Power to the People: Public Access and Reporting in the Texas Air Emissions Repository (TexAER)

Grace Kitzmiller, William Gerber, Jason Veysey, Bob Jacoby
Eastern Research Group, Inc., 14555 Avion Parkway, Suite 200, Chantilly, VA 20151-1102
Grace.Kitzmiller@erg.com

Contributor
Julie Westphal
Texas Commission on Environmental Quality, 12100 Park 35 Circle, Austin, TX 78753
jwestpha@tceq.state.tx.us

ABSTRACT

To support its state implementation plan (SIP) and EPA reporting processes, Texas must collect, develop, and integrate emissions data from numerous sources for 254 counties. The Texas Commission on Environmental Quality (TCEQ) and Eastern Research Group (ERG) have met this challenge with the Texas Air Emissions Repository (TexAER): a centralized, Web-based system that stores, reconciles, and reports nonpoint emissions data. With tools for uploading raw emissions estimates, collating and enhancing data, and extracting inventories in various forms, TexAER provides a single, flexible platform for managing nonpoint information.

Through its publicly available reporting functions, TexAER provides public access to emissions data and inventories for the state of Texas. Interested citizens can log on to TexAER at http://www4.tceq.state.tx.us/texaer/ and explore its dynamic, user-driven reporting system. The Emissions Data Report allows queries against any emissions inventory in the TexAER repository. Users can focus the content of their report by adding filters—on source classification code (SCC), pollutant, geographic location, and more—and can choose how to summarize, sort, index, and title the report's output. Reports can be run to screen, saved, or exported for offline analysis. At all points the user controls the slices of data presented. Users can also view and apply growth factors and control strategies to baseline emissions inventories.

Underlying the TexAER reporting system is a high-performance data warehouse. TexAER takes advantage of the latest data warehousing techniques and technologies to provide rapid, on-line reports that query and summarize millions of inventory records for reports with good response time.

In addition to the core Emissions Data Report, TexAER lets public users browse the SCC hierarchy, as well as the Document Catalog, a collection of reference documentation for emissions inventories. Together these functions provide significant access to TexAER data, opening TCEQ's nonpoint repository to the general public.

INTRODUCTION

The Texas Air Emissions Repository (TexAER), formerly the State Implementation Plan Emissions Data Management System (SIP EDMS), is a web-based computer system that TCEQ uses to store, secure, and access nonpoint (or, more specifically, area, non-road mobile, on-road mobile, and biogenic) emissions data. As a central, reliable repository of nonpoint air emissions information, TexAER is an important part of TCEQ's emissions data infrastructure. TCEQ requires a secure, central repository in order to maintain the integrity and reproducibility of nonpoint source data to address the large increases in the volume of data managed by TCEQ staff in recent years, and to meet data management requirements associated with the promulgation of the federal Consolidated Emissions Reporting Rule (CERR). Reliability and accessibility of emissions data is essential for the development

of state implementation plans (SIPs). TexAER also provides the general public with access to air emissions inventory data.

TCEQ uses TexAER to automate a number of data processing functions to quickly, and with confidence, provide the standardized emissions data necessary for the development of SIPs, the assessment of control strategies, and to the public and other governmental entities. In addition, TexAER allows TCEQ to make emissions and modeling data available to local entities in non-attainment and near non-attainment areas to support their efforts to address local air quality strategies. The full set of TexAER functionality is available through TCEQ's Intranet. A subset of TexAER's functionality is available externally through the Internet. This paper describes the reporting interface available to the general public, and details the implementation of a data warehouse to support TexAER's on-demand reporting.

BODY

Accessing TexAER

TexAER is available to the general public at http://www4.tceq.state.tx.us/texaer/. Figure 1 presents the TexAER login page.

Figure 1. TexAER Login Page

To begin using TexAER, users must first request a user account from the TexAER home page by entering a valid email address; optional information can be provided at the discretion of the user. After submitting the information, TexAER creates a username and password and forwards it to the email address provided by the user. After receiving login information, users can access TexAER and:

- Create detailed emissions reports;
- Create summarized emissions reports;
- Apply control strategies to emissions data;
- Project future-case emissions inventories;
- Search and retrieve information about control programs and techniques applicable to nonpoint sources; and
- Review and download supporting documentation associated with the emissions inventory data.

TexAER Reporting

TexAER's reporting functionality provides users a number of options to view and report on emissions inventory and other related data. Users can access all of the TexAER report-related functionality by clicking Reports in the top navigation menu. The Reports home page is shown in Figure 2.

