Cyberinfrastructure for Emissions Data & Tools

Stefan Falke
Washington University in St. Louis
stefan@wustl.edu

Gregory Stella
Alpine Geophysics, LLC

Terry Keating
US EPA - Office of Air & Radiation

Brooke Hemming
US EPA – Office of Research & Development

ABSTRACT

The Networked Environmental Information System for Global Emissions Inventories (NEISGEI, pronounced "nice-guy,") is an EPA supported initiative to develop a web-based global air emissions inventory network. Part of this effort is the development of a Web portal that provides description and access to 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. This paper presents the portal infrastructure and select data and tools available through it.

The NEISGEI portal is designed so that emissions community members can contribute to its growth and evolution by posting data, reports, tools, and other content, and contributing to discussion forums, announcements, and feedback on web portal design and content. Primary design considerations for the portal are the use of accepted web and data standards, the reuse of existing web infrastructure, and that its content is openly available for use in other applications or portals. Initial data sets "registered" with the portal include national and regional emissions inventories and activity data. Web browser tools have been developed and/or applied for visualizing and comparing data in maps, time series, and tables. Examples using emissions inventory and fire occurrence data are presented.

INTRODUCTION

Data and information issues encountered by air emissions researchers and managers are representative of those faced by the research and management community as a whole; vast quantities of heterogeneous data, a wide range of tools, many diverse organizations, and no straightforward way to pull the relevant pieces together. There is an important need to develop air quality information networks that foster the sharing and integration of data and tools while imposing minimal burden on users and contributors. New terms, such as service oriented architecture, web services, and cyberinfrastructure, have been added to our vocabulary to capture information technologies that may address these issues. They represent the application of information sciences and technologies to support distributed and interoperable data and tools for supporting science and engineering research and management.

The emissions community is looking to advances in information technology to help achieve its next generation of emission inventory systems. A recent NARSTO Emissions report highlights new database management approaches and information systems in envisioning a future inventory that "includes all significant emissions from all sources, time periods and areas, with quantified uncertainties, and timely accessibility. From this vision, the overall goal is to make inventories complete, accurate, timely, transparent, and affordable."[1] The hope is that cyberinfrastructure can make multi-spatial, temporal and composition scale air emissions data and tools easier to find, use and integrate. The paper addresses two components of an emissions cyberinfrastructure, 1) a web portal to organize, describe and connect to emissions inventories, models and analysis tools and 2) adding web tools and applications that allow data access, visualization, analysis, and dissemination of emissions related data.

CYBERINFRASTUCTURE

Cyberinfrastructure is defined as the information sciences and technologies, including distributed computer, information and communication technologies, used to build new types of scientific and engineering knowledge environments with the goal of pursuing research more effectively and efficiently. A recent National Science Foundation report on cyberinfrastructure examined the future of research and concluded that "contemporary projects require effective federation of both distributed resources (data and facilities) and distributed, multidisciplinary expertise and cyberinfrastructure is a key to making this possible."[2] Aided by new information science tools and computer science networks, scientists are poised to exchange information more effectively, integrate data and analyses more efficiently, and interact more actively. Over the past few years, workgroups have formed to study the potential of cyberinfrastructure in a variety of science and engineering domains, including environmental sciences, atmospheric sciences, and environmental monitoring. The emissions community is favorably situated to reap the benefits of cyberinfrastructure because of its rich databases, the transboundary nature of air quality issues that require collaboration among states, regions, and countries, and the relevance of satellite imagery to understanding emissions.

Service Oriented Architecture

A first step in building a cyberinfrastructure is making data and tools accessible through the Web. Many emission databases are already accessible through Internet-based methods either through direct data-file download or web query tools. The query systems allow users to filter and access data at multiple levels of detail. These systems meet the needs of individual end users who log in to the online system, complete forms for defining their query, and then view the results in tables/graphics or download the data for use in other tools. While these systems serve the individual user, they do not easily come together to form a distributed emission inventory network where automated computer-to-computer interaction among services is possible. This computer level interaction is possible through web standards and services. The availability of services that adhere to accepted XML-based standards for describing themselves and their interaction with other services provides a foundation for a service oriented architecture (SOA). The benefits of SOA are loosely coupled and interoperable services that are independent on any single platform, programming language or host.

