Next Steps Towards a Shared Emissions Database for North America

Prepared by

Alpine Geophysics, LLC

and

Washington University in St. Louis
Center for Air Pollution Impact and Trend Analysis (CAPITA)

For the

Commission for Environmental Cooperation

15 April 2004

Table of Contents

1 Introduction

Recognizing the value in sharing emissions information, the three environment ministers of North America who comprise the Commission for Environmental Cooperation (CEC) Council agreed, in 2001, that their governments would work towards the development of a shared North American emissions inventory for criteria air pollutants and greenhouse gases. Working towards this goal of a shared inventory, the CEC began a pilot project in 2003 to explore the availability of emissions data for the electricity generating sector and the feasibility of sharing that data electronically between the three countries. The availability of emissions data has been summarized in an earlier report (CEC, 2003) and a demonstration of a distributed database of this emissions information has been prepared (Falke, et. al, 2004). The purpose of this report is to discuss the future direction of electronic data sharing and to recommend next steps along the path towards a shared emissions inventory for North America.

One vision for the future of a North American emissions inventory is provided by the Networked Environmental Information Systems for Global Emissions Inventories (NEISGEI), a U.S. EPA initiated effort to create a web-based global air emissions inventory network that provides a catalog of distributed emission inventory data, tools for processing and analyzing the data, means for registering new data, and an environment for collaboration among air quality researchers, policy-makers, and the interested public.

The development of this integrated network of emissions data, tools, and community is being pursued incrementally. The efforts under the CEC to explore a shared inventory for the electricity sector are one of several pilot projects that the U.S. EPA hopes will contribute to the development of the network. In the initial phase of activity, projects were designed to incrementally develop the information technology required to support user access to existing local, regional and global pollutant inventories, and to provide interoperable tools for merging and manipulating these heterogeneous data-sets for modeling and policy analysis. A design goal in the infrastructure development is that interfaces will be built on top of existing inventories to make the inventory data accessible by network, while the data remain distributed and under the control of the institutions responsible for developing the emission inventories.

The successes of these initial projects, as well as advances made in related distributed information system research and development will be leveraged to begin building operational networks of distributed emissions data and analysis tools among a consortium of users and inventory developers. Related efforts include EPA's Environmental Information Exchange Network, efforts by the Regional Planning Organizations in support of the Regional Haze Rule, and broader interagency efforts such as Geospatial One-Stop.

After the initial infrastructure has been developed and deployed, attention will be placed on the research and development of data search and manipulation capabilities that will further advance the state-of-the-art in web-based database management and data manipulation and make possible the tools needed for more powerful environmental data analysis.

Ultimately, the goal of NEISGEI is to build capacity in developing countries by providing technical information and support for inventory development according to international

standards. The inventory network will partner with international aid programs to maximize the use of the network's resources for new inventory development.

Opportunities currently exist for participating CEC partners to take advantage of, and supplement, existing and planned activities in the development of an integrated North American emissions inventory and inventory query and distribution tools. These activities could immediately place the three partners on even footing allowing for the review, distribution, and integration of inventory development information and experience.

Section 2 of this report summarizes the primary features and benefits of the NEISGEI approach for implementing the commitment to a shared emissions inventory for North America. Section 3 discusses whether this approach is feasible given the state of the science and technology and discusses some of the non-technological challenges that must be addressed. Section 4 provides some specific recommendations on next steps to improve data sharing in North America. Section 5 discusses the long term prospects for fulfilling the NEISGEI approach in North America and globally.

2 Networked Environmental Information Systems for Global Emissions Inventories (NEISGEI)

The Networked Environmental Information Systems for Global Emissions Inventories (NEISGEI) refers to both a conceptual framework for and a step-wise development and implementation effort towards the realization of a fully integrated, distributed air emissions inventory. The conceptual framework describes a system that ties together data at all spatial and temporal scales using emerging Internet distributed database technologies; provides for the seamless merging, manipulation and analysis of any Internet accessible – and locally available resident – air quality-relevant data through the development of emerging Internet-oriented technologies; and proposes the development of a broad-based NEISGEI user community of scientists, regulators, policy analysts and the public. The technology development and implementation philosophy calls for the assembly of the network and toolkit in a piece-wise manner, allowing immediate functionality and value to the air quality community.

The resulting network of data and tools will lend itself to strategic planning of air quality and environmental management capacity building projects. Simplified, multi-party, cross-border collaboration in air quality studies and management will be made possible with a fully functional NEISGEI structure. Additionally, the development of environmental indicators, with the inclusion of data available on other environmental media will be facilitated. A fully populated network will be an invaluable resource for identifying important missing datasets for regional, hemispheric and global scale studies.

2.1 Network Infrastructure Development

At present, air quality information cannot be easily shared between parties because of technical incompatibilities. However, air quality management requires information at increasingly larger geospatial scales. By leveraging the efforts of individual institutions to collect and process air quality data and the technological advances made by those working at the leading edge of the distributed database access field, NEISGEI will enable collaborative and complimentary efforts of scientists and regulators working in related air quality disciplines and adjacent regions to share and analyze data.

2.1.1 Universal Access and Interoperability

In our global environment, where the success of air pollution management efforts depends on the constructive involvement of many stakeholders, freely accessible air quality information is crucial. Recent advances in information technology have made simple access to public data possible for all interested parties.

Of even greater interest is the potential for electronic submittal and retrieval of data, cooperative data exchange, consistent reporting protocols, and more powerful query tools which could greatly facilitate the international air quality decision making process.

2.1.2 Distributed Data Sources

The distributed nature of future information systems opens the channels to information access by removing the requirement of a single system or authority to monitor and distribute environmental data and knowledge. The flexibility that this principle brings to the NEISGEI framework will allow for the limitless, and immediate, access to environmental information the moment it is released to the public domain.

The increased need for real-time access to information has driven technology to reinvent itself by becoming scaleable and finding ways to communicate and collaborate across the artificial boundaries of organizations and geography. Individuals are becoming more empowered and seek the ability to become involved in decision-making in arenas they had never before considered. This increased involvement leads to greater potential for effective, sustainable, and equitable data development and analysis.

