Data Management Challenges in Developing a Network of Distributed North American Emissions Databases

Stefan Falke, Washington University in St. Louis

Gregory Stella, Alpine Geophysics, LLC

Terry Keating, U.S. EPA - Office of Air & Radiation

Brooke Hemming, U.S. EPA – Office of Research & Development

Background

Air pollutant emission inventories for the US, Canada, and Mexico are compiled, stored and disseminated using different methods

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

In support of this longer term goal, the Commission on Environmental Cooperation (CEC) and the US EPA have initiated a project to develop a prototype web tool for enabling uniform access to distributed emissions data from North American electricity generating power plants.

Distributed Data and Management Networks

Advances in information science and technology are driving the trend toward distributed networks and virtual communities for science and management.

Cyberinfrastructure

NSF’s initiative to apply new IT to building new ways of conducting collaborative research

http://www.cise.nsf.gov/sci/reports/toc.cfm

Earth Observation Summit

International effort to build comprehensive, coordinated, and sustained Earth observation systems

http://www.earthobservationsummit.gov

Ecoinformatics

EPA’s vision for national and international cooperation in data and technology development

http://oaspub.epa.gov/sor/user_conference$.startup

Integrated Ocean Observing System

International network of ocean related monitoring, assessment, and communication

http://www.ocean.us/

Linked Environments for Atmospheric Discovery

Network of high-performance computers and software to gain new insights into weather

http://lead.ou.edu/

Virtual Observatory

Network for astronomical data sharing and distributed analysis

http://www.us-vo.org/

Emissions Community Collaborative Activities

- **NIF Data standards**
  - Standard format and submission
  - NEI XML schema
- **Environmental Information Exchange Network**
  - Network linking EPA, States, and other partners through the Internet and standardized data formats
- **Facility Registry System**
  - Standard facility codes and locations
- **Data Sharing Efforts**
  - States, Tribes, Local agencies, RPOs
  - North America

NEISGEI: *Networked Environmental Information Systems for Global Emissions Inventories*

…is both a conceptual framework and implementation effort for the development of a fully integrated, distributed air emissions inventory – and the foundation for an all-media environmental information network

- ❖ Tie together data at all spatial and temporal scales using emerging distributed database technologies
- ❖ Provide shared, online tools for processing and analysis
- ❖ Provide for the seamless merging, manipulation and analysis of Internet accessible air quality-relevant data through the development of emerging Internet-oriented technologies
- ❖ Make use of existing resources – partner/link with others and their related projects
- ❖ Build a broad-based air quality user community: scientists, regulators, policy analysts and the public
- ❖ Create the network and toolkit via modular projects

Ongoing Efforts:

NSF-EPA Digital Government Funded Projects:

The California Air Resources Network

Future Effort:

EPA OAR RFA on Distributed Air Quality Data in Support of NEISGEI

Current Project Objectives

- Recommend and demonstrate to the CEC approaches for the comparability of techniques and methodologies for data gathering and analysis, data management, and electronic data communications for promoting access to publicly available electric utility emissions

- Identify, collect, and review existing sources of electric generating utility (EGU) emissions and activity databases, and provide a summary of the state-of-science

- Build a prototype web browser tool to query, retrieve, and explore emissions data from heterogeneous databases

The project's focus is on criteria pollutants and toxics because of their availability and accessibility.