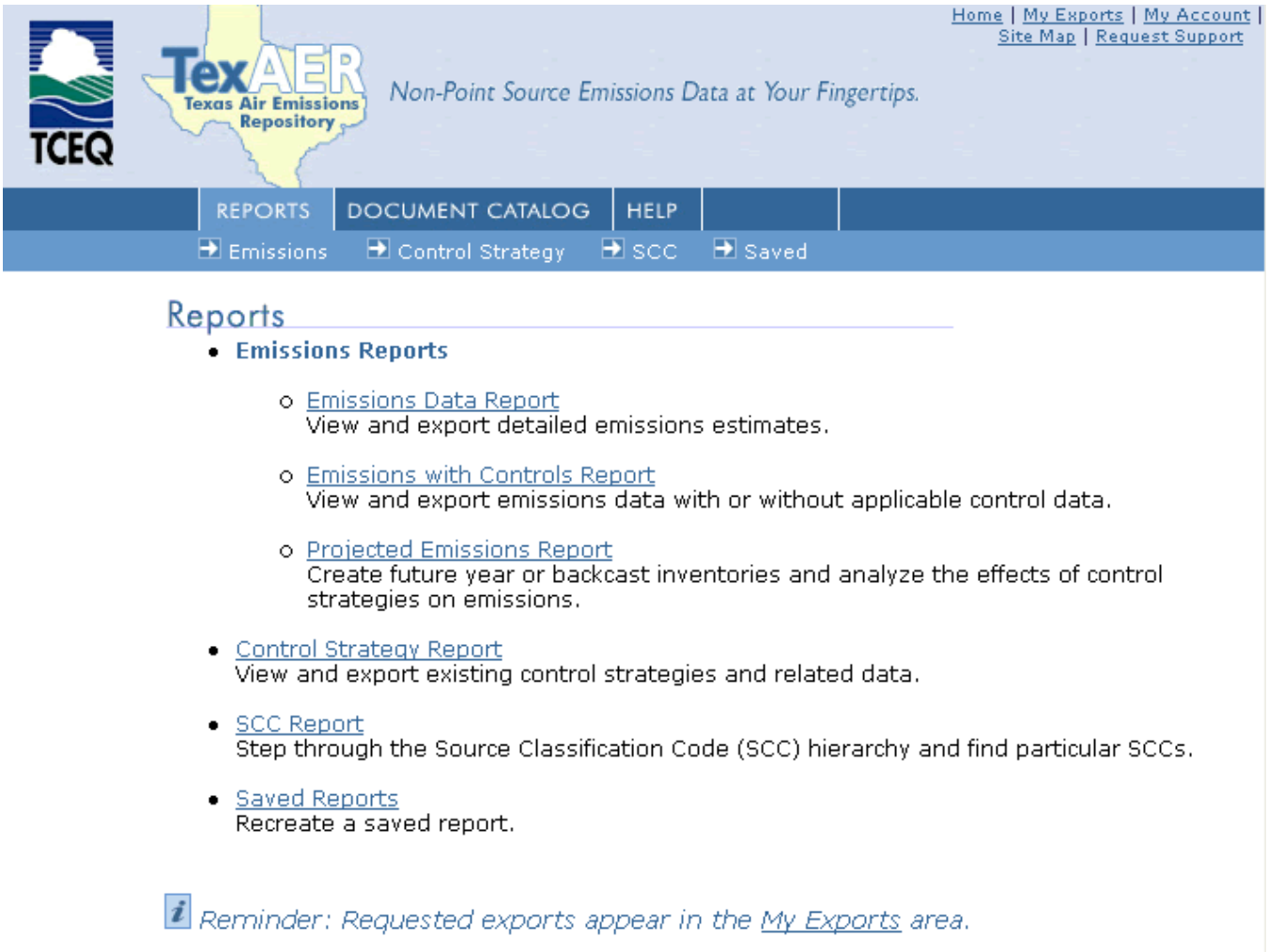


Figure 2. TexAER Reports Home Page

The available reporting options include:

- Emissions Data Reports: Present emissions data for individual or summarized emissions estimates. There are two Emissions Data report types: Summary and Detail.

- Emissions with Controls Reports: Present emissions and related NIF control information exactly as stored in the database, as well as information about potentially applicable controls.

- Projected Emissions Reports: Calculate and present estimated future emissions based on a target year, set of growth factors and emissions controls. There are seven Projected Emissions Report types: Base Detail, Base Summary, Projected Detail, Projected Summary, Delta Detail, Delta Summary, and Base Emissions with Controls.

- Control Strategy Reports: Present information about proposed, hypothetical, and adopted control programs, control strategies, and control techniques stored in the database. These controls are not necessarily linked to an inventory, subinventory, or emissions estimate.

- SCC Reports: Present a view of selected SCCs, where additional SCCs can be viewed by drilling-down through the SCC tree or searching for an SCC.

Public users can also save reports, access shared reports, and export several types of reports using TexAER.

Report Customization Options -- Filter Criteria

Public users can apply a variety of customization options when creating TexAER reports. The actual customization options available depend on which report is being generated and include:

- Emissions Inventory – Emissions inventory data generated by TCEQ, Councils of Governments (COG), and other organizations are uploaded into TexAER by TCEQ. Emissions inventories may include data for one or more counties, groups of SCCs, the entire state, or other subsets of information. These inventories provide emissions estimates for nonpoint sources.

- Emissions Subinventory (i.e., source type) – Emissions subinventories are a subset of a larger inventory. Area, Non-Road Mobile, On-Road Mobile, and Biogenic are all examples of the emissions subinventories in TexAER.

- Location Groups – Location Groups are a set of one or more geographical locations that are used for reporting or analytical purposes. For example, the list of counties that comprise a near non-attainment area, TCEQ Regions, Regional Planning Authority (RPA), and the Council of Governments (COG) are all Location Groups.

- Locations – Locations specify geographic areas or points associated with emission information. County, facility, and offshore are all examples of Locations.

- Pollutant Types – Pollutant Types represent a group of pollutants. For example, criteria, HAPs, greenhouse gases, ozone precursors, and VOCs are Pollutant Types.

- Pollutants – Pollutants are compounds releases to the air and are reported in emission inventories. For example, SO_2, NOx, CO, and ozone are Pollutants.

- Time Period – Time Periods are associated with emissions inventory data. Emissions data may be collected over certain time periods and in various units (e.g., annually or daily in tons per year or tons per day).

- SCC Class – A source classification code (SCC) class is a group of related SCCs. Aircraft, Gasoline Engines, Lawn and Garden Equipment, and Off-road equipment are all examples of SCC classes.

