Final Report for PGA 582-5-65494-FY08-05
An Advanced Database Resource for Ambient Atmospheric Data

Submitted to

Texas Commission on Environmental Quality

by

Dr. Bernhard Rappenglueck
Department of Geosciences
Institute of Multi-Dimensional Air Quality Studies
University of Houston
Houston, TX 77204-5007
Phone: (713) 743-1834, Fax: (713) 748-7906
brappenglueck@uh.edu

Dr. Dean B. Atkinson
Department of Chemistry
Portland State University
Portland, OR 97207-0751
Phone: (503) 725-8117, Fax: (503) 725-9525
AtkinsonD@pdx.edu

Overview of Database Products

Phase 1 Database (earlier support from TCEQ)

The data sets from the TRAMP and TexAQS 2006 campaigns were uploaded and stored at our secure web location http://texaqs.research.pdx.edu/. The PSU ARC team created a flexible upload interface that could accommodate the structured data format of the ICARTT file or a more user-friendly text-based upload tool, but still maintained the useful *metadata* provided in the ICARTT format, developed by the NASA Tropospheric Chemistry Integrated Data Center for the International Consortium for Atmospheric Research on Transport and Transformation. Any number of datasets could be merged onto a common time interval with a single fixed time range or with selectable time ranges for "first look" visualization. The merged data could also be downloaded to the host computer in a versatile comma or tab delimited text file. These text data files can easily be imported into a variety of software packages (*e.g.*, Excel, IgorPro, PSIPlot) for further statistical analysis or presentation. The database also provided a way to create a "wind rose" – a graphical wind directional analysis that can indicate the primary direction from which a pollutant is seen to originate. Finally, datasets could be extracted from the database in ICARTT format (whether or not it was originally in that format) for use in other organization's database or analysis software.

Phase 2 Extended Capabilities (this work)

Although the data from the TRAMP study was one dimensional in the sense that it was all taken at the same location as a function of time, there were several datasets from the TexAQS 2006 study that were multidimensional in space and time. Examples include the data from the NOAA Research Vessel *Ronald H. Brown*, the pollution monitoring SOF box analyses, and the research aircraft (NOAA NP3, CIRPAS Twin Otter, etc.) that participated in the TexAQS studies. The Houston Triangle study, although stationary, was designed to probe an area of Houston and thus would profit from a consideration of the spatiotemporal aspects of the data, especially in connection with other datasets. There are also several other data sets that could be very useful in the analysis of the air quality data – for example, the TCEQ meteorological, ozone, and PM monitoring networks, the traffic modeling and monitoring information from the department of transportation, and census information.

The second phase of the database project concentrated on constructing a new user friendly and highly functional online mapping interface. In order to provide the mapping interface, the phase 1 database was redesigned to be a truly four-dimensional (3D plus time) GIS-enabled database and to include some desirable user-interface upgrades, as explained below. The mapping interface includes a data export tool, some basic attribute, spatial and temporal query interfaces and addition statistical analysis and visualization tools. A larger subset of the TexAQS 2006 datasets was uploaded to the new database (*vs.* Phase 1) to allow for more testing and use of the mapping features. The new database has not yet been moved to a final URL location. The PSU PI (Dean Atkinson) can provide a demonstration and a provisional link for beta users.

Although it may seem a bit redundant, in the discussion below we will refer to the time-based data side of the database (note that the two functionalities are, at present, two distinct application sites) as the Data interface and to the other as the Map interface.

Data interface

Rationale

The phase one project that allowed for a flexible data merge and some rudimentary visualization was quite successful, so the new data interface was constructed to mimic most of its functions, with a few notable upgrades. It is noteworthy that although the look and feel of the earlier database was preserved in the current incarnation, it is quite different at a technical level. The new database is based on a truly four-dimensional coordinate system. Where the placement of data on a common time-scale was supported by the earlier version, the current version can be interrogated for spatial overlap irrespective of temporal overlap.