By involving the ideas and data of interested parties to the NEISGEI framework, the potential benefits are related to the advantages of individual agencies: greater understanding of local problems, better control and accountability during execution and operation of projects, and better ability to involve the air quality community and utilize existing resources to further their own localized needs.

2.2 Community of Providers and End-users

As individuals, great strides have been taken to improve the quality and access to environmentally related data. However, as an international community, much work still lies ahead. Air quality researchers and managers are often frustrated by their lack of access to high quality emissions information. Information technology researchers have interests and abilities in distributing information through the internet. By bringing these two communities/groups together under a consortium of cooperating mechanisms, they will be able to express themselves in terms that everyone will be able to comprehend and immediately benefit from.

Without the benefit of an international user community, however, the ability of the system to coordinate international efforts will never be achieved. Local and national scale programs have been established and are currently in practice among the North American environmental agencies. The combined efforts of these agencies to develop a single concept of environmental data access is essential for the coordinated reporting, policy development, transport analyses, and socio-economic studies that create an environment for collaboration among international researchers, policy-makers, and the interested public.

Realizing this vision is necessary to maintain and improve air quality on an international scale. Achieving the vision is possible, but by no means guaranteed. In addition to carrying out their current functions well, individual partners will need to develop new skills and approaches, become more efficient, and work together more effectively to solve problems. Participants in the consortium will have to continue to dedicate themselves to conducting sound research and promote new advancements in information technology transfer; respond creatively to new challenges and emerging issues; and improve the working partnership of all participating parties to the concept of a shared network of environmental information.

In addition to the strengthening and further development of the scientific and policy user community in environmental and air quality sciences, the pursuit of the principle of best science at an international and global level and promotion of applications by teaming together researchers from academia and private industry could be realized through an interactive consortium of users. The concept would promote research in the environmental field from the general public and in particular the next generation of environmental investigators through important education and outreach opportunities.

2.3 Capacity Building

2.3.1 Ability to Evolve

With the sponsorship of international organizations, these needs could further be identified and strategies outlined with an overall goal of developing the high-quality, accurate, and comprehensive data sets necessary for the various aspects of air quality analysis.

The NEISGEI concept is an example of a strategy that involves the creation of a web-based global air emissions inventory network in the form of a Web portal. Through international programs, consortium members could assist each other in identifying the capacity building and technology transfer needs required to achieve a uniformly high degree of quality in data collection, processing, and distribution among all nations. This network provides a catalog of distributed emission inventory data, tools for processing and analyzing the data, means for registering new data, and an environment for collaboration among international researchers, policy-makers, and the interested public (Hemming, et al. 2003).

The capacity building element of this initiative includes making readily available to developing countries the technical information and support for inventory development according to international standards. The inventory network will partner with international aid programs to maximize the use of the network's resources for new inventory development.

The concept involves the initial creation of the underlying network infrastructure, using newly available off-the-shelf and web-based distributed database management tools, and the creation of a consortium of users and inventory developers. Under this model, each of the participating members could provide or benefit from the underlying data used to quantify air quality from participants in the user community.

The continued design of this framework promotes the development of innovate strategies to expand the network while maintaining a standard that will allow others to utilize what already works.

2.3.2 Common Interpretation and Exchange

The ability to make air quality data publicly available for modeling and analysis requires a common, flexible, and well-documented process for reconciling differences in data formats and structures. To date, individual nations, States, Provinces, local and Tribal air agencies have

collected and maintained data in formats which best suited their individual needs. Some of these agencies may have only required summary data for general reporting requirements, others may have needed more individualized facility reporting to gain permit approvals, while others may have required more temporally and spatially refined data in order to support regulatory and policy analyses.

A set of common processes for disseminating data would allow individuals to maintain responsibility and control over their own data and reporting schedule, while at the same time making it available to other interested parties in structures already understood and documented. These structured formats should provide a minimum set of information on methods, factors, and activity data, as well as relevant assumptions that underlie the estimates for each element. Information provided through a common process is aimed at enhancing the comparability and transparency of air quality data by facilitating activity information among consortium partners, and easy identification of possible mistakes, misunderstandings and omissions in the data. However, commonality should not override maintaining the dynamic nature of ever changing data needs.

Examples of common reporting paradigms currently in use include the U.S. EPA National Emission Inventory (NEI) Input Format (NIF) and the United Nations Framework Convention on Climate Change (UNFCCC) for emissions inventory information.

The current versions of the NIF and all user documentation are posted online for users to download (EPA, 2003). In addition, there is an associated software program for download to help NIF users perform quality control checks on their files to ensure correct format specification. This format is widely used by State and local agencies to transfer data to the EPA's NEI. The format has maintained its usefulness by adapting itself to meet the needs of the user community while retaining the same variables necessary to define the emission inventory process.

Under the UNFCCC, reporting guidelines and requirements to submit an annual inventory report on greenhouse gases are increasing the need for more detailed inventories. Countries are also being encouraged to report on additional air contaminants through this framework. The revised Intergovernmental Panel on Climate Change (IPCC) guidelines and common reporting format developed under this framework might be considered in the efforts of a consolidated inventory of environmental data.

The importance of a common methodology derived through international coordination counts on a uniform reporting structure, consistency among reporting entities, the ease of information submission and modification, and reduction of burden for information exchange.

The UNFCCC summary of greenhouse gas inventories for individual countries demonstrates the importance of a common methodology derived from an inter-governmental agreement which produces emission estimates that are comparable at a national level. These methods, designed to develop and refine an internationally-agreed methodology and encourage the widespread use of this methodology by countries participating in the IPCC and by signatories of the UNFCCC have been considered credible, objective, and transparent (EPA, 2002).

However, although the UNFCCC guidelines have promoted the uniformity and transparency of these internationally reported data, the rigidity of the formats have prevented them from continued adaptation to ever changing reporting needs. This is an example of one of the major challenges that the development of a truly integrated North American emission inventory will have to overcome; well-defined and flexible reporting.