- SCCs – Source Classification Codes (SCCs) are 8- or 10-digit codes used by EPA to categorize individual processes or unit operations that generate air emissions. Each SCC represents a unique process or function within a source category that is associated with a specific air pollution emission event. One or more SCCs can also represent any operation that causes air pollution. SCCs are used to identify and classify processes and operations to prepare emissions estimates. For example, an SCC may correspond to a gas station, particular boiler type, a process heater, or a reactor vent.

- Control Data – Control data are NIF reported control information applicable to emissions estimates reported for an inventory.

- Control Strategies – Control strategies are legally applicable emissions reductions requirements for a particular emissions record in a selected subinventory. Users can filter on strategies that legally cover one or more of the estimates in a selected subinventory (excluding any other filter criteria). TexAER determines the list of strategies from which to choose based on inventory year, location, SCC, and pollutant combinations.

- Authority – Authorities are the entities that developed the control programs and strategies.

- Control Status – Control status indicates whether an emissions record was reported as controlled or uncontrolled in NIF.

- Control Techniques – Control techniques are specific equipment or actions that impact emissions rates.

Emissions Data Reports

The TexAER Emissions Data Report presents public users with emissions data exactly as stored in the database. After choosing an inventory and subinventory for a report, users can select from a variety of options to customize, or filter, the report to a subset of the available emissions data, including: location, location group, pollutant, pollutant type, SCC, SCC class, time period, and control status. The two types of Emissions Data Reports that public users can generate for the emissions data that match selected filtering options include:

- Emissions Summary Report: Displays emissions estimates summarized by location, SCC, or SCC class. The information displayed on the report depends on the type of summarization that selected, but always includes: pollutant, time period, unit of measure, and summed emissions estimate.

- Emissions Detail Report: Displays emissions estimates at the lowest level of detail for an inventory and subinventory selected by the user, including: location, SCC, pollutant, time period, unit of measure, and emissions estimate. Each row in a Detail Report

displayed by TexAER includes a Show Detail option. Figure 3 presents a sample Detail Report display.

— *Show Detail: History Display*

The History display presents a summary of any edits made to emissions estimates for the pollutant and time periods shown in the selected row of the Emissions Detail Report. The history includes the emission estimate (Qty), the name of the user who made a change, the reason for the change, and the date the change was made. Figure 4 presents the Show History Display.

— *Show Detail: Origins Display*

The Origins display presents a summary of the sources of the emissions estimate for each pollutant and time period in the selected row of the Emissions Detail Report. The origin information includes the inventory name, inventory year, the number of generations, and whether the inventory is active.

