Design of an Inter-RPO Emissions Inventory Warehouse System

Megan Schuster
MARAMA, 711 West 40th Street, Suite 312, Baltimore, MD 21211
mschuster@marama.org

Grace Kitzmiller, Hui Zhou
Eastern Research Group, Inc., 14555 Avion Parkway, Suite 200, Chantilly, VA 20151-1102
Grace.Kitzmiller@erg.com

John Haasbeek
ENVIRON, 1616 Seventeenth Street, Suite 468, Denver, Colorado 80202
jhaasbeek@environcorp.com

ABSTRACT

The purpose of this paper is to discuss the design of an emissions inventory data warehouse system to manage, store and report emissions inventory data including ancillary modeling files. The system is web-based with the capability of performing versioning and limited quality control. The system will allow users from across the country to input and export data. This paper will also discuss challenges associated with the development and design of the system due to project requirements. This project is funded by all RPOs; guidance and decisions have been made by an Inter-RPO technical oversight committee.

INTRODUCTION

The Regional Planning Organizations (RPOs) were established in 1999 under the Regional Haze Rule to assist states in developing plans to improve visibility in Class I areas. In order to effectively address and curtail regional haze, each of the five RPOs has developed a 2002 baseline inventory for use in photochemical modeling. The inventories include point, area, nonroad and onroad sources of emissions and will be used in $PM_{2.5}$, Regional Haze and Ozone State Implementation Plans (SIPs).

Emissions inventories are associated with an inherent degree of uncertainty given the extensive use of models, limited field measurement data, and continuous improvement of estimating techniques. The data included in inventories are extremely detailed and updated on a continual basis. Each RPO expects to prepare several versions of their 2002 baseline inventory and several versions of future year inventories to reflect updates in estimation methodologies and other improvements. Due to the ever-changing nature of the inventory, detailed versioning is critical. When an inventory is updated, it is important for users of the data to be notified, and for documentation to exist. The RPOs have found it challenging to track and share the most up-to-date inventory versions for use in modeling.

In 2004, the RPOs began the design and development of a system to store inventory data, to track changes in each inventory, and to ensure that quality assured inventories are being shared across the country. The goal of this project is to develop the Inter-RPO Emissions Inventory Warehouse, a system to facilitate emissions modeling inventory sharing between RPOs, and to assist in ensuring shared data is quality assured across the country. Each RPO may assign one or more staff members to upload data, however many people from across the country are expected to download data on a continuous basis.

The system has to meet several criteria in order for it to succeed. These criteria include ensuring secure data transmittal and storage, straightforward upload and download procedures, and low

system operating and maintenance costs. Due to the system requirements and these specific criteria, several challenges and important decisions are being made during the design phase. This paper will discuss the system requirements, challenges, and solutions associated with the design of the system.

SYSTEM DESIGN

The Inter-RPO Emissions Inventory Warehouse web site will facilitate sharing of up-to-date and quality-assured emissions inventories and model input files between counties, states, RPOs, EPA, tribes, the modeling community, and other stakeholders such as the general public. The web site will be publicly available over the Internet to reach a large, geographically dispersed stakeholder community and will be able to receive, store, and output EPA-compliant and modeling emissions inventories (including model ancillary files), and perform versioning and limited quality control.

Currently the warehouse system is in the Design Phase. Development and implementation of the system are planned for the summer of 2005. A final warehouse system will be online by January 2006. During the Design phase, several challenges were identified and solved. The requirements and challenges of the system are described below, followed by the design solution and description.

Requirements and Challenges

The Inter-PRO Emission Inventory Warehouse will serve as a repository for a variety of inventory and ancillary modeling files. Table 1 provides a detailed listing of sources of data and types of files expected to be managed using the Inter-RPO Emission Inventory Warehouse. The File Type column describes the type of files that will accepted to the warehouse system (NIF, RPO DEP, etc.) and the File Listing column lists the more detailed tables associated with the File Types that will be accepted by the system.