2.3.3 Partnership

The NEISGEI concept establishes the framework to build better partnerships among common goal partners. Through the development of more effective and efficient opportunities to share environmental information, a consortium of users will benefit from each others advancements. The reliance of one user on another promotes a more collaborative environment that offers individual confidentiality, confidence, and responsibility for one's own information.

Since the consortium of users would be working together for a common objective, similar data, goals, and information needs will predicate the reliance of one party on another. No one member's information would be more important than another's. A commitment to align common objectives would strengthen working relationships and promote sound science. From the start, the consortium would define priorities for data standards, development, and documentation and acknowledge that individual programs may not be as far advanced as most. A cooperative solution would be found recognizing the needs of the consortium to advance their efforts and adapt to differences.

For the first time, many agencies will be taking the responsibility of managing their air quality data programs for purposes of international distribution. In these cases, those groups should work to build the capacity, expertise, and infrastructure necessary to support not only their local program needs, but the eventual needs of the international community. These agencies will rely on the experiences and programs already working at other agencies and will immediately benefit from the cooperative exercise.

The creation of an international partnership trusting of each other will not come easy. However, through the successful and iterative development of the NEISGEI concept, initially utilizing existing frameworks and building new and modified ones when identified, the reliance of one party on another will strengthen and grow in time. Eventually, the participation of new members will be recognized as confidence in the framework and its users.

3 State of the Science and Technology

The potential benefits of distributed and shared data systems have driven a broad base of research and development among information technology organizations and environmental information users. These efforts are beginning to address the numerous technological and organizational barriers to the fulfillment of distributed environmental information networks.

3.1 Distributed Information System Initiatives

The inability to share and integrate environmental information among agencies and organizations has lead to inefficiencies and ineffectiveness in environmental management and policy. A recent U.S. General Accounting Office (GAO) report attributed inefficient and ineffective collaboration among the agencies involved in the management of forest fires to the heterogeneous information systems employed by the numerous agencies and their inability to adopt new information technology (GAO, 2003). While the report singles out forest fire management, its critique of effective information systems extends to most other environmental areas and is reflected in new efforts among federal and international organizations to address similar issues.

The National Science Foundation, in acknowledging advances in information technology and their potential in revolutionizing the way in which research is conducted, has established new programs focused on developing cyber-infrastructure and applying it to science domains. The program has just begun to fund projects that apply information technology advances within science and engineering domains, including environmental disciplines.

The EPA has struggled in the past with sharing information among its multiple offices and in 1999 established a new Office of Environmental Information to improve its information coordination. In addition to devoting resources to the internal information exchange with the agency, EPA is also developing an Environmental Information Exchange Network to reduce the complications in transmitting data collected by the States and EPA's central data exchange server.

The Earth Observation Summit is an international effort to build comprehensive, coordinated, and sustained Earth observation system that consists of interoperable databases among the participating countries. The Earth Observation Summit is being lead by the G-8 countries and aims at identifying, and filling, observation data gaps in environmental knowledge.

Other efforts are also in their initial stages and a subset of these are briefly described in Appendix A. A common goal among these programs is to make environmental data easier to access in forms relevant for a variety of end applications. In general, these programs are centered within specific sub domains of environmental science. A step-wise method, beginning with domain-specific applications and gradually expanding in scope by connecting with other parallel efforts is likely to be the selected approach to a multi-domain information network. Every community, whether air quality managent in general or emissions inventories in particular, must define its set of standards and protocols for unifying its data and its infrastructure. Only after each community reaches internal equilibrium can the larger issues of inter-discipline linkages truly be addressed. In many cases, the linkages will be straightforward as

interoperability standards adopted by individual communities are equally applicable outside their societies. Nonetheless, the road to fulfilling this broad goal is long and is far from an immediate certainty.

3.2 Challenges in Fulfilling Distributed Systems

The challenges in attaining distributed data and tool networks encompass both technological and organizational aspects. It is generally argued that technological innovation has reached a state where distributed data networks could be meaningfully implemented (Falke, 2002). Assuming that argument to be true, organizational and cultural barriers prevent such a network to be deployed on a substantial scale. The institutional history of government organizations tends to be defined by information systems and applications designed for specific purposes. In many cases, these systems are developed and maintained by contractors with narrowly defined contracts. Changing these contracts to reflect a distributed data sharing approach is a non-trivial and potentially costly endeavor.

A summary of identified barriers to distributed emissions inventories is provided in Table 1. The data actor column in Table 1 indicates a stage in the access and use of emissions data. Data providers are the organizations that store and maintain the emissions inventory database. Mediators are at the intermediate stage and provide the necessary processing and tools for translating data into uniform access. The users are the end customers of the data who apply the data into their specific applications. Each stage has its own set of requirements for a successful distributed data network.

Table 1. Distributed Emission Inventory Barriers

Data Actor	Distributed Network User Issues / Requirements
Provider	Technology implementation Server and database security Bandwidth limitations Data misuse
Mediator	Consistent and stable access to data provider Database mapping
User	System performance and responsiveness Easy to use interface

3.2.1 Organizational Challenges

In building a consortium of data providers, it is imperative that the providers can see a benefit, or return on investment, from joining the network. Joining for the sake of the community at large is not sufficient. Data providers must stand to gain from sharing their data. Some potential

benefits include increased exposure and use of their data, access to other data sources, access to tools that add value to their data, and easier methods for collaborating with other organizations.

Another critical issue is maintaining appropriate acknowledgement and recognition for the data provider's information. Even though the data remain physically within the purview of the data provider in a distributed network, the front end to that data can be located anywhere on the network. A third party interface to an organization's data can potentially give the impression that the data are being served from the third party and in the process lose the credit due to the data provider. Ensuring credit for contributions should be a priority in the design of distributed data networks.

An additional hurdle to be addressed is data misuse. Data have inherent limitations in their relevancy to questions they can answer. In a centralized system, data distribution can be limited and therefore inappropriate applications of those data controlled. An openly shared system could potentially lead to greater use of data in contexts not intended by the original providers. On the other hand, a shared system would lead to greater use of data and improved community-wide recognition of a data set's limitations.