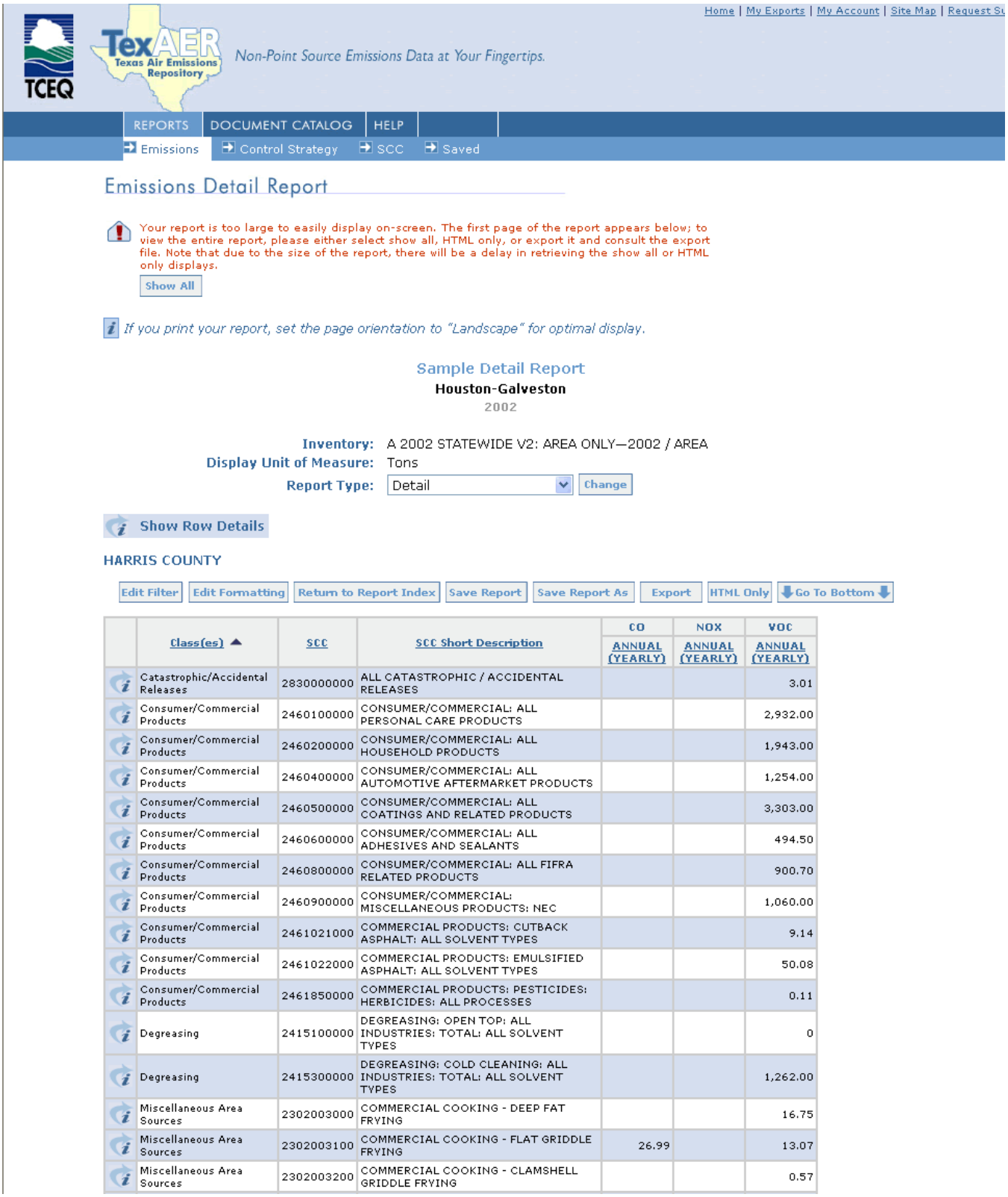
Emissions With Controls Report

Public users can generate, view, and export Emissions with Controls Reports using the TexAER reporting interface. The Emissions with Controls Report presents users with emissions and related NIF control information exactly as stored in the database, as well as information about potentially applicable controls. After choosing an inventory and subinventory for the report, users can select from a variety of options to customize, or filter, the report to a subset of the available emissions data, including: location, location group, pollutant, pollutant type, SCC, SCC Class, time period, and control status. The Emissions with Controls Report (Detail only) displays emissions estimates at the lowest level of detail for the selected inventory and subinventory, including any control data associated with the emissions estimate, such as rule penetration, rule effectiveness, and total capture control efficiency, as well as potentially applicable controls based on the inventory year, location, SCC, and pollutant being viewed.

Projected Emissions Reports

Public users can generate, view, and save Projected Emissions Reports using the TexAER reporting interface. Projected Emissions Reports present users with estimated future emissions based on a set of growth factors and any selected controls. After choosing an inventory and subinventory for a report, users can select from a variety of options to customize, or filter, the report to a subset of the available emissions data, including: location, location group, pollutant, pollutant type, SCC, SCC class, time period, and control status. In addition, to create projected emissions reports, the user must also choose a target year and growth factor data set, and can select from a variety of potentially applicable control strategies. The growth factor data sets available in TexAER were extracted from EGAS (http://www.epa.gov/ttn/chief/emch/projection/index.html). The two primary Projected Emissions Reports that users can generate for the emissions data that match the user's filtering selections include:

- Projected Emissions Detail Report: Displays projected emissions estimates at the lowest level of detail for the inventory and subinventory selected by the user, including: location, SCC, pollutant, time period, unit of measure, and emissions estimate. The projected emissions values are calculated using the target year, growth factor data set, and potentially applicable emissions controls selected by the user. Users can request to make a Projected Emissions Report a new inventory in TexAER through the Make Permanent function.



Home | My Exports | My Account | Site Map | Request Su

TCEQ TexAER Texas Air Emissions Repository — Non-Point Source Emissions Data at Your Fingertips.

REPORTS | DOCUMENT CATALOG | HELP

Emissions | Control Strategy | SCC | Saved

Emissions Detail Report

Your report is too large to easily display on-screen. The first page of the report appears below; to view the entire report, please either select show all, HTML only, or export it and consult the export file. Note that due to the size of the report, there will be a delay in retrieving the show all or HTML only displays.

Show All

If you print your report, set the page orientation to "Landscape" for optimal display.

Sample Detail Report

Houston-Galveston

2002

Inventory: A 2002 STATEWIDE V2: AREA ONLY—2002 / AREA

Display Unit of Measure: Tons

Report Type: Detail Change

Show Row Details

HARRIS COUNTY

Edit Filter | Edit Formatting | Return to Report Index | Save Report | Save Report As | Export | HTML Only | Go To Bottom

Class(es)	SCC	SCC Short Description	CO ANNUAL (YEARLY)	NOX ANNUAL (YEARLY)	VOC ANNUAL (YEARLY)
Catastrophic/Accidental Releases	2830000000	ALL CATASTROPHIC / ACCIDENTAL RELEASES			3.01
Consumer/Commercial Products	2460100000	CONSUMER/COMMERCIAL: ALL PERSONAL CARE PRODUCTS			2,932.00
Consumer/Commercial Products	2460200000	CONSUMER/COMMERCIAL: ALL HOUSEHOLD PRODUCTS			1,943.00
Consumer/Commercial Products	2460400000	CONSUMER/COMMERCIAL: ALL AUTOMOTIVE AFTERMARKET PRODUCTS			1,254.00
Consumer/Commercial Products	2460500000	CONSUMER/COMMERCIAL: ALL COATINGS AND RELATED PRODUCTS			3,303.00
Consumer/Commercial Products	2460600000	CONSUMER/COMMERCIAL: ALL ADHESIVES AND SEALANTS			494.50
Consumer/Commercial Products	2460800000	CONSUMER/COMMERCIAL: ALL FIFRA RELATED PRODUCTS			900.70
Consumer/Commercial Products	2460900000	CONSUMER/COMMERCIAL: MISCELLANEOUS PRODUCTS: NEC			1,060.00
Consumer/Commercial Products	2461021000	COMMERCIAL PRODUCTS: CUTBACK ASPHALT: ALL SOLVENT TYPES			9.14
Consumer/Commercial Products	2461022000	COMMERCIAL PRODUCTS: EMULSIFIED ASPHALT: ALL SOLVENT TYPES			50.08
Consumer/Commercial Products	2461850000	COMMERCIAL PRODUCTS: PESTICIDES: HERBICIDES: ALL PROCESSES			0.11
Degreasing	2415100000	DEGREASING: OPEN TOP: ALL INDUSTRIES: TOTAL: ALL SOLVENT TYPES			0
Degreasing	2415300000	DEGREASING: COLD CLEANING: ALL INDUSTRIES: TOTAL: ALL SOLVENT TYPES			1,262.00
Miscellaneous Area Sources	2302003000	COMMERCIAL COOKING - DEEP FAT FRYING			16.75
Miscellaneous Area Sources	2302003100	COMMERCIAL COOKING - FLAT GRIDDLE FRYING	26.99		13.07
Miscellaneous Area Sources	2302003200	COMMERCIAL COOKING - CLAMSHELL GRIDDLE FRYING			0.57