Features retained from Phase 1

- All of the data from the Phase 1 database was moved to the current database. Only about 30% of the original data was found to be spatially variable.
- A menu-based data search utility and temporal average and merge capability. Single or multiple Timespan selections may be used. The multiple Timespan selection can be used to focus on particular time segments (Mondays, all afternoons, etc.)
- Rudimentary time series plots of multiple data sets from the temporal merge.
- ASCII download of averaged and merged datasets for analysis in other software packages.
- ASCII data uploads in comma or tab-delimited file formats. The user is required to input the ICARTT metadata, using a data entry form, and the ASCII file must adhere to certain specifications (data in columns, time first, followed by column-labeled data.)
- Fully ICARTT compliant download of data (that may have been uploaded in either ICARTT or ASCII format) for use on other data platforms.

Upgrades for the current version

- New options were added for locating data using the Merge and Analyze function. Old options were PI, Organization, and Table Name (which was rarely used and therefore eliminated.) New options are Collection, Sample Type, and Subtype (*e.g.*, chemical or aerosol.) An automatic filtering of the available data occurs as selections are made.
- The metadata stored in the database was systematized and streamlined to be more useful for the current project but still be consistent with the ICARTT format.
- Statistical analysis/visualization functions (implemented by the open source R program) have been added to both the data side and the map interface side of the database to analyze and visualize selected data. (The data is selected as before by time in the merge and analyze section of the data side and by circle or polygon on the mapping side.) These functions include Q-Q plots, Histograms, Line plots of time-series data, Scatter (multiple series vs. selected series) plots (Linear Regression to be added soon), and Scatter Density plots. These are relatively low quality plots that are intended for first look analysis. It is

anticipated that the data download would be used to port the data into more sophisticated data analysis and presentation software for eventual presentation.

- All of the publicly available data from the NOAA *R/V Brown* and NP-3 research aircraft were downloaded and placed in the database by ARC staff. This provided a large set of spatially variable data relevant to the map interface.

Map interface

Rationale

Invoking the map interface menu tab opens a graphical overview of the study area. The eleven base layers for the map (plus the general city and land contours) include a set of demographic and air quality relevant layers for the Houston-Galveston metropolitan Area (HGA.) The default selection includes the city boundaries, parks and green spaces and major roadways. We also have included a subset of Points of Interest including historical landmarks, hospitals, and community centers. Users will be encouraged to submit shapefiles for other layers of broad interest.

The data may then be added to the map (each dataset is treated as a separate layer) and multiple sets can be interrogated using circle or polygon area selection tools. Each time a new data layer is added to the interface, it is cached to improve the display speed on the next access. At present, there are 45 layer types (equivalent to the Collection designation on the data side and usually defined as a combination of a PI with a platform) comprising 246 individual data layers (some meteorological, some chemical, some aerosol related.) As noted above, we hope to continue adding data layers, including the emissions inventory and measurement networks from the TCEQ in the future.

Map interface features

- Intuitive interface for base and data layer selection with layer caching for improved speed after first access
- Standard zoom and pan features for map display
- Selection of displayed spatial data by user-interactive circle or polygon drawing tool
- Statistical summaries of data at full layer or selection level
- As noted above, first-look graphs of selected points may be produced in the map interface (Q-Q, Line, and Histogram)
- Shapefile download (ESRI compatible) of selected areas or complete layers
- Image downloads of areas and layers

Two key planned features of the database proved to be more challenging than anticipated: Task 4.2 from the original statement of work - Working Maps - Spatial Analysis: create contour maps using the Cubic Spline, IDW, and Krigging methodology and Task 6.1 - Data Tables - Spatial Analysis: create contour maps using standard mathematical procedures (Cubic Spline, IDW, Krigging methodology) have not been completed to date. Although it was possible to create a user-friendly tool with default settings that could interpolate data, the sparseness of the data examined to date require more active participation by and experience in the user in order to provide a useful interpretation. When more spatially dense data are incorporated into the

database, we will attempt to develop a contour development tool with both basic (for new users) and more advanced functionality.