3.2.2 Technical Challenges

Despite the many technological advances, significant needs remain before distributed networks will be accepted within scientific and policy communities. Perhaps most significant is security. A data provider must have assurance that making their data available through a distributed network will not adversely impact their operations. Concerns include an increased volume load on their servers which could lead to disruption of their mission-specific operations as well as security breaches due to opening their databases to the outside world.

Consensus derived standards and protocols are still missing in many aspects of distributed computing, particularly in describing and defining the services for making data available and accessible through distributed tools. The hope is that, as distributed computing becomes commonplace, these standards will stabilize and promote the expansion of distributed data networks.

An effective distributed system should be responsive to the user. Accessing large datasets, such as multi-dimensional national emissions inventories with thousands of emission point locations, is currently too cumbersome for efficient user interaction. This performance limitation should not be considered insurmountable. The continually expanding bandwidth of internet networks and more efficient algorithms for handling distributed data promise to make distributed systems fast enough for everyday use by researchers, managers, and the public.

The underlying data presents its own set of challenges. In many cases, data are not inherently accessible. They are primarily handled "off-line" and then made available for simple download. Particularly, in the case of air emissions data and their complex relational structures, it is non-trivial to convert static files into dynamically accessible datasets.

The technical and organizational challenges outlined above point to the need for a possible "hybrid" solution to distributed databases that adheres to the principles of distributed databases

but simplifies the technical and organizational requirements by avoiding direct end-user access to the underlying databases. Section 4.4. outlines possible approaches for such a solution.

4 Continued Development of a North American Emission Inventory

The development of a single comprehensive and accurate emissions inventory is essential for the coordinated reporting, policy development, transport analyses, and socio-economic studies that create an environment for collaboration among international researchers, policy-makers, and the interested public.

4.1 Status of Integrated Emissions Inventory

The CEC recently supported a pilot project in which emissions data from North American power plants were compiled and provided for the development of data and infrastructure for a distributed electronic database of emissions information (CEC, 2003). Although the power plant source sector is thought to be one of the most consistently reported among the three countries, it can be seen that there still is a way to go before the data are considered transparent enough to represent a coordinated North American inventory.

A coordinated inventory should be based on the fact that emissions information is universally available for the purpose of developing local, regional, national, and international plans that are comparable on a source sector basis. Data presented in the CEC report shows that although individual local or regional inventories are prepared and exchanged in common ways, additional international coordination still needs to occur. Reporting periods, unit specific information, confidential data, and methods for estimation continue to be opportunities where improvement can be achieved.

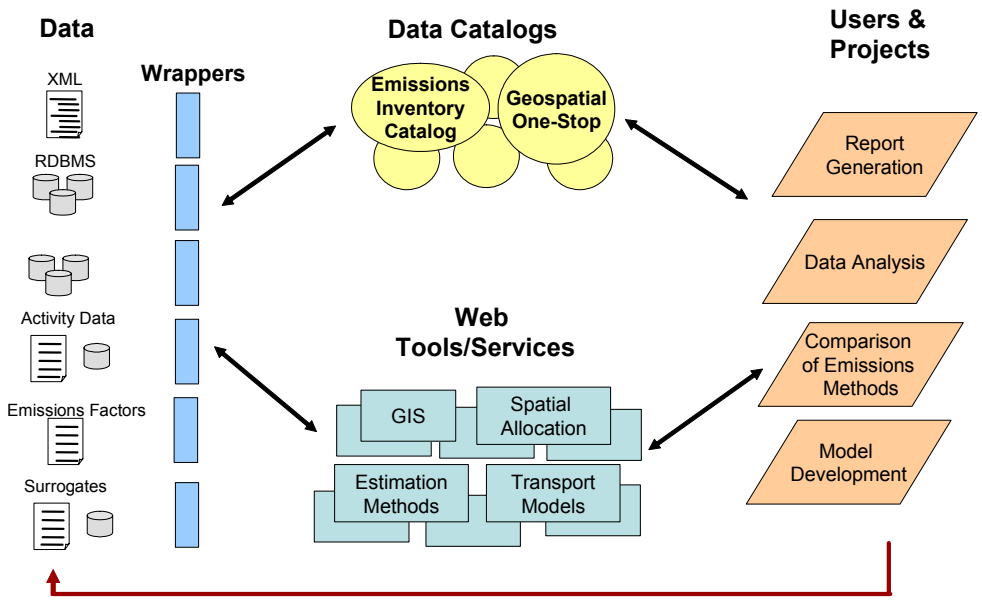
The current state of an integrated North American emissions inventory is a prototype tool that allows users to browse heterogeneous emissions databases. This prototype tool is an important first step that has demonstrated the feasibility of the NEISGEI concept and has introduced individuals involved in North American emissions inventories to the benefits of a distributed data network. Numerous technical and political challenges remain before an operational distributed emissions network will be realized but the stage has been set to begin to address these challenges and progress from the prototype to an operationally implemented system.

The envisioned end state of a distributed emissions inventory is depicted in Figure 1. Distributed data sources (emissions estimates, activity data, surrogates, etc) in a variety of formats (relational database management systems, text files, etc.) are registered in data catalogs. The data from these catalogs can then be uniformly accessed with the aid of data wrappers (translators) and connected with web tools and services to support a variety of end applications.

4.2 Benefits of an Integrated Inventory

The importance of a common integrated emission inventory derived through international coordination counts on a uniform reporting structure, consistency among reporting entities, the ease of information submission and modification, and reduction of burden for information exchange.

Figure 1. – Distributed Emissions Inventory Flowchart



Our long term goal as North American inventory data developers and users should be the advancement of an emissions inventory that is universally available to all who want to access its information, is of high-resolution and source and facility specific, is comprehensive with respect to pollutants and sources, is well documented and easily duplicated, and is based on comparable methodologies and factors. This inventory would allow for ease of technical and policy evaluations by data providers, stakeholders, and other interested parties.