Figure 3. Sample Detail Report

Figure 4. Show History Display

- Projected Emissions Summary Report: Displays projected emissions estimates summarized by location, SCC, or SCC class. The information displayed on the report depends on the type of summarization selected by the user, but always includes: pollutant, time period, unit of measure, and summed emissions estimate. The projected emissions values are calculated using the target year, growth factor data set, and potentially applicable emissions controls selected by the user.

In addition, users can also create other projections-related reports, including:

- Base Detail: Displays emissions estimates for the base, or starting, year at the lowest level of detail for an inventory and subinventory selected by the user, including: location, SCC, pollutant, time period, unit of measure, and emissions estimate. The base year data represent the emissions estimates before the growth factor data set and potentially applicable emissions controls selected by the user are applied.

- Base Summary: Displays emissions estimates for the base, or starting, year summarized by location, SCC, or SCC class. The information displayed on the report depends on the type of summarization selected by the user, but always includes: pollutant, time period, unit of measure, and summed emissions estimate. The base year data represent the emissions estimates before the growth factor data set and potentially applicable emissions controls selected by the user are applied.

- Delta Detail: Provides the user with a view of how the selected controls applied during the projections calculations impacted emissions by displaying emissions estimates for the base inventory year, the projected emissions estimates for the target year with only base inventory controls applied, the projected emissions estimates for the target year with new controls applied, and the difference between the controlled base and projected emissions estimates. The emissions estimates are presented at the lowest level of detail for the inventory and subinventory selected by the user, including: location, SCC, pollutant, time period, unit of measure, and emissions estimate.

- Delta Summary: Provides the user with a view of how the controls you selected to apply during the projections calculations impacted emissions by displaying emissions estimates for the base inventory year, the projected emissions estimates for the target year with only base inventory controls applied, the projected emissions estimates for the target year with new controls applied, and the difference between the controlled base and projected emissions estimates summarized by location, SCC, or SCC class. The information that is displayed depends on the type of summarization selected by the user, but always includes: pollutant, time period, unit of measure, and summed emissions estimate.

- Base Emissions with Controls (Detail only): Displays a report that presents base, or starting, year emissions data at its lowest level of detail for an inventory and subinventory selected by the user, including any control data associated with the emissions estimate, such as rule penetration, rule effectiveness, and total capture control efficiency, as well as potentially applicable controls based on the inventory year, location, SCC, and pollutant being viewed. The base year data represent the emissions estimates before the growth factor data set and potentially applicable emissions controls selected by the user.

Control Strategy Data Report

Public users can generate, view, and export Control Strategy Reports using the TexAER reporting interface. These reports present users with information about proposed, hypothetical, and adopted control strategies, control programs, and control techniques stored in the database. These controls are not necessarily linked to an inventory, subinventory, or emissions estimate. Control strategies are legally applicable emissions reductions requirements for specific combinations of SCCs and locations and/or SCCs and pollutants. Control programs are groups of control strategies that have been proposed or adopted as an approach to meeting emissions reductions goals. Control techniques are specific equipment or actions that impact emissions rates. The control strategies in Tex AER were collected from a variety of resources, including Federal, State of Texas, and local Texas resources.

Users can search for and customize, or filter, the report using a subset of the available controls information, including: authority, control program, control technique, location, location group, pollutant, pollutant type, SCC, and SCC Class. The TexAER Control Strategy Report and related details that users can generate for the controls that match filtering criteria selected by the user include:

- The Control Strategy Report: Displays summary information for all of the control strategies that meet the filter criteria selected by the user, including:

  — Name
  — Description
  — Authority
  — Status
  — Effective Date
  — Percent Reduction
  — Control Program
  — Control Technique
  — Location Group
  — Location
  — Pollutant Type
  — Pollutant
  — SCC Class
  — SCC

  A sample Control Strategy Report is presented in Figure 5. This report also provides options for the user to view additional detail for each Control Strategy.

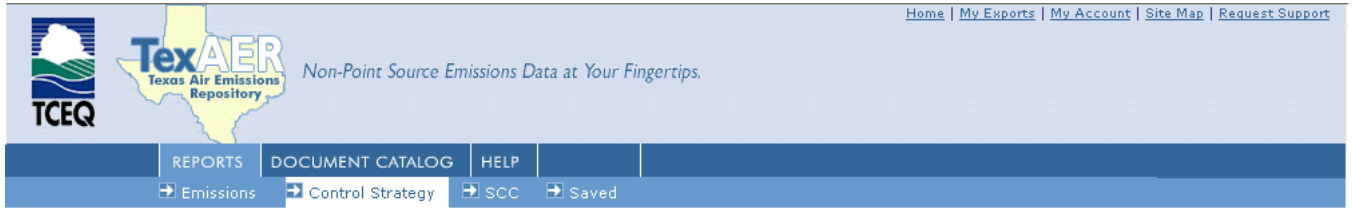
- Control Strategy Detail Displays: Present additional information about the locations, SCCs, pollutants, control techniques, and control programs associated with the control strategies selected by the user. The additional detail include:

  — Locations and associated rule penetration (RP) values for the Control Strategy.