Quality assurance/quality control procedures

Preliminary verification of the database resource is being performed by the PSU team. Matching existing data (external to the database) with processed data has been the primary method of QA/QC. Secondary testing will be accomplished by beta testing by other TRAMP/TexAQS 2006 study participants in the near future.

Technical Specifications of Database

Server/Base

Apache HTTPD

Apache HTTPD 2.2.3 is used to serve client requests to both the data and mapping sites. Instances of Apache running on separate servers is used — one hosting the mapping application and one for the data application.

PHP

The server side application was built using PHP 5.1.6, a high versatility scripting language.

R

R 2.6.1 is being used to execute statistical calculations.

ARC Executioner

The ARC Executioner is used to run R jobs that are too large to run directly in real-time within the database.

MooTools

MooTools is a JavaScript library that facilitates the rapid development of interactive user interface elements. It was used to provide basic support for animated transitions and a palette based user interface.

The Yahoo! User Interface Library

The Yahoo! User Interface Library (YUI) contains several components that allow for simpler cross-browser compatible development. YUI was used to provide JavaScript controls and CSS definitions that created a consistent look across major browsers.

Database

PostgreSQL

The database is hosted using PostgreSQL 8.1. PostgreSQL allows for many advanced database features, such as transactional support, and also permits the use of extremely large tables, necessary for this project.

PL/R

PL/R 0.6.2.8 is the procedural language used that allows for the creation of stored procedures in PostgreSQL that execute R code. This allows for data to remain in the database and avoid the extraction and transmission of data to an external R process.

PL/Perl

PL/Perl is the procedural language built into PostgreSQL that allows for the creation of stored procedures in PostgreSQL executing Perl code. This was used to facilitate the creation of graphics and data files from within the database. PL/Perlu is the “untrusted” version of PL/Perl

which grants users access to the filesystem and thus must have all procedures created by the superuser.

Geospatial Tools

MapServer

MapServer 5.0.2 was used to generate cartographic output of data. Interfaces to MapServer were built using PHP/MapScript, MapServer's built-in PHP API.

PostGIS

PostGIS is an extension to PostgreSQL that allowed for the storage and manipulation of projected and unprojected geospatial data. This project was built against PostGIS 1.3.2.

Grass

Map overlays were generated using Grass 6.2.2 allowing for processing of data with R and the pending generation of interpolated maps.

ARC Map-Fu

ARC Map-Fu was used as a codebase in the development of the interactive mapping application. Map-Fu is built by the PSU ARC team using PHP, MapServer, and PostGIS and provides basic cartographic functionality, such as the ability to zoom on a map and query against specific layers.

Time-based Data Side

Data Retrieval Tool

The data merge and download tool was modified to accommodate more dynamic selection of data. Users were given the options of searching by "Platform," "Principal Investigator," "Sample Type," or through "All Data." Any selection made by the user triggers a Javascript event that uses the XMLHTTPRequest method (XHR) to request necessary data to filter the other items. Autocompletion facilities were added to each panel to ease in the selection of data. This functionality was provided by the YUI AutoComplete component.

Operations Manual for Advanced Database Resource for TexAQS 2006 data

Dean Atkinson and Monica Wright
Chemistry Department
Portland State University
Portland, OR 97207-0751

The database home page offers three main functionalities to the user and a contact button (Email Us) to follow up on any issues that the user may encounter. The user would select "Upload Data" to add data to the database. This function is generally only performed by Principal Investigators (PIs) or their representatives. More details on using the file upload utilities are available on the website under "Help" and are included here. "Analyze Data" is used to search for data and to access the data merge, download, and visualization and analysis portions of the data side of the database. The ICARTT-Formatted File Download link may be used by simply selecting a PI/Organization from the drop-down box and "Get PI/Organization Info" and then by selecting a date and finally the "Download" link. Most users would be likely to use the "Analyze Data" or "Interactive Map" links from the home page. The "Interactive Map" link leads to the map site, which is presently a separate site from the data side. Opening the map in a separate browser tab is the simplest method for retaining ready access to both sides of the database. The three user-active links are described in greater detail below.