Table 1. Inter-RPO Emissions Inventory Warehouse Data Sources

File Type	File Listing
Files in the RPO Data Exchange Protocol* format that are also consistent with EPA's National Emissions Inventory Format (NIF)	✓ Transmittal (TR) ✓ Emissions (EM) ✓ Emissions Process (EP) ✓ Emissions Period (PE) ✓ Control Equipment (CE) ✓ Site (SI) ✓ Emissions Unit (EU) ✓ Emissions Release Point (ER)
Mobile emissions files in the RPO Data Exchange Protocol* format	✓ Transmittal Mobile (MT) ✓ Network (MN) ✓ Attribute (MU) ✓ Attribute Lookup (MA) ✓ County Cross Reference (MX) ✓ Link coordinates (ML)
Continuous emissions monitoring (CEM) files in the RPO Data Exchange Protocol* format	✓ Continuous Emissions Monitoring (EM CEM) ✓ Measure Values (CEMEV)
Temporal allocation and profile assignment files in the RPO Data Exchange Protocol* format	✓ Temporal Source (TemporalTS) ✓ Temporal Monthly (TemporalTM) ✓ Temporal Daily (TemporalTD) ✓ Temporal Hourly (TemporalTH)
Spatial surrogate files in the RPO Data Exchange Protocol* format	✓ Grid Description (SpatialAG) ✓ Surrogate Description (SpatialAD) ✓ Spatial Surrogate (SpatialAS) ✓ Surrogate Ratio (SpatialAR)

File Type	File Listing
Chemical speciation files in the RPO Data Exchange Protocol* format	✓ Compound List (SpeciateCL) ✓ Profile Cross Reference (SpeciateCR) ✓ Speciation Profile (SpeciateSP) ✓ Pollutant to Pollutant Conversion (SpeciateCP) ✓ Lumped Profile (SpeciateLP) ✓ Compound Assignment (SpeciateCA)
Growth and control factor files in the RPO Data Exchange Protocol8 format	✓ Growth and Control Factors (GrowthGC)
Emission factor model input files	✓ Nonroad model inputs ✓ MOBILE6 model inputs ✓ NMIM inputs

*For more information on the RPO the RPO Data Exchange Protocol format, see http://64.27.125.175/tech/emis/protocol/protocol.html

The requirements associated with storing, accessing, and downloading the data and files that will be managed by the Inter-RPO Emission Inventory Warehouse are described below.

Requirements Summary

The requirements associated with the Inter-RPO Emissions Inventory Warehouse can be organized into the following categories:

- Upload - Upload and store emissions inventory data, ancillary model files, and other relevant files.
- File Management – Store files and associated metadata.
- Download - Extract and download data stored in database, retrieve files managed by the web site, allow users to download the data and files.
- Versioning – Designate and track inventory versions for any data or files uploaded to the web site.
- Security - Provide secure access to the data and files stored in the Inter-RPO Emissions Inventory Warehouse.
- Reports – Allow users to query and summarize emissions inventory data stored in the Inter-RPO Emissions Inventory Warehouse database.
- Maintenance – Streamline data, file, and user management, version control, and limited site content updates while considering the impact on licensing fees and long-term maintenance costs.
- ListServ and Notify - Implement and manage a ListServ; provide other alternatives to notify users of important updates.

Challenges

In designing a system that meets all of the requirements for the Inter-RPO Emissions Inventory Warehouse, the Inter-RPO Technical Oversight Committee (RPO TOC) had to balance competing requirements, available funding and staffing resources, and technology limitations with the need to make a readily accessible, easy-to-use web site that would meet the data needs of the emissions modeling community. Several of the key challenges considered during this design process include the following:

Volume of data. Uploading, storing and providing access to, and allowing users to download several versions of a nationwide inventory and ancillary modeling files each year requires a tremendous amount of server storage and database resources. To reduce the requirements for database storage, only those files and data needed for reporting purposes will be stored in the database. All other files will be stored in their native format, compressed if appropriate, on a file management server. To speed upload and download times, file transfer via http and ftp will be enabled.

Costs. Considering the requirement to provide Internet access to the data and files managed by the Inter-RPO Emissions Inventory Warehouse and the significant file/database server resource requirements, the licensing fees, hosting fees, and the costs of staff with specialized expertise required to prepare and maintain proprietary database technology such as Oracle or SQL Server could be prohibitive. To balance these competing needs, the RPO TOC chose to implement the website using reliable, proven, open-source technologies that are capable of handling large volumes of data. Additional details on the selected open source technologies and site architecture are provided later in this paper.