Guiding Principles

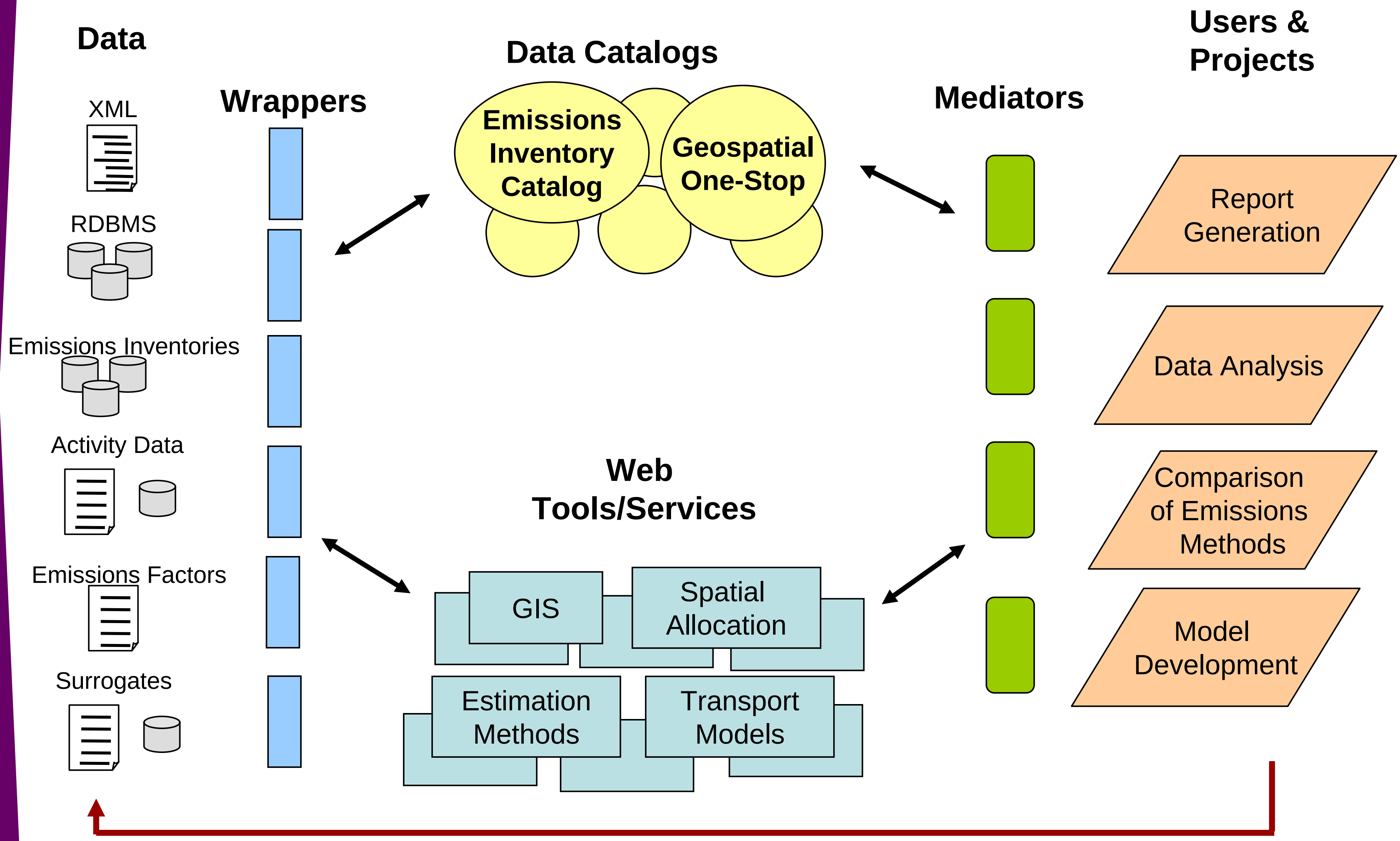
- **Distributed.** The data sources remain distributed and in the control of their providers. The data are dynamically accessed through the internet rather than through a central repository.

- **Non-intrusive**. Data providers are more likely to participate if joining an integrated network does not impose new or additional burden on them.

- **Transparent.** The distributed data should appear to originate from a single database to the end user. One stop shopping and one interface to multiple data sets are desired without required special software or download on the user’s computer.

- **Flexible/Extendable**. An emissions network should be designed with the ability to easily incorporate new data and tools from new nodes joining the network so that they can be integrated with existing data and tools.