Canada and the United States currently have operable inventory management systems designed to receive, store, process, display and output emissions data. These systems are also being modified to receive, store, process, and display combinations of the activity data and emissions calculation methods used to estimate emissions inventories. Continued work is underway to provide inventory display capabilities which will include GIS functionality, tabular and flat file data formats, graphs and charts, and the ability to capture these displays in user-defined report formats.

The Mexican National Emission Inventory (MNEI) effort is in a position to take advantage of the experiences from the Canada and U.S. emissions inventory development and database management system implementations. The MNEI already has benefited in terms of compatibility with the other North American inventories as it has adopted the NIF data format standard allowing it to implement existing web interfaces for its database that allows users to query and download subsets from the emissions inventory.

There is an opportunity for the Mexican data management team to supplement an online system with web services that allow dynamic data access. This additional capability can be implemented without altering database design and would place the Mexican emissions group in a

lead position for North American data sharing and dissemination. The CEC power plant integration project has shown that while the Canadian and U.S. emission inventories are easily downloaded through online interfaces, they do not yet lend themselves to the type of dynamic access that is needed for a distributed emissions database. The process of establishing a Mexican emissions data access service would provide a template and test case that Canada and the U.S. could use to implement similar services for their databases.

An envisioned system would be capable of:

- Receiving and storing emissions data in emissions reporting formats commonly used by various agencies and sources with little or no additional effort;
- Producing user-specified reports (including model-ready input);
- Performing user-selected quality control and assurance tests;
- Allowing data queries and graphic display; and
- Presenting this information as geographic information system (GIS) displays, flat files, air quality model-ready input files, and in other formats.

4.3 Existing Demonstration of Integrated Inventory Development

The feasibility of a future integrated North American emissions inventory has been demonstrated as a web tool for browsing heterogeneous emissions data sources. From a general distributed database technology perspective, we are at a point where distributed database concepts can be applied to actual implementations. However, when focusing on the development of a distributed air emissions inventory, the CEC pilot project encountered numerous technological and organization challenges in dynamically accessing the currently available emissions inventories.

A key difficulty encountered was that the data access of emissions data is currently not at a state that allows dynamic, distributed access. Emissions inventories are currently not designed for such application and while most of the emissions inventories used during this project were available through the Internet; their web access methods only support single user access.

Attempts were made to automate a manual approach through an internally hosted web server but those attempts failed to produce a stable, reliable method for accessing the data. Most access software utilizes some authentication technology that prevents dynamic, server-side access. For example, the products employed by U.S. EPA's databases are designed for single user access through a desktop computer. This limitation prevented automation of the dynamic access of U.S. EPA databases through a web server interface. Security is and will continue to be an important concern in distributed data access and is one of many issues yet to be resolved.

These challenges are certainly not insurmountable and collaborative efforts in the future are likely to generate an operational distributed North American emissions inventory. The development of this inventory would benefit from a step-wise approach that initially focuses on the most readily available and multi-country comparable data. The preliminary versions of the inventory would help clarify issues related to handling complex queries. Building and using initial versions will assist in creating consensus approaches to issues as straightforward as data

naming conventions that could make the exchange of emissions data among the three countries even simpler.

The benefits of a distributed emissions inventory was recently referred to in comments made during a demonstration of the pilot prototype tool. Potential users mentioned the ability for the tool to save the time and effort involved in finding, collecting and formatting international emissions data generated outside of current programs. Additionally, the tool's usefulness in exploring these data and comparing emissions from multiple sources was commended.

A distributed data network will not replace the currently implemented systems for querying and downloading emissions inventory data but will rather supplement the already functioning systems with additional capabilities for disseminating these data. A distributed system will open the emissions inventories to a broader set of users and will increase availability, review, and ultimately the quality.

4.4 Next Steps

In progressing toward a distributed emissions inventory, it is important to keep the goals of a distributed emissions inventory at the forefront:

- Minimum burden on data providers;
- Shared and distributed data; and
- Uniform and transparent user interface to data.

We contend that the progression toward a distributed emissions inventory can be promoted by continuing to upgrade the state of a distributed emissions inventories using technologies and techniques that do not impose additional burdens on data providers and by encouraging emissions inventory managers to adopt new technologies that foster the sharing of their data with external clients. Among these technologies are web services and related standards, such as the OpenGIS Web Map Server that Canada NPRI is planning to implement.

The implementation of an operational distributed emissions inventory will be more quickly realized by focusing on defining a process for dynamically linking emissions inventories rather than imposing data format, software, and hardware standards. As part of this process, defining a finite list of focus points would aide in the development of integrated inventories. A few of these items include:

- Metadata- More complete description information about emissions databases would help in relating heterogeneous data;
- More complete access to distributed datasets - A process is needed for creating trusted provider-user agreements that would, in turn, help address issues of security and data misuse;
- More comprehensive content - Networked data and tools that spark additional interest in the technology's potential;

- Tool development - Advance prototype tools to make them more representative of what an emissions inventory would ultimately use;
- Integration - Links to other web services (Geospatial One-Stop, TerraServer) and other relevant data sources, including Web Map Servers;
- Partnerships - Establish collaborative partnerships with other researchers and agencies developing related networks and tools;
- New capabilities - Build tools that add value to the distributed data and provide incentives for data providers to join networks; and
- Network performance - Clarify the handling of distributed data for optimal system performance.

Progress will be made in a step-wise fashion as a combination of technology "push" and "pull" efforts. The adoption of new approaches to data management and use of information technology will be driven from many sides, such as the desire to make federal data available through Geospatial One-Stop. The plan for a distributed emissions inventory should adjust to these changes as they occur.

This progression has both a near term and long term approach. The near term proposal uses currently implemented technologies that allow a distributed database to be created through the application of data caching. The longer term proposal focuses on web service technologies that would require additional technology implementation by the data provider in order to become a node on the network. Both of these approaches embrace the concept of mediated data access; a middle component between the data provider and data user that adapts to the data and user needs in fostering their interaction.

4.4.1 Near Term Opportunity – Cached Data Access

Emissions databases are already available through online interfaces and many can be automatically queried through single user access accounts. These allow single users to download data but do not allow other distributed servers to handle multiple user queries and pass them along to the emissions database. A solution to this problem is to use a mediator structure that dynamically accesses the data and store it as a cached dataset on the mediator's server. This mediator could then supply the data using web services that allow distributed access while still maintaining the original link to the data source.