  — Control Techniques that could be used to comply with the strategy, including technique name, SCC, SCC short description, pollutant, control efficiency (CE), and cost per ton.

  — SCCs and pollutants associated with the Control Strategy, including the associated effectiveness (RE), RE Method, and total capture control efficiency (TCCE).

  — Programs associated with the Control Strategy, including the program name, adoption date, begin date, end date, description, and TCCE.



Control Strategy Report

Options for Detail Information View	
☑ Locations	☑ Techniques
☑ SCCs and Pollutants	☑ Programs

1 2 3 4 5 6 7 8 9 Next 25

Select All | Clear All | Show Detail | Export | HTML Only | Refine Search | New Search

Search Results

Show Details	Name ▲	Description	Authority	Status	RP	RE	TCCE	Effective Date	Compliance Date	Expiration Date
☐	Abrasive Blasting - General	Dry Abrasive Blasting Applies to any abrasive cleaning operation; controls emissions statewide by a fabric filter	TEXAS STATE REGULATION	Adopted	1	0.80	0.99	08/09/2000	09/04/2000	
☐	Abrasive Blasting - Water Storage Tanks	Abrasive Blasting of Water Storage Tanks The operation of cleaning or preparing a surface by forcibly propelling a stream of abrasive material against the surface; controls emissions statewide by a shroud or vacuum device equipped with high effici...	TEXAS STATE REGULATION	Adopted	1	1	0.95	09/20/1991	11/01/1991	
☐	Agreed Order - Northeast Texas FAR	http://www.tceq.state.tx.us/implementation/air/sip/may1999net.html	AGREED ORDER	Adopted	1	1	1	05/01/1999	05/01/2003	
☐	Agreed Order - Northeast Texas FAR	http://www.tceq.state.tx.us/implementation/air/sip/may1999net.html	AGREED ORDER	Adopted	1	1	1	05/12/1999	05/12/1999	
☐	Agreed Order-Harris County SOx controls	http://www.tceq.state.tx.us/assets/public/implementation/air/sip/sipdocs/1997-05-HAR/96181sip_ado.pdf	AGREED ORDER	Adopted	1	1	1	08/01/1997	08/01/1997	
☐	Airport Ground Support Equip - DFW	Dallas/Fort Worth GSE agreement GSEs emissions are to be reduced by 90% by 2007. Applicability: GSE used by American and Delta airlines operated in the Dallas/Fort Worth area (at DFW airports)	LOCAL ORDINANCE	Adopted	0.80	1	0.90	04/19/2000	04/19/2000	12/31/2007

Figure 5. Sample Control Strategy Report

SCC Reports

Source classification codes (SCCs) are 8- or 10-digit codes used by EPA to categorize individual processes or unit operations that generate air emissions. Each SCC represents a unique process or function within a source category that is associated with a specific air pollution emission event. One or more SCCs can also represent any operation that causes air pollution. SCCs are used to identify and classify processes and operations to prepare accurate emissions estimates. For example, a code may correspond to a gas station, particular boiler type, a process heater, a reactor vent, etc. The SCC Report provides users with a view of selected SCCs. Users can view SCCs by drilling-down through the SCC tree or searching for an SCC.

Saved Reports

Users can save filter criteria and formatting options for TexAER Emissions Data Reports, Emissions with Controls Reports, and Projected Emissions Reports. The user enters report metadata and accesses an option to save the report. The TexAER Saved Report functionality allows users to:

- Search for saved reports;
- View, modify, and retrieve saved report criteria and corresponding Emissions Data Reports, Emissions with Controls Reports, and Projected Emissions Reports;
- Access information describing whether underlying emission data have changed since the report was saved;
- Share saved reports; and
- Delete saved reports.

Export Reports

Users can export many TexAER reports in one of the following three formats: comma-delimited, tab-delimited, and NIF 3.0. Users make format selection for the export, enter additional metadata, and submit the export. Exports are created off-line. Users can access reports and other files created for export from any page in TexAER. Figure 6 presents a sample display of the My Exports.