In the SOA based on web services approach, mediators serve the role of brokers, providing data providers with web service interfaces onto the network and 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 complete off-the shelf software solution. However, many required components are considered standards in web programming applications and therefore make it possible to create an operational data web service network.

NEISGEI

Networked Environmental Information Systems for Global Emissions Inventories (NEISGEI, pronounced "nice-guy,") is an initiative to develop a web-based global air emissions inventory network that uses a Web portal to provide 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.

Ideally, emissions tools and the data sets would be available to a wide variety of users who would be able to access, integrate, compare, and generate emissions inventories at different spatial and temporal scales for assessing intercontinental transport or any other air quality management challenge. A complete service oriented architecture for emissions inventories is a long term goal that requires both continued development in information science and technology and next generation development of emissions inventories that support these new approaches and technologies. In the meantime, the approach to integrating distributed emissions data and tools can be achieved using a combination of web services where they are available and manual integration where they are not.

In moving toward this goal, NEISGEI is based on principles of service oriented architecture, uses the DataFed environment for making data available as services and building analysis tools, and uses a web portal to catalog and describe the data, tools, and organizations from which emissions applications are built.

NEISGEI Portal

The term 'web portal' has been used in a variety of contexts making it difficult to provide a comprehensive definition but, for our purposes, the most pertinent role of portals is as aggregators of distributed content driven by their user community. We envision the NEISGEI web portal as a community resource providing an array of content and services that allows air quality scientists and managers a place to explore and share emissions data, tools and ideas.

The NEISGEI web portal is more than static content supplemented with a collection of web links. It offers interactive tools that allow users to expand its content and that facilitate community building through the use of online dialogue. The design philosophy of the NEISGEI web portal is to maintain a rich set of finable content that are distributed over the Web and to enrich those content with descriptions of the contexts in which they are used. The intent is not to build a central repository of the data, tools, reports, and models themselves, but rather to primarily store metadata (data about data) augmented with community supplied annotations.

The portal includes many of the common components found in content management systems, such as document libraries, event lists, and discussion forums. It also provides mechanisms for dynamically adding "resources", such as data, tools, reports, information systems, and websites (Figure 1). These resources are registered with the portal as a description of the resource and where to find it. The registered portal entry is available for annotation by other users who may expand the resource's description, pose questions, and provide examples where they have used the resource, or cite other related resources.

Figure 1. NEISGEI portal front page. Accessible at http://www.neisgei.org.
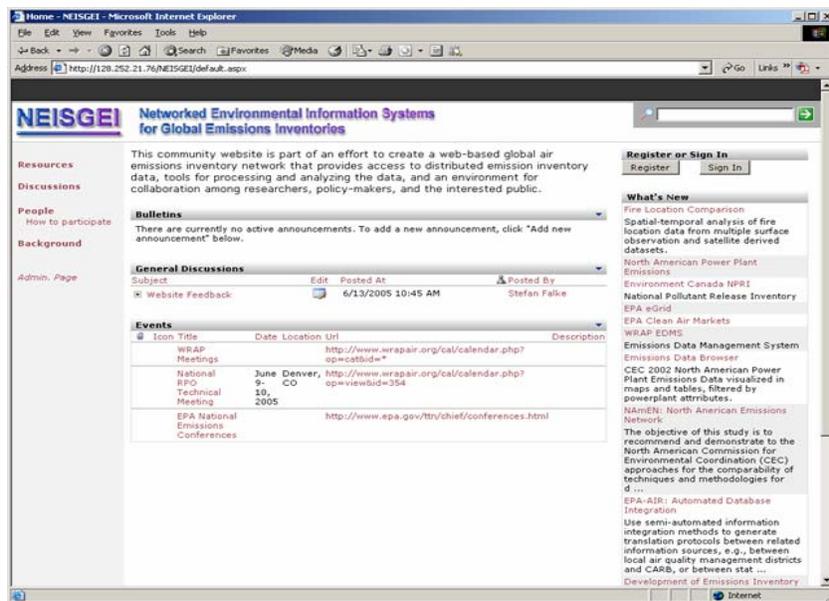
Portals are designed and constructed through a layout of portlets, modular components that provide a focused service or function with a dedicated graphical user interface. A portal page is created by assembling a set of portlets, thereby offering dynamic, reconfigurable content.