Envisioned Emissions Community Resource of Data & Tools



Process of Building a Web Application

- ✓ Identify and access relevant data (build wrappers)
- ✓ Build relational database to temporarily store the data that are not accessible in a distributed manner
- ✓ Acquire authorization and access to those data that are dynamically accessible through internet interfaces
- ✓ Create field name “mappings” among datasets
- ✓ Identify available web technologies for building a distributed emissions tool
- ✓ Develop new components necessary for the prototype
- ✓ Build web tool prototype for demonstrating the feasibility of exploring emissions data

Available Internet Accessible Emissions Data

Data Source	Time Coverage	Pollutants	Reporting Level
NEI (US)	1985-1999 (criteria) 1996-1999 (HAPs)	NOx, SO2, CO, PM, VOC, HAPs	Boiler
eGrid (US)	1996-2000	NOx, SO2, CO2, Mercury	Boiler & Generator
Clean Air Markets (US)	1980, 1985, 1988-1999	NOx, SO2, CO2	Generator
NPRI (Canada)	1994-2001	HAPs (Criteria starting in 2002)	Facility

These are publicly available, on-line accessible emissions data. Other data resources are available, but at this time only in hard copy form and therefore not usable in demonstrating distributed database concepts.

- 1999 BRAVO Mexican emissions data were obtained in electronic format files
- 2000 Canadian Mercury Emissions were obtained in electronic format files

Emissions Data Characteristics

NPRI

Web browser query
Web map server
Ongoing database structure upgrade

NEON NEI EMISSIONS ON THE NET

Web browser query
Remotely accessible using SecuRemote
Not yet publicly accessible

Clean Air Markets

Web browser query
Remotely accessible using SecuRemote
Oracle database

eGRID

Downloadable Excel Spreadsheets
Plans for a dynamic web system were shelved

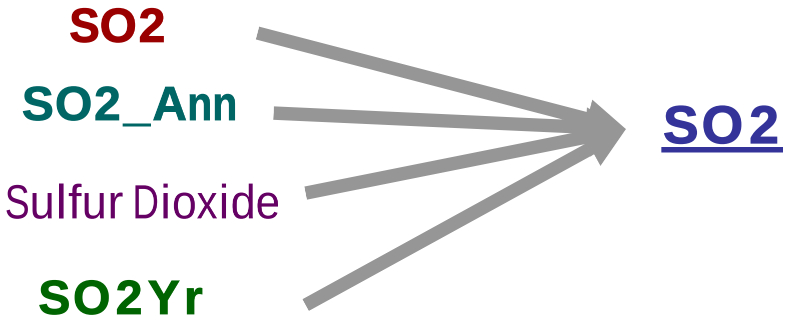
Mapping Multiple Database Fields

Emissions inventories are based on different underlying data models.

Each inventory uses a uniquely defined set of field names. However, many of these field names are similar to (or their content is similar to) fields in another country’s inventory.

Some of the key relationships among the inventories have been captured by developing a “mapping” among fields.

These mappings provide a set of connections that can subsequently be applied to automated query and integration of data from multiple inventories.



SO2

SO2_Ann

Sulfur Dioxide

SO2Yr

SO2

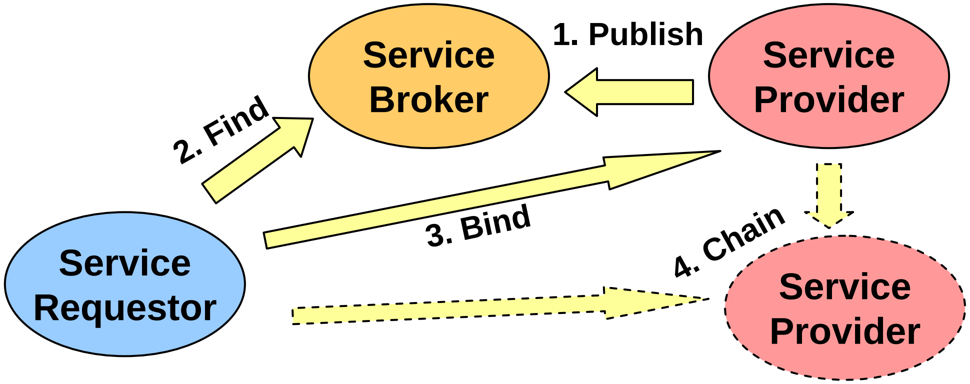
Web Services

Web services are generally defined as software applications invoked over the Web with eXtensible Markup Language (XML)-based standards. They are self-contained and use XML for describing themselves and communicating with other web resources, thereby allowing them to be reused in a variety of independent applications.