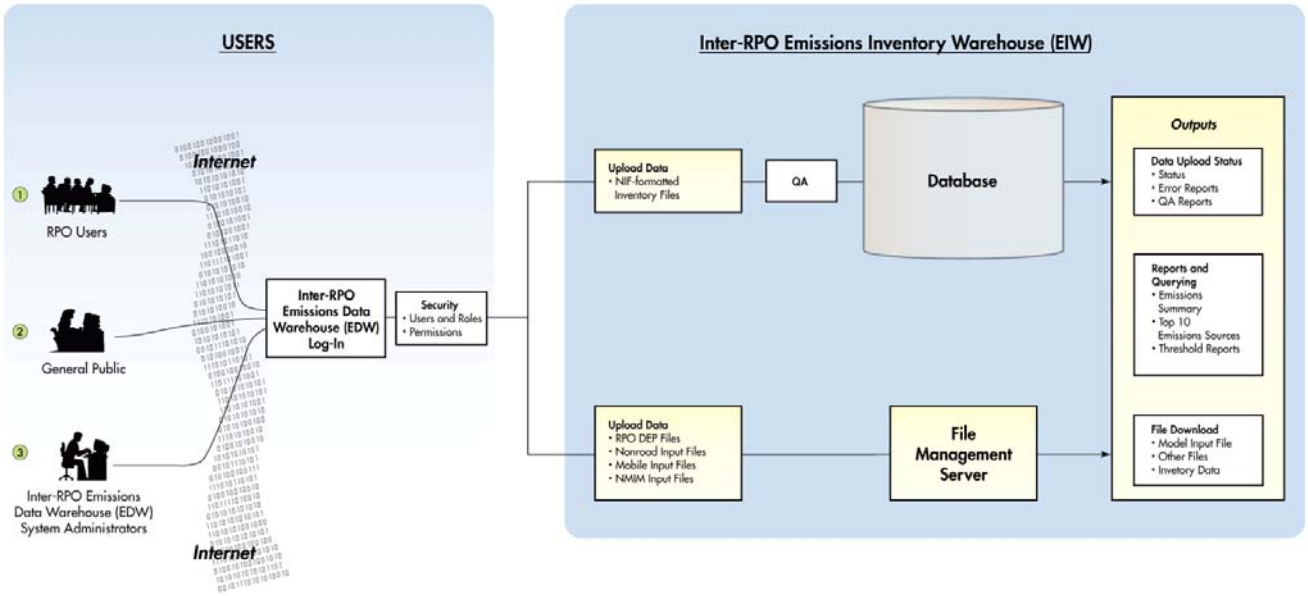
Versioning: One of the greatest challenges to the RPO community was managing and tracking inventory versions. The Inter-RPO Emissions Inventory Warehouse will automate the process of assigning versions to inventory data. Specific inventory version 'names' will be added to the database. Each RPO will have access to an upload area for their data files only. As RPO users upload data, they will be required to designate the inventory version for the file. Each RPO will be able to upload as many updates to an inventory as desired until a predetermined number of days before the inventory release date.

Reporting Performance. One of the challenges of using an open-source database tool is the potential impacts on the speed of real-time, on-line reporting when working with extremely large data sets such as those that will be stored in the Inter-RPO Emissions Inventory Warehouse. The RPO TOC decided to address this challenge in two ways: 1) after users select report criteria, the report will be scheduled and generated off-line; the user will be able to download the completed report at a later time, and 2) the speed of off-line report generation will be improved through the use of pre-aggregated data tables that can support most reporting requirements. Stored procedures will be used to populate these pre-aggregated tables after each version of an inventory is finalized and designated as ready for release.

Overview of Web Site

The Inter-RPO Emissions Inventory Warehouse is expected to be accessed by a wide variety of users from across the U.S., as well as Canada and Mexico. It is expected to provide access to up-to-date, quality assured inventories. Figure 1 presents an overall flow of information through the web site.

Figure 1. Inter-RPO Emissions Inventory Warehouse Flow



Conceptual Model & High-Level Data and Process Flow

Users are expected to access the site to upload, download, or prepare reports using the information stored by the Inter-RPO Emissions Inventory Warehouse. The design of the site balances the need for easy access to files and data by a geographically dispersed user community, and the need for intuitive and user friendly site navigation, and security. The types of users, site functionality and navigation, and security are described below.

User Groups

The different types of users expected to access the Inter-RPO Emission Inventory Warehouse web site include: Uploaders, General Public, and Administrators.