A number of commercial and open source portal building software are available. The NEISGEI Portal is built using LifeRay, an open source portal platform that is flexible and easily customized (www.liferay.com). A primary consideration for using LifeRay is its commitment to portlet standards. The Java Portlet Specification (JSR-168), defines a standard application programming interface for J2EE (Java) based portal platforms. A portal developer can find collections of JSR-168 portlets on the web and simply embed them in their portal. Web Services for Remote Portlets (WSRP) is an XML and web services specification that allows the remote sharing of portlets. WSRP allows portlets running on one portal to be displayed in another portal without requiring any additional programming by the portal developers. To the end-user, it appears that the portlets are running locally within their portal, when they may actually reside in remotely-running portals. Whereas web services allow interoperability at the service, or process, level, WSRP adds interoperability at the graphical user interface, or presentation layer, level. The plug-and-play nature of standards based portlets enhances a portal's capabilities without additional programming or development on the portal side.

The hope is that by adhering to these standards, the NEISGEI portal will eventually be able to share content with other related portals, such as the EPA agency and science portals. The NEISGEI portal is accessible through http://www.neisgei.org.

DATAFED

Many of the data and tools available through the NEISGEI portal are found and/or developed within the DataFed framework. The federated data system, DataFed (http://datafed.net), aims to support air quality management and science by facilitating more effective use of air quality-relevant data. DataFed is a web infrastructure that provides the foundation for accessing distributed air quality data and for processing and visualizing these data through web services.[3] The key role of DataFed is to mediate the flow of data between data providers and users. Specifically, it (1) facilitates registration of distributed data; (2) homogenizes, on the fly, all the data into a physically-based global data model (space and time); (3) supports interoperability through the use of standards-based protocols (OGC, XML-SOAP); and (4) provides a set of basic tools for data exploration and analysis. DataFed provides mediator software for creating "views" of data, including maps, time series, and tables, that are distributed among multiple

web servers. The views are created using web services thereby allowing them to be used and reused in custom applications with standard web programming languages. Over 50 distributed air quality-relevant datasets are accessible through DataFed including North American emissions inventories and data related to fire emissions.

Geospatial Standards

One of DataFed's strengths in serving as a mediator between data provider and user is its implementation of geospatial standards that promote the exchange of data for geospatial applications. Geographic information systems (GIS) and geospatial analysis are widely used in analyzing, visualizing and sharing emissions data, models, and analysis. Standards for finding, accessing, portraying, and processing geospatial data are defined by the Open Geospatial Consortium (OGC).[4] The most established OGC specification is the *Web Map Server (WMS)* for exchanging map images, but the *Web Feature Service (WFS)* and *Web Coverage Service (WCS)* for accessing databases are gaining wider implementation. WFS retrieves discrete feature data encoded in Geography Markup Language (GML) format while WCS allows access to multidimensional data that represent coverages, such as grids. While these standards are based on the geospatial domain, many are designed to be extended to support non-geographic data "dimensions," such as time and the many other dimension tables found in emissions inventories. Some web information systems are making air quality data accessible through OGC interfaces. On the client side, many map browsers and other visualization tools support OGC specifications, thereby allowing interoperability between the data and the applications that use those data.

EXAMPLE APPLICATIONS

Three examples of connecting emissions data with services in a cyberinfrastructure are presented. They highlight not only new applications built using the cyberinfrastructure but also the reuse of data and tools. In all three examples, neither the data nor the tools that served as building blocks for the applications were specifically designed for these particular purposes but openness and standards made their reuse possible.

North American CEC Emissions

The North American Commission on Environmental Cooperation published a report on fossil fuel–fired power plant emissions across North America for 2002.[5] The report includes North American maps for pollutant emissions (SO_2, NO_x, Hg, and CO_2) and tables for a handful of power plant's listing their specifications (location, capacity, fuel type, etc.). These maps and tables are effective in quickly summarizing, illustrating and explaining the emissions data. In addition to providing the report, CEC also made its emission inventory available through Microsoft Excel tables. We registered the CEC emissions inventory in DataFed, making it available to DataFed tools and its development environment and accessible more broadly through geospatial interfaces (e.g., Open Geospatial Consortium specifications).