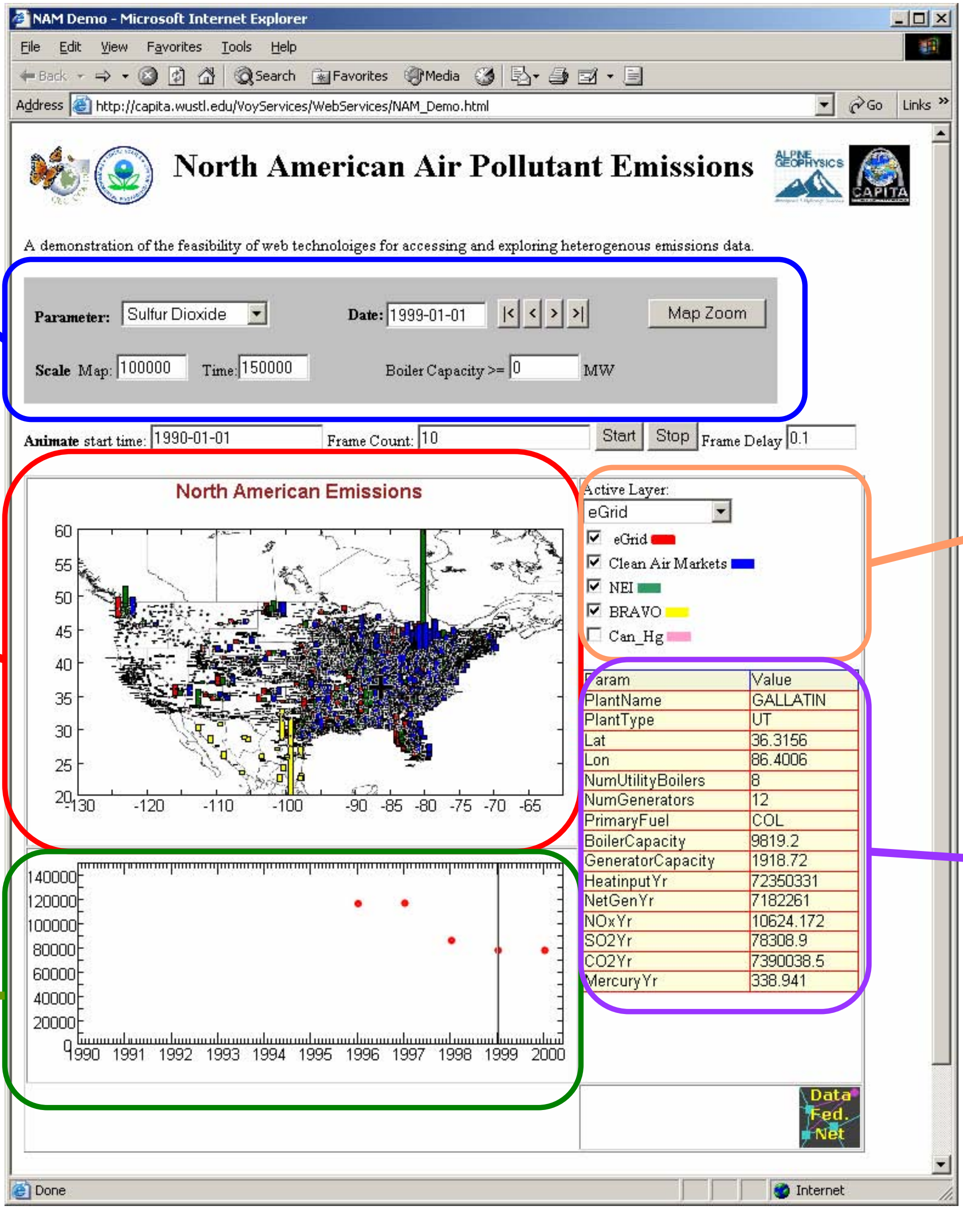
Because they are designed to be independent of any particular database platform, they are ideally suited for building a distributed database and tools network.

Many of the analysis and processing tools used by the air emissions management community could benefit from web service technology. ***Not only can their data be shared but heterogeneous, distributed tools that operate on that data can be shared as well.***

A longer term vision for web services is to be able to “orchestrate” or “chain” multiple services from multiple providers so that new web applications can be constructed.