"Upload Data" link (for PIs or representatives)

ICARTT Formatted Data upload

If a fully ICARTT-compatible file is used for upload (typically with the ICARTT file extension *.ict) the procedure is simple. The user just selects the "ICARTT Formatted Data" link, then the "Browse" link and navigates to the directory and file on their computer. After the file is selected and the "upload" button is pressed, a dialog reports on the status of the file upload. If the upload is successful, no further action is required. Note that the FScan utility available at the ICARTT website at http://www-air.larc.nasa.gov/missions/etc/IcarttDataFormat.htm can be used to confirm that a given file is formatted properly and details on the format are also available.

Delimited Data Upload Help

After selecting the "Upload Data" and "Delimited Data" links the user selects a PI/Organization from a drop-down menu, or selects the top option to add a new one. The "Browse" link is then selected and the user navigates to the file folder and file that they want to upload. The delimiter (tab or comma) is also selected in a drop-down menu box and finally the "upload" link is selected.

Note: When uploading data that is to be associated with a PI and Organization that does not already exist in the application's database, you will be prompted to construct an ICARTT-type header that will be included with your data. You will only need to do this once, so it is worth making it fairly close to conformance with the ICARTT format. The information will be useful to other data users and will eventually allow you to output your data in the full ICARTT format.

This document contains basic information about the required fields. If you are unsure about what to enter in each field, please consult the ICARTT format page: http://www-air.larc.nasa.gov/missions/etc/IcarttDataFormat.htm

1. **Format for Time Variables:** Choose the correct format for the time variable in the first column of your data file. The first choice is a date/time stamp that is recognizable by UNIX, the second choice is the number of seconds from some date and time that you will be asked to provide.
2. **Data ID:** A short descriptor for the type of data (or the instrument used) contained in the file [e.g., CRTDN, BasicMet]. This short string and the location ID are combined to form an identifier in this database and are concatenated with a date string to create the ICARTT filename.
3. **Location ID:** Location or platform of measurement [e.g., UHMT (moody tower), rhb (Ron Brown), np3 (NOAA P3)].
4. **Time Zone Confirmation:** *You must confirm the time zone that the time column in the data (the first column) refers to and check the confirmation box before you can upload the file.*
5. **PI name:** last name, first name/initial
6. **Organization/affiliation of PI:** department or unit/organization [e.g., Chemistry Department/Portland State University]
7. **Data source description:** Text description of instrument name; platform name; model name and number, etc. [e.g., Various - TECO NOx & Ozone, 3-axis Sonic Anemometer, Capacitive Humidity Sensor, NTC Thermistor]
8. **Mission name:** Study name, platform name, year [e.g., TexAQS II TRAMP Moody Tower Campaign 2006]
9. **UTC date when data begin; UTC date of data reduction or revision:** Date of first entry in data file, space, date when last revision was completed (probably today's date) [e.g., 2006 08 26 2007 08 05]
10. **Data Interval:** This value describes the approximate time spacing (in seconds) between consecutive data records. It is the (essentially constant) interval between values of the independent variable. [For 1 Hz data the data interval value is 1; for 1 minute data the value is 60; for 2 Hz data the value is 0.5. If the data records include start AND stop times, then the data interval value is 0, since the data interval is clearly established.]
11. **ULOD (Upper Limit of Detection) Flag:** A value that indicates data have exceeded the ULOD [usually -7777].
12. **LLOD (Lower Limit of Detection) Flag:** A value that indicates data have exceeded the LLOD [usually -8888].
13. **Variables:** The independent and dependent variable names are populated using the labels from the first line of the uploaded file. In most cases, you will not need to edit the label or unit for the independent variable. You should fill in the units for the dependent variables and the other fields as follows:
    - **Variable (names):** Name or description of data in that column. These must match the column names in the file you are uploading. [e.g., ToFAMS]
    - **Units:** Units for given dependent (or independent) variable. [e.g., ppm]
    - **Is Time Unit:** Check the box if the data column is another time axis. [e.g., MidTime, EndTime]