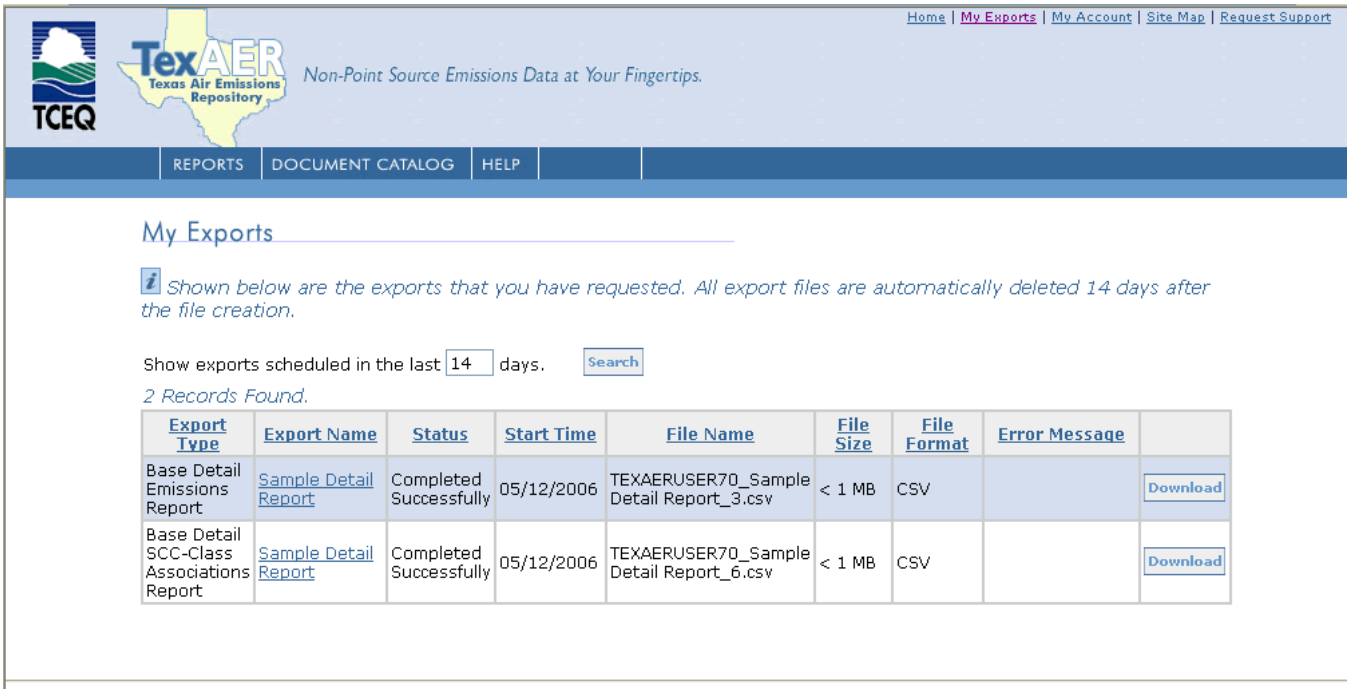


Figure 6. Sample My Exports Page

NIF Export

The TexAER NIF Export Tool exports emissions data in version 3.0 of EPA's National Emissions Inventory (NEI) data gathering and Input Format (NIF 3.0). EPA requires this format for submitting air emissions inventory data to meet reporting requirements. You can find comprehensive documentation of the NIF 3.0 format and required data elements for each inventory type at http://www.epa.gov/ttn/chief/nif.

TexAER Data Warehouse

One of the key features of TexAER is on-demand reporting. Users can select a report type and apply a wide variety of filtering and formatting criteria; TexAER then dynamically generates the report. By early 2005, the TexAER database had grown to include approximately 130 tables with over 29 million records and 7 gigabytes of data. In addition, it was estimated that TexAER would grow to approximately 35 gigabytes within three years. The database is currently approximately 26 gigabytes. Preliminary requirements analysis and database design activities for TexAER had been completed by one contractor, and requirements refinement and implementation was being completed by another contractor. The inherited database model was transactional in nature and optimized for frequent data insertion, updates, and deletions rather than to support reporting.

As the work on TexAER progressed data were updated on a regular basis; however large volume updates on an hourly or even daily basis were not expected to be typical. On the other hand, the type, number, and complexity of TexAER reports had grown significantly. Given the amount of data accessed and extracted by TexAER to create a report, the transactional database model was impacting performance. Several Oracle database optimization techniques were applied to address the performance issues, including the use of additional indexes, materialized views, and SQL hints. Although these techniques improved performance, report generation slowed to several minutes for some of the most complicated sets of reporting criteria.

In order to meet the goals of on-demand reporting, the application of dimensional modeling and data warehousing techniques in additional to the TexAER transactional structure was investigated. A data warehouse is "a copy of transaction data specifically structured for querying and reporting." (Kimball). Well-designed data warehouses provide for significantly improved performance of large queries such as those used to generate TexAER reports, improved scaling, easier maintenance and archiving of data (due to use of partitioned tables), and a standardized design that results in a short learning curve and streamlined use of third party reporting tools.

In a data warehouse, data are modeled as tables that contain facts and dimensions**.** *Fact tables* typically include quantitative information. One of the primary facts tables in TexAER stores air emissions values. *Dimension tables* typically contain information that provide context, or metadata, for the facts. In TexAER, examples of the dimensions associated with the emissions facts include pollutant, SCCs, location, time period, inventory, and subinventory/source type. Use of *aggregate tables* that summarize data can also speed performance. Aggregate tables that may be applicable to TexAER in the future could include summarizing hourly emissions to daily values.

One optimal implementation of a data warehouse results in a star schema, which can be visualized by picturing the fact table surrounded by its many dimensions. To achieve a star schema, the fact table is joined to its many dimensions through careful creation and implementation of primary and foreign key relationships. Several good internet resources on dimensional modeling and data warehousing can be found at:

- Kimball University: http://www.kimballuniversity.com/
- The Data Warehousing Institute: http://www.dw-institute.com/
- Creative Data http://www.credata.com/research/dwsites.html

Many vendors offer commercial data warehousing products, including Oracle, SAS, IBM RedBrick, Hyperion, Business Objects, Microstrategy, and Cognos. For TexAER, TCEQ chose to implement the data warehouse using Oracle's built-in data warehouse support. This decision was based on the following:

- Oracle data warehouse functionality is native to the database package and does not require use of a third party tool, or another layer, that would overlay Oracle. In addition, because the data warehouse functionality is built into Oracle, it was already included in TCEQ's Oracle licenses and did not require additional software or licensing investments.