Components of an Emissions Tool



North American Air Pollutant Emissions

A demonstration of the feasibility of web technoloiges for accessing and exploring heterogenous emissions data.

Parameter: Sulfur Dioxide **Date:** 1999-01-01 |< < > >|

Scale Map: 100000 Time: 150000 Boiler Capacity >= 0 MW

Animate start time: 1990-01-01 Frame Count: 10 Frame Delay 0.1

Active Layer: eGrid

☑ eGrid 🟥
☑ Clean Air Markets 🟦
☑ NEI 🟩
☑ BRAVO 🟨
□ Can_Hg ▨

Param	Value
PlantName	GALLATIN
PlantType	UT
Lat	36.3156
Lon	86.4006
NumUtilityBoilers	8
NumGenerators	12
PrimaryFuel	COL
BoilerCapacity	9819.2
GeneratorCapacity	1918.72
HeatinputYr	72350331
NetGenYr	7182261
NOxYr	10624.172
SO2Yr	78308.9
CO2Yr	7390038.5
MercuryYr	338.941

Control Panel
controls the views

Map View
displays tons of emissions
as proportional bars

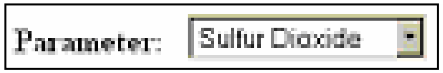
Time View
displays a time series of
emissions for a selected facility

Data Layer Control
controls the layers to
display in the map and
which layer is active
(displayed in the
time and table views)