Uploaders will be specifically designated by each RPO and will have read and write access to the web site. The Uploaders are the gatekeepers that control the information and files that will comprise the Warehouse and be collated, transferred, stored, and made available to the General Public. Uploader responsibilities include:

Verifying that the content of the files and data uploaded represents the best quality inventory information available at the time of the upload;

Determining if and when higher quality files and data are available and should replace previously uploaded information;

Designating which files and information are associated with a specific version of an inventory;

Determining when a set of inventory files is complete and of quality sufficient to be made accessible to the General Public;

Entering information indicating why an inventory changed from a previously available version; and

Performing all functions available to the General Public.

General Public users will have read-only access to the web site, and will be able to view, report on, and download those data and files Uploaders have approved and made available.

Administrators will have both read and write access privileges and will be responsible for maintaining the web site hardware, software, and database; implementing web site updates, troubleshooting Uploader and General Public issues, and resolving user access issues.

Warehouse Functionality

The Inter-RPO Emissions Inventory Warehouse web site functionality will include four main components:

- Upload – to load files to be stored in the database or file management server.

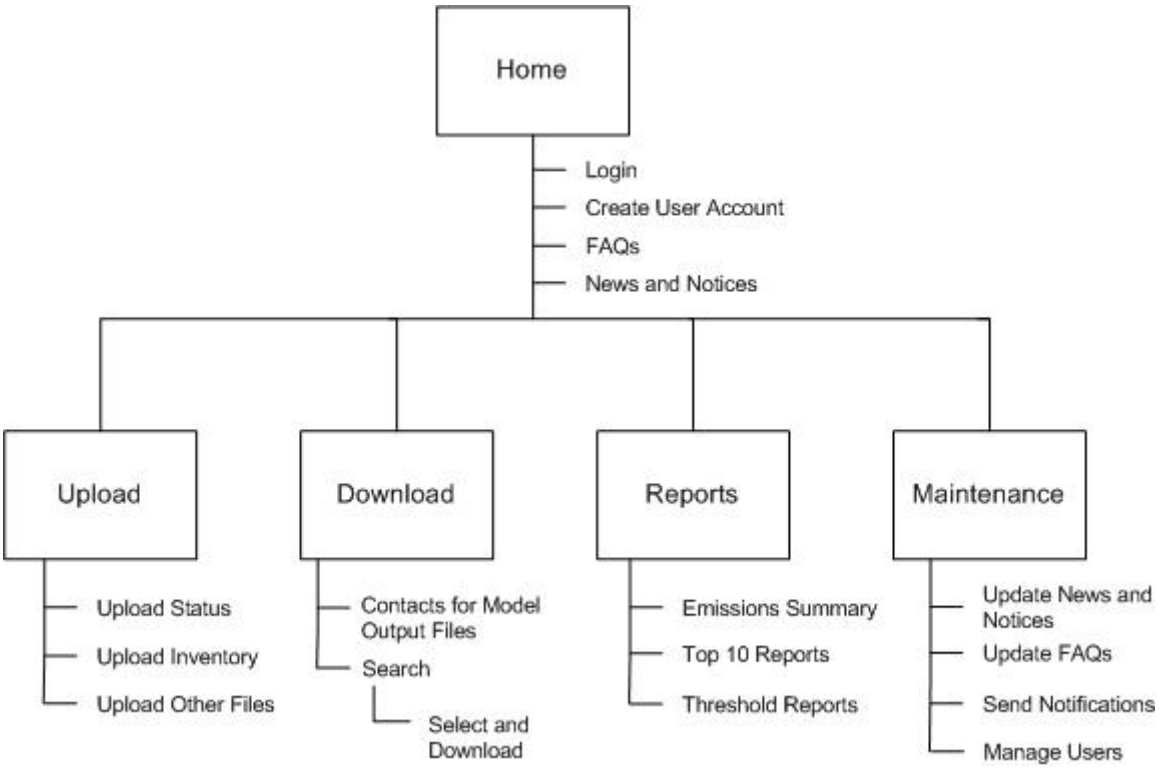
- Download – to search for and download files stored and managed by the Inter-RPO Emissions Inventory Warehouse.

- Reporting – to summarize and retrieve inventory data stored in the database. The reports will include: emissions summaries, a listing of top 10 sources, and sources that exceed pollutant and stack thresholds. Users will be allowed to generate these reports for a subset of an inventory based on source type, geographic location, SCC level 1 or SCC level 2, pollutant, and stack parameters (Threshold report only).

- Maintenance – to manage users and modify limited portions of the website content.

The Inter-RPO Emissions Inventory Warehouse web site map is presented in Figure 2.