Instead of accessing the data directly from a database each time a query is executed and then discarding the results after the user is finished with that query, this data system would store the retrieved data in a cache and in a format that allows for efficient access, querying, and analysis of large, multidimensional emissions data.

A cached data solution would avoid the data provider issues outlined in Table 1 as it would construct an intermediate, virtual instance of the data as designed for efficient user access. Instead of a user query going directly to (and burdening) an emissions inventory database, the query would instead only interact with an intermediate form of the data.

The cached data would contain a relevant subset of data and would be dynamically linked to the initial emissions database (thereby benefiting from the advantages of the original data) so that when updates occurred in the original data, the host system of the intermediate information would be notified and updated with appropriate changes. This would ensure that a shared, single version of the data was continually available to end users. Maintenance of the data cache would be automated with minimal associated cost. Human interaction would be required during setup to provide the mapping to the data provider's database but once the link with the data cache was established it would automatically update to reflect changes in the source database.

4.4.2 Long Term Opportunity – Web Service Technology

In the longer term, it is feasible to think about a peer-to-peer type of distributed access network. Such a network would allow direct access to each emissions database on the network after each data server implemented web services or some alternative web interface method of dynamically accessing the data. Because these web services are self-contained and use Extensible Markup Language (XML)-based standards for describing themselves and communicating with other web resources, they can be reused in a variety of independent applications.

In the web services network approach, mediators serve the role of brokers, providing users with the interfaces for finding available data, dynamically retrieving it, and integrating it with other distributed data sources. These network users can function on an independent level, each addressing local issues of importance. These individual components can then be integrated or modified to handle differing data types dynamically on demand.

Web service technology is still evolving and does not currently provide a convenient off-the shelf software solution. However, many required components are considered standards in peer-to-peer web programming applications and therefore make it possible to create an operational data web service. These components allow computer-to-computer communication in a platform- and programming language independent manner. Additionally, web service technology provides existing software applications with service interfaces without changing the original applications, allowing them to fully operate in the user's existing environment.

Both the near term and long term opportunities provide solutions to the challenges encountered in developing integrated emission inventories in a distributed nature. What continues to remain unclear is which method, combination of the methods, or other methods will best offer the final resolution to an operational distributed emissions inventory on a North American scale.

5 Coordination

In the absence of the adoption of a shared framework of reporting regulations and formats for environmentally related information, there will always be differences in the way air quality data are prepared, presented, and used. Ideally, through cooperative efforts, stakeholders in the information sharing process can work together to enhance data comparability and accessibility; hence minimize potential errors and misunderstandings arising from incompatible data sets.

An opportunity to provide resource savings, minimization, and cost-effective solutions to individual issues is provided by the NEISGEI concept which can be used as a resource to make planned projects easier to accomplish. Using existing applications to drive future activities allows for individual tools to improve. The focus can continue to be directed toward data enhancement, expansion, and quality while knowing that the framework to make the data distributable is already in place. Agreed upon formats, protocols, and standards eliminate the need and associated resources required to develop these items.

An example of this coordination comes from the United States Environmental Protection Agency (EPA) Consolidated Emission Reporting Rule format requirements. The National Emission Inventory Input Format (NIF), described in an earlier section, is the result of decades of lessons learned in the collection, verification, quality assurance and control, and redistribution of emission inventories from and among the fifty United States. This structure and associated protocol was prompted by the need to have a consistent format for data to allow a single entity (EPA) to process emissions inventory information it received from multiple, individual agencies.

During the initial data submittal requirements of the EPA, each state would provide its inventory in a format developed under its own mechanisms. This format, although easily interpreted by each individual submitter, was unique in some fashion to the State. This made the direct comparison of multiple states' data difficult, at best. Although there were shared elements of each submitter's data, these fields may have been identified by different headers, reported in different units, or been optional for state-specific reporting regulation.

Capitalizing on the common elements of the emissions-related data, EPA took the initiative to create a standardized reporting format that could be shared and to make multiple data sets directly analogous. This standardization not only reduced the burden on EPA's inventory program, but also generated the ability for individual states to compare their data with those of a neighboring or regional state.

Following the development of a consistent format, a single repository of emissions data was established knowing that all EPA collected emissions data could be made available through the site with a consistent set of standards applied to all the information. This immediately made the information worth more as a collection than it was as individual elements. Initial reports developed from these data were static summaries of aggregate information predetermined for ease of distribution.

Currently, EPA is refining an online tool that will allow public access to these nationally developed emissions sets though active data query options available through its website. The

tool, once complete, will promote the local, regional, or national development of information exchange opportunities individualized for each unique user.

Through a very similar process, Environment Canada produces the National Pollutant Release Inventory (NPRI) to distribute information on annual releases to air, water, and land, and off-site transfers for disposal or recycling. This information is also made available in an annual public report issued by Environment Canada and can be accessed and searched through an on-line database.

These projects are in no way unique. In fact, many ventures just like them are in development every day. The opportunity to combine the information from these systems is at our fingertips. It is now up to the user community to make the most of this chance.

5.1 Active Opportunities

Currently, there are multiple efforts underway that individual programs or agencies can immediately learn from and utilize to promote a distributed database shared framework. As examples, two of these projects relevant to the work of the CEC are presented here.

5.1.1 Mexico National Emissions Inventory

The Mexico National Emissions Inventory is the result of significant efforts by many participants from Mexico, the United States, and Canada. A draft inventory and associated report contains preliminary emissions estimates for the six northern Mexican states of Baja California, Sonora, Chihuahua, Coahuila, Nuevo León, and Tamaulipas (ERG, 2003).

A second "draft final" report will contain the interim emissions inventory for the entire country (i.e., 6 northern states and the remaining 25 Mexican states, plus the Federal District). The third and "final" report will contain the final emissions inventory for the entire country as well as estimates at the municipality-level for all sources of air pollution occurring in 1999 and is expected to be released in July 2004.