- Oracle supports a variety of indexes including bitmap and b*tree indexing. Traditional, or b*tree indexes can be thought of like the index at the back of a book. An index sorts the information in the book and indicates the pages (rows) on which a particular item can be found. Although useful, these types of indexes can require a large amount of storage space and are only useful when the query conditions exactly match the conditions used to create the index. B*tree indexes are not typically useful for finding records that relate two or more keywords at the same time. Bitmap indexes are 'smart' indexes and not found in all database management systems. These indexes use a series of ones and zeros (bits) to indicate where data are located (the map). The power of this approach is that the maps from multiple indexes are combined dynamically during a query to form one consolidated bitmap that is then used to quickly locate the rows that contain the data of interest. This improves performance by significantly reducing the time required to perform complex queries.

- Oracle supports star transformations. Star transformations allow the query optimizer to join the fact table to multiple dimensions that represent multiple report selection criteria into a

single composite before accessing the fact table. Oracle then accesses the fact table just once to retrieve the records required to create a report. Databases operating without star transformation may repeatedly access a fact-type table to meet each report criterion, requiring significantly more time to complete the operation.

- Managed parallelism. Oracle can take advantage of the database server hardware and disk array to perform multiple operations simultaneously. Oracle can take a complicated query and reporting operation, break it into discrete chunks, or multiple threads, and execute those threads simultaneously. Parallelism can be achieved even if the database is stored on one server.

- Oracle also provides native support for table partitioning. Partitioning a large table and breaking into smaller, separate physical structures speeds up query and reporting time because the code can access one smaller structure rather than an entire large table. Most databases require manual partitioning; the database administrator decides on a partitioning strategy, creates separate tables for each partition, and moves data to each partition. While Oracle allows for manual partitioning management, it can also automatically partition a table based on criteria that the database administrator specifies. Note that there is not really a limit to number of partitions Oracle can support. The TexAER fact table is partitioned using a combination of inventory and subinventory. TexAER has over 16 million emissions records; the average number of records in each partition is 181,000. Therefore, the typical maximum number of records accessed by any one query is 181,000, quickly speeding the reporting process.

Once the data warehouse structure is created, a process must be created to populate the data warehouse with the data from the transactional database. This process is typically broken into three steps: extraction of data, transformation of data, and loading of data (ETL). The ETL process can be implemented using a variety of tools and program languages. The TexAER ETL was written by the TexAER team using Oracle's native PL/SQL language and is executed on a nightly basis.

Lastly, optimization and tuning of the Oracle server settings for the data warehouse is critical to achieving the best performance. The primary settings are those for the memory and swap space requirements. These space requirements for a data warehouse are much larger then those for a traditional transactional database. While recommendations for the appropriate sizing can be found from Oracle and in reference books, the optimum levels can only truly be achieved through testing. Another important setting is the degree of parallelism which can be used for queries. This setting should be based on the available hardware and typical number of concurrent users for the system.

After completing implementation of the data warehouse in the TexAER database, an average 20 percent improvement in query performance was recognized. These performance improvements are also scalable. The performance of the data warehouse will be relatively stable whether an additional 16 million or 116 million emissions records are added.

To summarize, TexAER includes both a transactional database and a data warehouse. The transactional database accepts data insertions, updates, and deletions. These data changes are transferred to the data warehouse through the nightly refresh, or ETL process. Implementation of the data warehouse has reduced report generation and query times by 20 percent.

Next Steps

TexAER development efforts this year will focus on expanding the reporting functionality to implement a GIS interface and a graphing interface. The GIS interface will allow users to visually display selected emissions data by county. Approximately 20 layers will be available to overlay the emissions map, such as counties, attainment status, roadways, water bodies, airports, railroads, monitoring stations, and others. The TexAER graphical interface will allow users to graph pollutant emissions data for multiple SCCs, SCC classes, locations, location groups, and time periods.

CONCLUSIONS

TexAER provides an intuitive, easy-to-use interface that allows public users to create reports of nonpoint emissions data for any of the 254 counties in Texas. Users can generate detailed and summarized emissions reports, apply control strategies to emissions data, project future-case emissions inventories, search and retrieve information about control programs and techniques applicable to nonpoint sources, and review and download supporting documentation associated with the emissions inventory data. Reports can be generated for virtually any combination of emissions inventory, emissions subinventory (i.e., source type), location groups, locations, pollutant types, pollutants, time periods, SCC Classes, SCCs, and control status. Reports requested by the public using TexAER are generated on-demand.

In order to meet these real-time reporting requirements and provide reasonable performance, a data warehousing component that complements the TexAER transactional component was added to the database. The TexAER data warehouse significantly improved the performance of the large queries required to generate reports, improved database scalability, and provides for easier long-term maintenance and archiving of data. When working with a data set expected to approach one million records or more, data managers should consider a data warehousing approach to better support data analysis and reporting.

REFERENCES

TexAER - http://www4.tceq.state.tx.us/texaer

EPA's National Emissions Inventory (NEI) data gathering and Input Format (NIF 3.0) - htttp://www.epa.gov/ttn/chief/nif

Economic Growth Analysis System (EGAS) - http://www.epa.gov/ttn/chief/emch/projection/index.html

Kimball, Ralph. *The Data Warehouse Lifecycle Toolkit: Tools and Techniques for Designing, Developing, and Deploying Data Warehouses*: John Wiley & Sons, 1998

KEYWORDS

Inventory
Emissions
Nonpoint
Area
On-road mobile
Non-road mobile
Biogenic
Report
Data Warehouse