Table View
displays the data record
for a selected facility

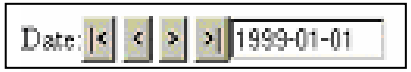
http://capita.wustl.edu/NAMEN/

Embedded Images and Controllers in a Web Page



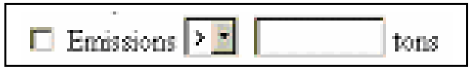
Parameter: Sulfur Dioxide

Parameter Controller



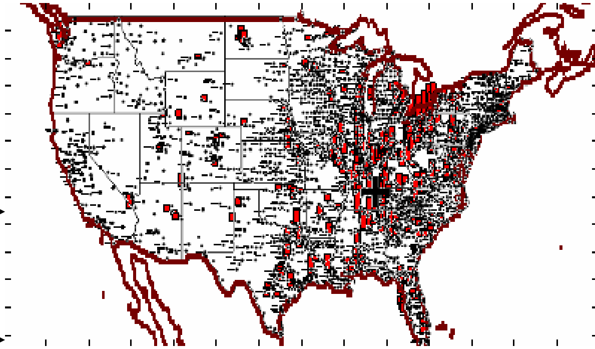
Date: |< < > >| 1999-01-01

Date Controller



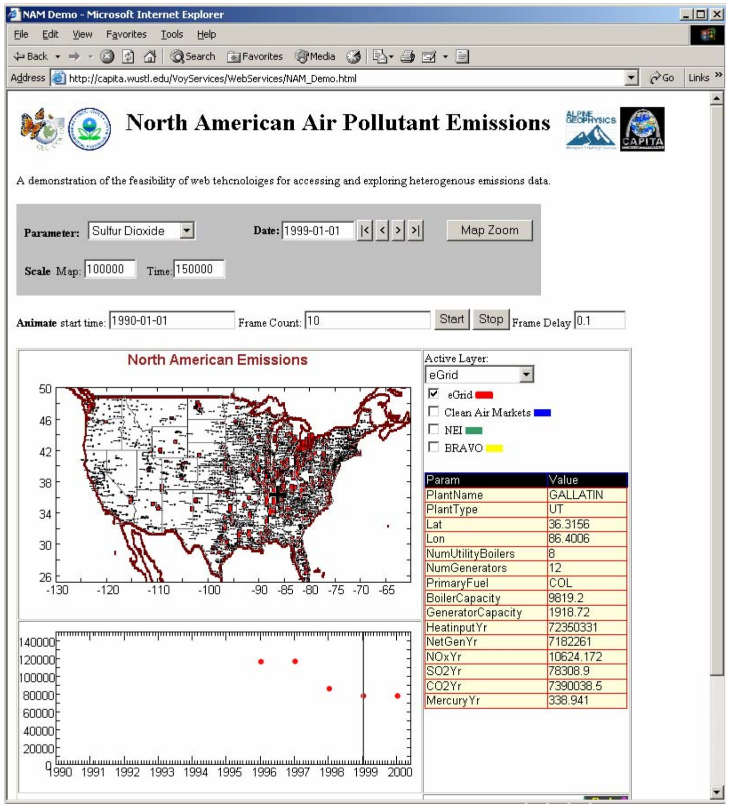
□ Emissions > [ ] tons

Query Controller



The controllers and map image view can be linked and assembled in a web page. Changing the settings of a controller changes the URL of the map image and updates the web page.

The web page can be constructed using standard web application programming languages, such as JavaScript and ASP.



Address: http://capita.wustl.edu/VoyServices/WebServices/NAM_Demo.html

North American Air Pollutant Emissions

A demonstration of the feasibility of web tehcnoloiges for accessing and exploring heterogenous emissions data.

Parameter: Sulfur Dioxide **Date:** 1999-01-01 |< < > >| Map Zoom

Scale Map: 100000 Time: 150000

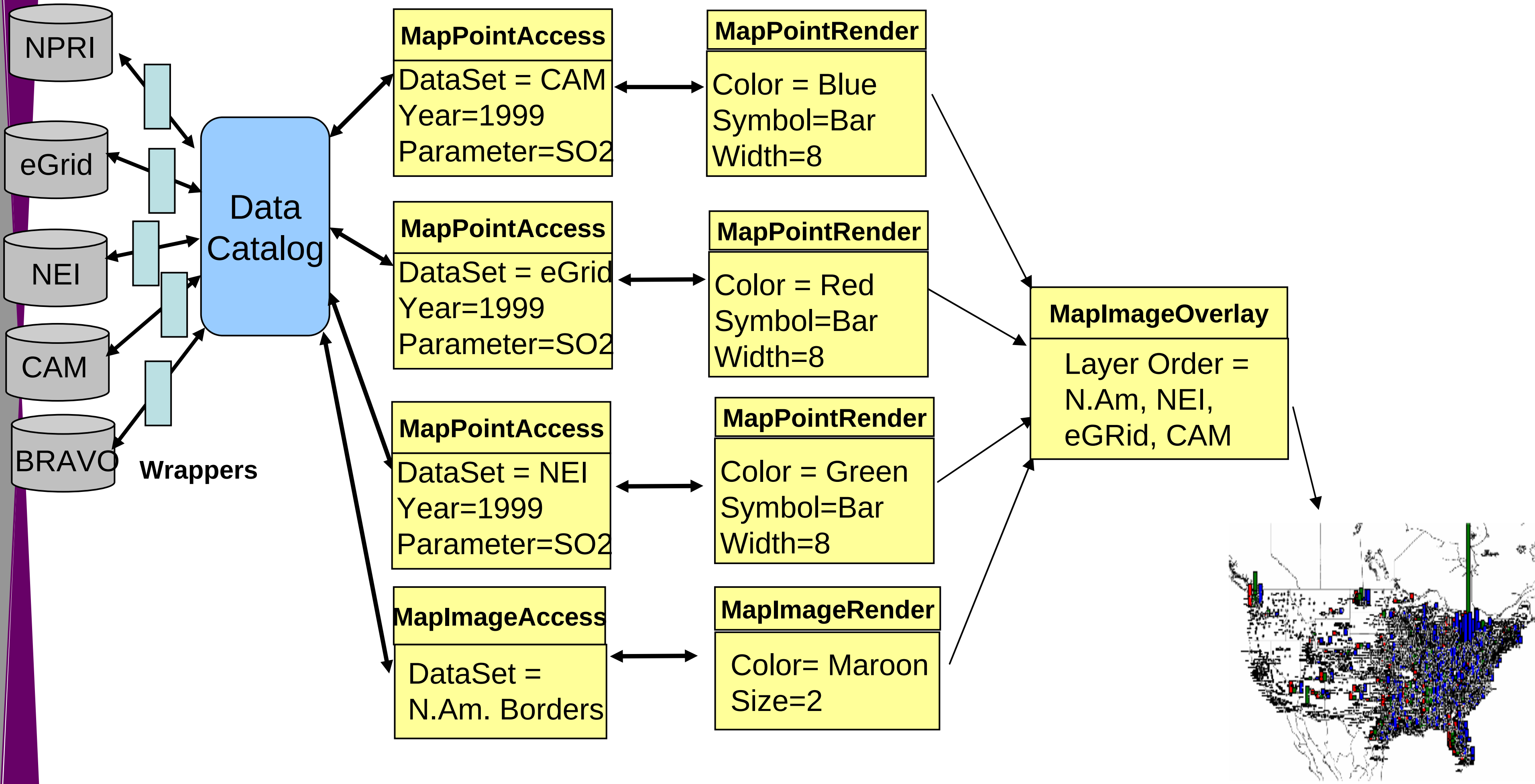
Animate start time: 1990-01-01 Frame Count: 10 Start Stop Frame Delay 0.1