Figure 2. Inter-RPO Emissions Inventory Warehouse Preliminary Site Map



Security

The web site will be secured through several mechanisms, including:

- Secure Login. All users of the Inter-RPO Emissions Inventory Warehouse web site, including Uploaders, General Public, and Administrators, will be required to enter a UserID and password to access the site. Users can request an account at any time; however a valid email address is required to activate the account. The UserID and password will be emailed to the address entered by the user. Email information may also be used by Administrator users to send notifications of updates.

- User-based access to web site functionality. The web site will be protected using a security module designed to ensure the integrity of collected data. Only authorized users (i.e., Uploaders) that enter a UserID and password cleared for write access will be able to upload a file. Other users (i.e., General Public) will also be required to enter a UserID and password, but will have read-only access to the web site. User permissions, roles, and access levels will be stored in the database. Table 2 shows the functionality that will be accessible to each type of user:

Table 2. Access to Inter-PO Emissions Inventory Warehouse Functionality

User Type	Upload	Download	Reporting	Maintenance
Data Uploader	X	X	X	
General Public		X	X	
Administrator	X	X	X	X

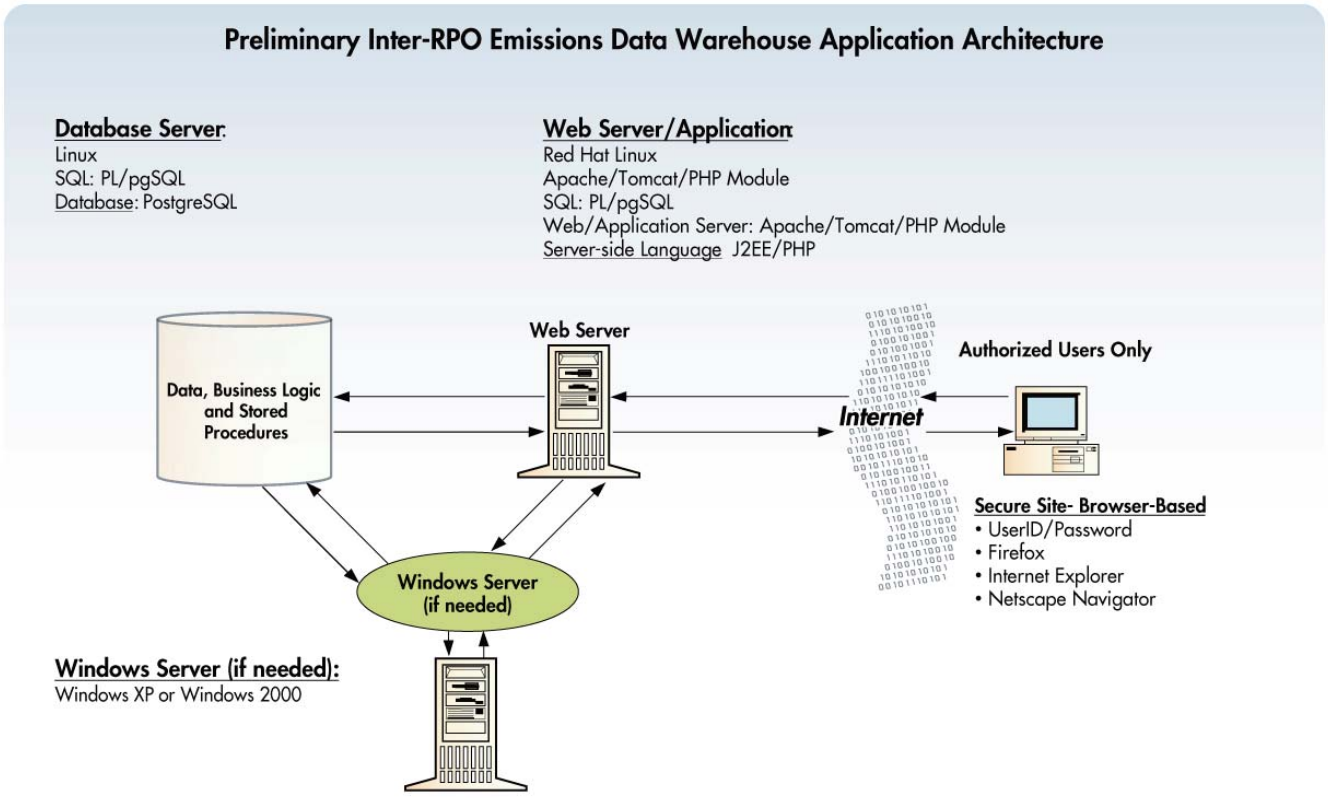
- System/Application Security - Best practices in web site security will be applied during development of the web site. These practices will prevent many potential security issues such as cross-site scripting attack and SQL injections that are currently among the most common modes of attack.