Because the files developed under this project are expected to be made available in NIF, these data immediately lend themselves to the tools and comparisons available for U.S. data in this same format. Mexican agencies (or other interested parties) can take advantage of ongoing improvements to models, methods, and tools used to analyze, quality assure, and compare emissions from various agencies and see where opportunities for continued improvement or coordination exist.

5.1.2 Canada – U.S. Air Quality Agreement

The purpose of this Agreement is to establish a practical and effective instrument to address shared concerns regarding transboundary air pollution. Using coordinated or cooperative scientific and technical activities, economic research, and information exchange, each Party agreed to establish specific emission reduction or limitation objectives of SO2, NOx, and VOC.

Through this Agreement, the Parties have agreed to exchange, on a regular basis, monitoring, emissions, technologies, measures and mechanisms for controlling emissions, atmospheric processes, and effects of air pollutants as they relate to the transboundary area.

Countless hours and dollars spent on data manipulation and translation could be saved by utilizing a framework of data exchange which would allow the individuals to collect, develop, analyze, and disclose information in its *native format* while immediately making it available and useful to outside agencies.

5.2 Continued Commitment

Developing agencies should be encouraged to work with their larger counterparts to structure cooperative arrangements that improve the cost-effectiveness and quality of environmental support programs. To achieve the agreements, stronger commitment to the concept of shared data networks is required. The commitments should be formalized in documented strategies, operative tools, and outlined plans that are the result of collaborative efforts among participating agencies. To the extent possible, resources should be committed to the development of programs that support the capacity and shared framework building process.

6 Summary

The development of a North American emission inventory that could be used for strategic planning of air quality and defining environmental management capacity building projects is closer today that it has ever been in the past. Coordinated efforts among local, state, provincial, national, and international agencies have led to the sharing of environmental data but at levels limited by current practices. Opportunities currently exist for participating CEC partners to take advantage of consistent mutually relevant activities in the development of an integrated North American emissions inventory and inventory query and distribution tools. These activities could immediately place the three partners on even footing allowing for the review, distribution, and integration of inventory development information and experience.

Through a continued effort to push the envelope of interactive data exchange, the NEISGEI concept could open the doors to simplified, multi-party, cross-border collaboration in integrated emission inventory development and air quality study and management. This web-based global network will have the ability to retrieve and query multiple data nodes of various formats while retaining the transparent elements of individualized access. This network will aide to further the coordinated advancements of data analysis, registration, and environmental collaboration to a level that individualized resources could not match. The first step involving the creation of the network's infrastructure has already begun.

Current coordinated inventory development projects are utilizing the initial framework of NEISGEI and new projects appear to easily lend themselves to inclusion. Opportunities have also presented themselves for immediate development of online distributed database access. The ability of NEISGEI to synchronize regional efforts prior to or during national or international efforts makes the tool portable to all users. The time and money savings that can be achieved in using an *existing* framework of data protocol, formats, and methods identify NEISGEI as a resource that makes current, planned, and yet to be developed projects more cost-effective to implement. Emissions analysts, modelers, and policy-makers stand to benefit from a system that makes exchanging and using these data easier. However, the concept is more than data exchange.

Without a consortium of active participants eager to review, revise, and utilize the framework, the process will falter. Users will be required to promote the usage of the system through work of their own and act as a community to further its scope. New standards will be defined for existing and future data exchange opportunities. More data will be available for studies involving various agencies. Multiple generations of data users and distributors will begin to speak a common language without having to lose their own. Openings for broader environmental-related activities will present themselves and make value added contributions to the science of air quality.

Additionally, this community must commit to capacity building or providing support to those who need it. Technical information and support for data development according to international standards should be readily available to developing countries. In-kind support is made more available when shared technologies are supported by all parties. NEISGEI will provide that

shared support network and allow individuals to build programs from the ground up while benefiting from the more structured programs already in place.

The NEISGEI model is far from complete, but by committing to developing technologies that will create a forum for information, inspiration, and collective understanding of the science of air quality and our environment; we can lead it to its full potential.

7 Bibliography

CEC, 2003. *Availability and Infrastructure of North American Electric Generating Utility Emission Inventories and Opportunities for Future Coordination*. Prepared for the Commission for Environmental Cooperation by Alpine Geophysics, LLC. October 2003.

EPA, 2002. Background on the U.S. Greenhouse Gas Inventory Process. *Quality Assurance / Quality Control and Uncertainty Management Plan for the U.S. Greenhouse Gas Inventory*. U.S. Environmental Protection Agency Office of Atmospheric Programs (6204N) Greenhouse Gas Inventory Program Washington, D.C. 20460 EPA 430-R-02-007A, Version 1.0 June 2002.

EPA, 2003. *NEI Input Format (NIF) Version 3.0 User's Guide Instructions and Conventions of Use*. Released April 2003; Revised November 2003 (Errata Included) http://www.epa.gov/ttn/chief/nif/index.html

ERG. 2003. *Mexico National Emissions Inventory, 1999—Draft.* Prepared for the Secretariat of the Environment and Natural Resources and the National Institute of Ecology of Mexico, United States Environmental Protection Agency, Western Governors' Association, and the North American Commission for Environmental Cooperation by ERG. July 2003.

Falke, S.R, 2002. *Environmental Data: Finding It, Sharing It, and Using It,* Journal of Urban Technology, 9(2), 111-124.

Falke, S.R., G. Stella, and T. Keating, 2004. *Demonstration of a Distributed Emissions Inventory using Web Technologies.* Presented to the Commission for Environmental Cooperation and U.S. EPA in Hull, Ontario. February 27, 2004. http://capita.wustl.edu/NAMEN/DemoPresFeb27.pdf

Hemming, B., S. Falke, and T. Keating. 2003. *Networked Environmental Information Systems for Global Emissions Inventories (NEISGEI).* Presented at NARSTO Workshop on Innovative Methods for Emission Inventories, Austin, TX. October 2003.