Active Layer: eGrid

☑ eGrid 🟥
□ Clean Air Markets 🟦
□ NEI 🟩
□ BRAVO 🟨

Param	Value
PlantName	GALLATIN
PlantType	UT
Lat	36.3156
Lon	86.4006
NumUtilityBoilers	8
NumGenerators	12
PrimaryFuel	COL
BoilerCapacity	9819.2
GeneratorCapacity	1918.72
HeatinputYr	72350331
NetGenYr	7182261
NOxYr	10624.172
SO2Yr	78308.9
CO2Yr	7390038.5
MercuryYr	338.941

North American Emissions Demonstration Data Flow



NPRI

eGrid

NEI

CAM

BRAVO

Wrappers

Data Catalog

MapPointAccess
DataSet = CAM
Year=1999
Parameter=SO2

MapPointAccess
DataSet = eGrid
Year=1999
Parameter=SO2

MapPointAccess
DataSet = NEI
Year=1999
Parameter=SO2

MapImageAccess
DataSet =
N.Am. Borders

MapPointRender
Color = Blue
Symbol=Bar
Width=8

MapPointRender
Color = Red
Symbol=Bar
Width=8

MapPointRender
Color = Green
Symbol=Bar
Width=8

MapImageRender
Color= Maroon
Size=2

MapImageOverlay
Layer Order =
N.Am, NEI,
eGRid, CAM

Name
Settings

= web service

The settings of each web service can be changed by the user, creating a dynamic application

Some of the Remaining Challenges

Organizational

- It is imperative that data providers see a benefit, or return on investment, from joining a network.
- Appropriate acknowledgement and recognition of the data providers once the data become part of a network must be maintained
- How do we balance the benefit of greater data use with the risk of data misuse?

Technical

- Security restrictions and systems designed for desktop-only access limit the ability to implement networks
- Consensus derived standards and protocols are still missing in many aspects of distributed computing
- Accessing large datasets, such as multi-dimensional national emissions inventories with thousands of emission point locations, remains cumbersome for efficient, dynamic user interaction through web browsers.

Future Goals

- More complete access to distributed datasets - A process for creating trusted provider-user agreements that would help address issues of security and data misuse.

- More comprehensive content – Current efforts in creating distributed information systems will make a diverse set of data and tools available that could spark additional interest in the technology’s potential;

- Integration – Linking current distributed database efforts together with one another will create a broad base of data and tools and will serve as important examples in testing and demonstrating the effectiveness of distributed databases;

- Metadata - More complete description information about emissions databases would help in relating heterogeneous data. Efforts to use FGDC metadata and the development of standard data catalogs, such as Geo Spatial-One Top are beginning to address this.

The Eight “fallacies” of Distributed Computing

Essentially everyone, when they first build a distributed application, makes the following eight assumptions. All prove to be false in the long run and all cause big trouble and painful learning experiences.

1. The network is reliable
2. Latency is zero
3. Bandwidth is infinite
4. The network is secure
5. Topology doesn't change
6. There is one administrator
7. Transport cost is zero
8. The network is homogeneous

-Peter Deutch
(http://www.aladdin.com/users/ghost/)

These are challenges we will all continue to struggle with and overcome