- **Scale Factor:** Use to apply a multiplicative scaling to all of the data – for example if a recalibration is needed [typically 1].
- **Missing data indicator:** A value that indicates data points are missing for some reason [usually -9999].
- **Requires Wind Speed Variable:** Used for averaging of wind direction (low wind speeds are given less weight in the average) – indicate the Wind Speed column, usually by copy and pasting the label. [e.g., WindSpeed].
- **Special comments:** Notes of problems or special circumstances unique to this file. An example would be comments/problems associated with a particular flight or cruise.
- **Normal comments:** *You will need to check a box indicating that you have inserted the correct information in this section.* You will not be able to easily modify this information, so you may want to use the list below to create a text document that can be pasted into the space provided. This is the place for investigators to more completely describe the data and measurement parameters. Note the **non-optional** information regarding uncertainty, the upper limit of detection (ULOD) and the lower limit of detection (LLOD) for each measured variable. The ULOD and LLOD are the values, in the same units as the measurements that correspond to the flags –7777 and –8888 within the data, respectively. The last line of this section should contain all the variable names on one line.

WHAT TO INCLUDE IN THE NORMAL COMMENT SECTION:

PI_CONTACT_INFO: Phone number, mailing address, email address and/or fax number.
PLATFORM: Platform or site information.
LOCATION: including lat/lon/elev if applicable.
ASSOCIATED_DATA: File names with associated data: location data, aircraft parameters, ship data, etc.
INSTRUMENT_INFO: Instrument description, sampling technique and peculiarities, literature references, etc.
DATA_INFO: Units and other information regarding data manipulation.
UNCERTAINTY: Uncertainty information, whether a constant value or function, if the uncertainty is not given as separate variables.
ULOD_FLAG: -7777 (Upper LOD flag, always -7's).
ULOD_VALUE: Upper LOD value (or function) corresponding to the -7777's flag in the data records.
LLOD_FLAG: -8888 (Lower LOD flag, always -8's).
LLOD_VALUE: Lower LOD value (or function) corresponding to the -8888 flag in the data records.
DM_CONTACT_INFO: Name, affiliation, phone number, mailing address, email and/or fax number.
PROJECT_INFO: Study start & stop dates, web links, etc.
STIPULATIONS_ON_USE: (self explanatory)
OTHER_COMMENTS: Any other relevant information.

REVISION: R#
R#: comments specific to this data revision. The revision numbers and the associated comments are cumulative in the data file. This is required in order to track the changes that have occurred to the data over time. Prepend the information to this section so that the latest revision number and comments always start this part of the header information. Note that FIELD data files have revision LETTERS, not numbers.
Indep_Var VarName_1 VarName_2 VarName_3 … VarName_n

"Analyze Data" link - Data Merging and Analysis

To merge and/or analyze data from the data side of the database, select the "Analyze Data" and "Merge and Analyze Data" menu links. This will direct you to the Data Merge and Analysis page.

Choose the data sets you would like to merge:

1. You may choose to display the available datasets by PI, Organization, or collection or you can narrow the search using the Subtype option.
2. Highlight one selection at a time and click the **Add ->** button.
3. You may make multiple selections by repeating the step above for each PI, Organization, or Table you want to analyze.
4. Once all of the selections you are interested in have been moved to the right hand side of the page, click on the **Get Variables** button.

   This will direct you to a new Data Analysis page that contains all of the collections that were selected or were applicable to the PI or organization you selected.

   Each Collection has a blue highlighted table name. Clicking this name will pull up the ICARTT header information, total number of data points in the table, and the first and last dates that contain data. This information should be used when determining appropriate merge dates, average intervals, and minimum sample sizes.

   Select the variables for analysis

   1. Check the box(es) next to the variable(s) of interest.
   2. For each table specify the minimum number of points required to calculate a mean. {The default is one, which obviously makes a standard deviation calculation impossible.} *For 10 minute averages of one minute data the user can indicate that only intervals with 8 points should be used to calculate an average, thus avoiding unreliable "edges" in datasets*.