A web application was built within the DataFed environment that displayed the emissions data on a map and table (Figure 2). The map displays emissions or emission rates for SO_2, NO_x, Hg, and CO_2 as circles proportional in size to the emissions or rates. The circles are color coded according to fuel type, capacity, or generation. The user can select the combination or emissions/rates and color code type to generate the map on the fly. The map viewer includes standard zoom and pan tools for honing in on areas of interest. Clicking on a particular power plant in the map, updates a table with the information for that plant, including the plant's name, code, fuel type, capacity, generation, emissions, and emission rates. The application is accessible at:
http://webapps.datafed.net/datafed.aspx?page=PowerPlant_Emissions

Figure 2. North American powerplant emissions data through a traditional report and standard based web application (http://webapps.datafed.net/datafed.aspx?page=PowerPlant_Emissions).
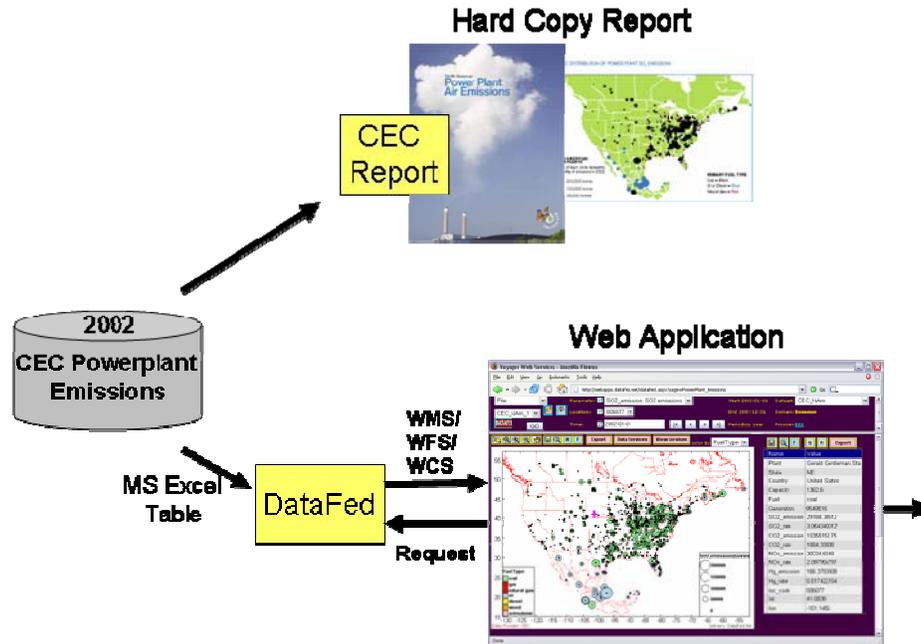

The web application serves as a supplement to the hard copy report by offering the ability to browse the 2002 CEC emissions inventory through maps and tables that can be dynamically changed based on user specified emissions type, area, and plant of interest.

Satellite Fire Locations and Smoke

Satellite derived fire location pixels are collected by multiple satellite sensors, including MODIS, GOES and AVHRR. NOAA's Hazard Mapping System (http://www.ssd.noaa.gov/PS/FIRE/) processes the fire detection data from these satellites by having an analyst identify false detections or missed fires and adjusting the fire pixel datasets accordingly.[6] The satellite analyst also analyzed the MODIS images for visible smoke plumes and creates a hand-drawn outline encompassing the boundaries of the plumes. The processed fire detection and derived smoke plume datasets are served through ftp servers and ESRI ArcIMS map servers. Two ArcIMS applications provide visualization of the fire detections, one for the latest fires and another that provides browsing older fires.

The ArcIMS applications are useful for assessing areas impacted by fires and the duration of their impact. The applications include related datasets for overlay in the map viewer. A limitation of ArcIMS applications is the inability for a user to dynamically add other, external datasets to the map that are not already embedded in the ArcIMS application. By NOAA providing the fire detection and smoke plume data as OGC Web Map Services (WMS), other web client applications are able to access and view the data with other data. Figure 3 illustrates this networked information flow that chains the satellite data sources with NOAA HMS and DataFed. NOAA HMS adds value to the data by processing satellite derived fire products and exposing those data through web interfaces while DataFed extends the HMS products to a framework where they can be combined with other air quality related data and tools.

