – Preliminary Draft*, CARB, <http://www.arb.ca.gov/airways/ccaqs.htm>