   Select the Interval Start and End

   3. Use the drop down menu to select the inclusive endpoints for data merging and averaging. These times are in Central Daylight Time [UTC - 5 hours]. If you

want to select data from multiple non-contiguous time segments, use the “Number of Timespans” drop-box to select the number of segments.

Select the Interval Size

4. Specify the number of seconds to be used for averaging the data. *(The default is 120 seconds which will give the user 2 minute averages of the data over the time span selected.)*
5. Specify the number of data points to display for user evaluation before download.
6. To display standard deviation and count for the averaging intervals, make sure the boxes are checked next to each option.

Compute Interval Averages

7. When the above options are correct, click on the **Compute Interval Data Averages** button.
8. Formatted Output Options and a preview of the averaged data will be displayed under the Compute Interval Data Averages button. You may use the Beginning, Previous, Next and End buttons to preview additional data.

Formatted Output Download Options

- Create Results File
    1. Select the desired delimited format (Comma or Tab).
    2. Indicate if column headers are desired.
    3. Select **Create Results File**
    4. A pop up window should appear.
    5. Name the results file in the space provided.
    6. Use the check boxes to indicate if standard deviation and/or counts should be included.
    7. If desired, data can be removed, column names can be changed and multiplicative conversion factors can be applied.
    8. When all the options are correct, select the **Create File** button

- Create Graph

  * The default maximum memory for this option is 32 MB. If the data selected for graphing exceeds this limit, a blank screen will appear instead of a graph after **Create Graph** is selected in step 9. If this occurs, the graphing request should be modified to a shorter time span, larger averaging interval, or fewer variables. You may have to log off and re-enter the database to restore proper function after exceeding the maximum limit.*

    1. Select the **Create Graph** button

2. A pop up window should appear.
3. Choose a type of plot.
4. Enter the graph title (Optional).
5. Enter the Y-axis label (Optional).
6. Specify graph width and height.
7. Check boxes to indicate legend and color preference.
8. Choose the data to include in the graph and enter any necessary conversion factors.
9. Once the correct options are selected, click on the **Create Graph** button.
10. The graph will appear in a new pop up window.

Note: These are relatively low quality graphs intended for "first look" comparisons. You may use the Create Results File option to download data to construct high quality data presentations in other software packages.

- Create file for using R to create a wind rose

"Interactive Map" link – spatial analysis

Selecting (or selecting "open in a new tab") the "Interactive Map" link opens a new window that displays the base level map of the Houston-Galveston metro area. In the upper right of the map is a set of tools links that are used to operate the spatial visualization and analysis functions.

- The "Layers" link is used to open a flyout that lists all of the available layers. The first two are all of the available Base layers and a few Points of Interest layers. Clicking on any of the check boxes activates the layer and unchecking it removes the layer. Depending on server and connection speed, the results of clicking on a layer may take anywhere from 30 seconds to several minutes to become visible.
- The remainder of the layers in the "Layers" flyout are data layers, sub-categorized by collection. Clicking on the plus sign (+) next to each collection opens the collection and clicking on the plus sign next to each data set produces an overview, including a statistical summary, of the data in the set. Checking the box next to the data set displays the data layer. For the spatially invariant data, this results in a single circle on the map at the location of the monitoring site.
- The "Tools" link opens a flyout menu that contains links for zooming and panning similar to those used in most map applications. The last link "Draw" is used to select data by superimposing circles or polygons on the map over the selected data. The polygon tool adds a node every time the left mouse button is clicked and terminates the polygon when the left button is double-clicked.
- The "Overview" link provides a small inset map showing the study area. This is useful when navigating at small or large scales.
- The "Info" tab is used to activate a flyout that contains statistical information about a data selection. Links below the summary activate the three plots (QQ, Histogram, and Line) and the shapefile download tool.
- The "Legend" link opens the data legend.

- The “Zoombar” link opens the zoombar tool that can be used to more rapidly switch from fine and broad levels of zoom on the map.

Beginning users should try activating a data layer or two and then selecting it with a circle (Draw menu). They can then activate the Info link and look at the different plot types.