Figure 3. Pathways involving multiple organization in web mapping fire detection data. NOAA HMS ArcIMS applications are accessible at: http://map.ngdc.noaa.gov/website/firedetects/viewer.htm and http://www.firedetect.noaa.gov/viewer.htm. DataFed access to the HMS fire products: http://webapps.datafed.net/datafed.aspx?dataset_abbr=GDSG_FIRE
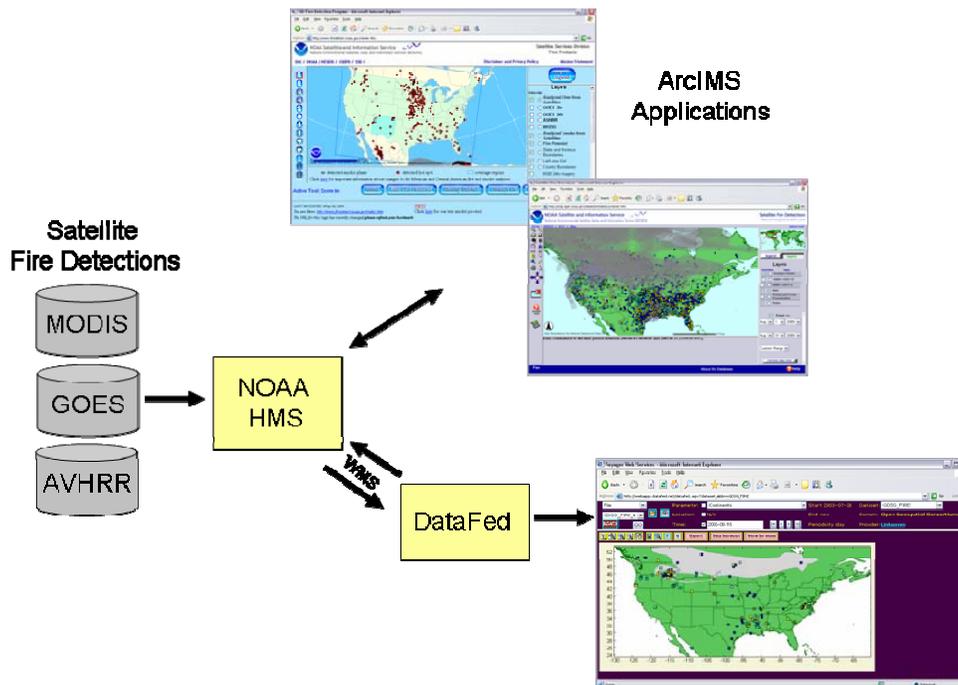

NASA World Wind and Google Earth

Perhaps no other application has thrust web mapping into the public spotlight more than the release of Google's mapping products, Google Maps and Google Earth. The Google applications not only give users the ability to zoom into a satellite image to view their own backyard but also tools to create their own mapping applications with their own data without having to be involved in any development. NASA's World Wind is an application similar to Google Earth in that it offers a three dimensional visualization environment but is aimed at the earth science community. These and other related mapping products have spawned customized mapping applications in every discipline imaginable.

While World Wind and Google Earth are not strictly web services, they are tools openly available with open interfaces that can be connected to service oriented networks. In fact, World Wind supports the OGC Web Map Service so that any WMS server can easily be added to its tools as a data source. Visualizing emissions data with these tools has a number of advantages: 1) very little programming is required to gain useful visualization tools, 2) they provide integrated access to some GIS and satellite imagery, and 3) they include some more advanced visualization tools, such as World Wind's temporal animation facility useful for viewing emissions patterns and trends. Figure 4 illustrates one possible information flow for bringing emissions data into these mapping applications. DataFed's wrappers translate heterogeneous data formats into a common multidimensional framework. Once in this multidimensional framework, the data can be passed through a number of data conversion services that make the data suitable for import into other applications, such as exporting into the XML-based KML format used by Google Earth, World Wind's point layer XML format, or as an OGC Web Map Service.