Appendix A

Additional Distributed Information System Initiatives

Advancements in information technology have spawned a number of efforts in community-based and integrated environmental monitoring, analysis and management activities. Most are in their initial stages and require extensive planning, research and management over the next few years before they become stable distributed networks striving toward shared objectives.

Section 3.1 outlined some of the efforts directly relevant to air quality management. In this appendix, we outline other efforts that serve as examples of distributed networks to be considered in developing an integrated North American emissions inventory.

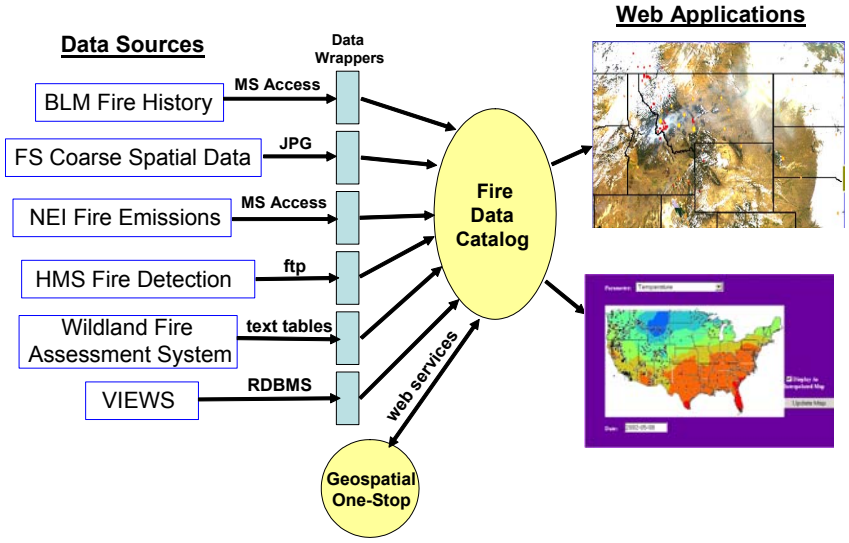
Fire, Smoke and Air Quality Network Pilot Project

Forest fire scientists and air quality researchers face the daunting challenge of working efficiently and effectively with a variety of data sources, data formats, analysis methods and information systems in generating a comprehensive understanding of the state of the environment. The multidisciplinary nature of forest fires and their air quality impacts require researchers in one discipline to rely on data and information from another discipline. A pilot project supported by the NSF Digital Government Program and U.S. EPA has conducted preliminary research into establishing a software infrastructure for integrating distributed fire related data.

Classes of data wrappers that homogenize data formats from multiple sources have been developed to accommodate common fire emissions data types and their particular data access requirements. Data types include monitoring network data, gridded model output, emissions inventory databases and satellite imagery. Each class includes a type of data access, including complete one-time ingest, dynamic caching (such as of daily fire location data that are only temporarily stored on a remote server), dynamic data access (such as to a ftp server with ASCII table files updated daily), direct access to a relational database, and web service access.

In progressing toward an integrated fire and air quality emissions data network, this project has leveraged research results from another project called DataFed.net, an infrastructure that supports collaborative atmospheric data sharing and processing services (www.datafed.net). The fire emissions data were registered datasets using the cataloging services in DataFed.net where the registered data access instructions can be interpreted for browsing and visualization. DataFed.net includes interfaces for rendering data "views" including maps, time series, and tables. The views are each created as their own web service thereby allowing them to be used in custom applications with standard web programming languages (such as JavaScript and ASP).

The figure below shows the components and data flow for the fire related data network that accesses distributed data sources, catalogs them, and provides user access through web interfaces in the form of maps and time series. New web services tools have been developed as part of this project to extend the capabilities of DataFed.net, including the ability to dynamically manipulate the data, such as by interpolating a point data set to a continuous surface grid. The project's website is: http://capita.wustl.edu/FSAN



Data Flow and Component Diagram for the Prototype Fire and Air Quality Data Network

Semiautomoated Integration of Heterogeneous Databases

The University of Southern California's Information Sciences Institute, with support from NSF's Digital Government Program and U.S. EPA is advancing technologies for automating the integration of heterogeneous databases. Current practices require substantial manual effort in translating and matching the fields from one database to another so that they data can meaningfully be integrated. The project is focused on air emissions data from State of California Air Resources Board, and Santa Barbara County Air Pollution Control District. Its objectives are to develop a suite of tools to map, transform, and re-aggregate data from one schema to another by leveraging techniques of statistical machine translation and web services. The project's website is: http://www.neisgei.org/epa-air/

Linked Environments for Atmospheric Discovery

The Linked Environments for Atmospheric Discovery (LEAD) is a recently funded NSF funded project with the objective of developing the components needed for a complete infrastructure for dynamically accessing and using meteorological data. The project proposes to provide data access and analysis services that can be linked to provide data in a format most relevant to the end user in real time. The client base consists of researchers, managers, and the public. The site's website is: http://lead.ou.edu/

Integrated Ocean Observing System

The Integrated Ocean Observing System is an US effort in linking monitoring and data sources for an integrated network of ocean related monitoring, assessment, and communication. The motivation for the integrated network is the need for the most complete data and analysis in support of regional and global science and policy issues. In fact, the US initiate is designed to operate within a global monitoring and assessment system in the future. The US effort is just getting underway is currently focusing on the development of an operational network of existing

monitoring and data systems before entering a future stage of developing new capabilities that address data gaps. The project's website is: http://www.ocean.us/

National Virtual Observatory

Now entering its third year, the national virtual observatory is a NSF funded project that aims to build the foundation for an integrated network of astronomers sharing their data and analysis tools. Like other data analysis fields, astronomy is inundated with data as it is generated at rates that exceed the capabilities for analyzing them. The virtual observatory is a research and development effort based on data standards and distributed computing technology to advance the astronomical community to a new generation of integrated analysis and tool building.

The national virtual observatory project has demonstrated the benefits of a shared network in prototype applications. To date three prototype applications have been built that demonstrate the potential capabilities for accessing, processing and analyzing astronomical observations through web interfaces using data collections and analysis tools residing on distributed servers. More information about the virtual observatory project can be found at the project's website: http://www.us-vo.org/