- Network/Hosting Security – The hosting provider will ensure the security and service of the entire architecture by applying network management security; network monitoring and intrusion detection; firewall, routers, ports configurations; virus protections, server log reports; network vulnerability assessments plan or protocol; incident response reporting and standards; and back up and restore standards and procedures.

Web Site Architecture

In balancing the need to handle a large volume of data/file uploads and downloads with licensing and maintenance costs, the RPO TOC determined that a web site solution based on open source technology would best meet the needs of the project. Figure 3 presents the preliminary Inter-RPO Emissions Inventory Warehouse web site architecture.

Figure 3. Preliminary Architecture for the Inter-RPO Emissions Inventory Warehouse



The main components of architecture include the following:

- Web Server Platform,
- Web Server,
- Application Server,
- Database Server,
- Database and Procedural Language,
- QA Server (if needed), and
- File Management Server.

A detailed description of each of these components is provided below.

Web Server Platform

The *Web Server Platform* is the operating system of the physical server supporting the website and networking. To allow for the use of open source technology for other components of the architecture, Linux, an open source operating system, will be used.

Web Server

The *Web Server* is a process that runs on the physical server that handles HTTP requests received from an Internet. An Apache HTTP Server with the Tomcat module of the Apache Jakarta Project for Java support will be used. Apache is an open-source server that is compatible with almost all Linux distributions (as well as Windows and several other operating systems), and is the most commonly used web server on the Internet.

Application Server and User Interface

The *Application Server* is where the business logic for the website is located and processed. This includes the code that: generates the dynamic content that is displayed to the user on their browser, communicates with the database to support querying and reporting, and interacts with the operating system to process uploaded files. The Tomcat module of the Apache Jakarta project will be used to support web page development using Java Server Pages (JSP), database communications using Java Database Connectivity (JDBC), and operating system interactions using the Java programming language. Java is a freely redistributable platform for dynamic web applications and other types of programming needs.

Database Server

The *Database Server* is a process that runs on the physical server that houses, operates, and manages access to the databases. To allow for the use of open source technology for the database, Linux, an open source operating system, will be used.

Database and Procedural Language

The *Database* is where the information accessed by website reporting and export functionality is typically stored. PostgreSQL, a relational open-source database that was first made available 15 years ago, will be used. PostgreSQL is robust, stable, and scalable, and runs on all modern Linux distributions.

The *Procedural Language* is the tool used to manipulate information in the database, or modify the database structure in some way. PL/pgSQL is the PostgreSQL procedural programming language, and will be used for internal database operations (such as moving data from one table to another).

Additional detail about PostgreSQL and PL/pgSQL can be found at http://www.postgresql.org

Windows Server (if needed)

A *Windows Server* that can be used to extract emissions inventory data from NIF formatted Microsoft Access files and possibly complete QA/QC of these files may be included in the site architecture if needed. This server would be used to communicate and retrieve data stored in the MS Access formatted inventory files, and run the EPA Format and Checker tool on files uploaded to the Inter-RPO Inventory Warehouse.

As an alternative to the Windows Server, the RPO TOC is investigating restricting data uploads to ASCII format only, and whether key formatting and referential integrity checks can be completed during the data upload process.

File Management Server

The *File Management Server* will be used to store model input files, non-NIF formatted RPO-Date Exchange Protocol files, and other ancillary files associated with an inventory. This server may be combined with the Web Server or housed on a separate server.

CONCLUSIONS

The Inter-RPO Emission inventory Warehouse is expected to provide the states, RPOs, EPA and the general public with quick and easy access to up-to-date, quality-assured, inventory data and ancillary model files. Use of open-source technology should minimize maintenance costs and allow for expanded site functionality to be contributed by the open source community. The project is currently in the Design Phase, and development and implementation of the system will take place during the summer of 2005. The system is expected to be completed by January 31, 2006. All decisions regarding the development of this system will be made by the Inter-RPO Technical Oversight Committee.

KEYWORDS
Inventory
Emissions
Inter-RPO
Warehouse
Data management