Figure 4. DataFed mediated information flow for emissions data to popular web mapping visualization applications.
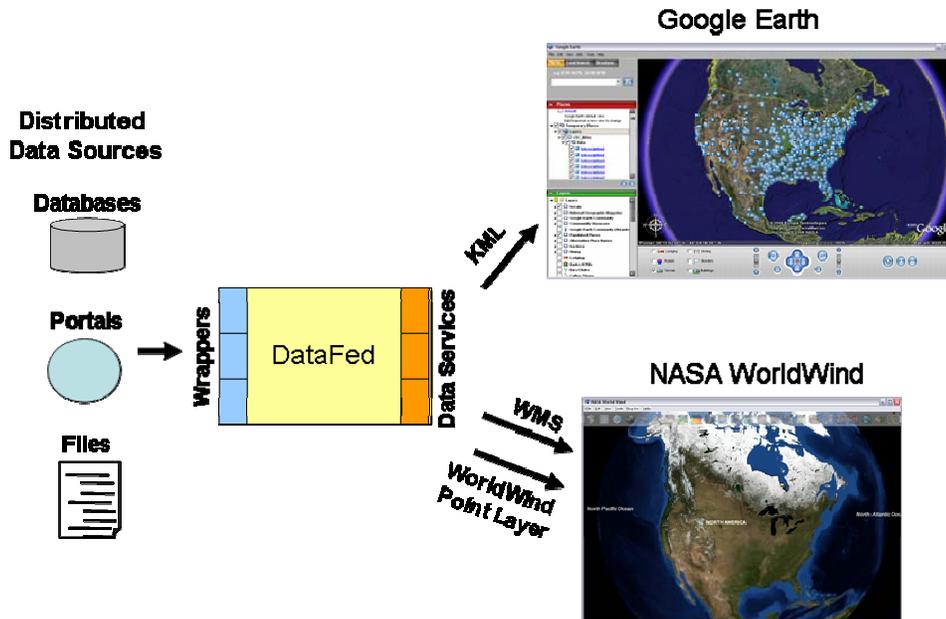
CONCLUSIONS

The proliferation of information systems, standards, and processing/visualization tools available through the web are making it easier to access and use emissions data. An ongoing challenge is how to find and properly connect these components into a networked set of services that provide the most flexibility and greatest capabilities for the emissions community. The NEISGEI web portal assists in bringing the web resources virtually together and provides a forum to explore their connections. DataFed offers an infrastructure for mediating the connection between data sources and applications. Information flow chains of openly accessible data sources, mediation technologies, and web applications are being built with current and evolving technologies. The technologies and applications discussed in this paper represent a subset of the active research and development efforts to address cyberinfrastructure challenges. The greatest contribution to interoperability and the construction of an emissions cyberinfrastructure will continue to be made through community-wide coordination and collaboration among these efforts.

ACKNOWLEDGEMENTS

REFERENCES

[1] The NARSTO Emission Inventory Assessment Team, *Improving Emission Inventories for Effective Air Quality Management Across North America: A NARSTO Assessment*, NARSTO 05-001, 2005.

[2] Atkins, D.E.; Droegemeier, K.K; Feldman, S.I.; Garcia-Molina, H.; Klein, M.L.; Messerschmitt, D.G.; Messina, P.; Ostriker, J.P.; Wright, M.H. *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*. Technical report, National Science Foundation, 2003.

[3] Husar, R.B., Hoijarvi, K., and Falke, S.R., "DataFed: Web Services-Based Mediation of Distributed Data Flow", In Proceedings of Earth Sun-System Technology Conference: Baltimore, MD, 2005.

[4] Buehler, K. and L. McKee, *The OpenGIS Guide: Introduction to Interoperable Geoprocessing and the OpenGIS Specification*, Waltham, MA, 1998.

[5] Miller, P and C. Van Atten, *North American Power Plant Air Emissions*, Commission for Environmental Cooperation of North America, 2004.

[6] Ruminski, M., Simko, J., Kibler, J., McNamara, D. and Kasheta, T. "The Hazard Mapping System (HMS) - A Multiplatform Remote Sensing Approach to Fire and Smoke Detection", In Proceedings of American Geophysical Union Fall Meeting, 2003.

KEY WORDS

Emission Inventories
Web services
Service Oriented Architecture
Portal
Web Mapping
GIS
Cyberinfrastructure