

# **Air Quality Data Integration from Heterogeneous Sources**

**2004 National Conference on Digital Government Research**

**Eduard Hovy**

**Andrew Philpot**

Digital Government Research Center

USC Information Sciences Institute

**Stefan Falke**

Center for Air Pollution Impact and Trend Analysis

Washington University, St. Louis

This poster presents preliminary results from research conducted at the University of Southern California's Institute for Information Sciences (ISI) and Washington University's Center for Air Pollution Impact and Trend Analysis (CAPITA) in addressing data integration challenges for heterogeneous air quality data sources. The research presented in this poster involves partnership with the US EPA, the State of California Air Resources Board, and Santa Barbara County Air Pollution Control District.

## **The Problem: Air Quality Data Integration**

The management of air quality involves local, state, regional, national, and international organizations. At each level, data are collected and used for analysis, assessment, and regulatory enforcement. Effective air quality management requires coordination among multiple organizations and, therefore, requires integration among their respective data sets. This integration remains a complex IT challenge due to the variety of collection, storage, format, and dissemination methods employed by each organization. In most cases today when organizations need to share data require very involved, specialized arrangements between the organizations in order to

## **Initial Results**

ISI has focused on advancing technologies for automating the integration of heterogeneous databases

- Focus on the human-centered process of identifying semantic equivalence between source and target data sets
  - Develop a suite of tools to map, transform, and re-aggregate data from one schema to another, and for visualization
  - Leverage techniques of statistical machine translation and web services
  - Address issues of scalability and maintenance
- 
- Replicated as example source and target the 2001 CEIDARS submissions (databases) from SBAPCD to CARB
  - Developed and evaluated cell-by-cell, column-by-column, and column-within-annotated-row methods of alignment
  - Identified the Expectation Maximization (EM) algorithm as possible approach

Washington University in St. Louis has been conducting research on the aspects of dynamically integrating heterogeneous data sources and has focused on data used in forest fire emissions research and management. 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](http://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).

Figure 1 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.

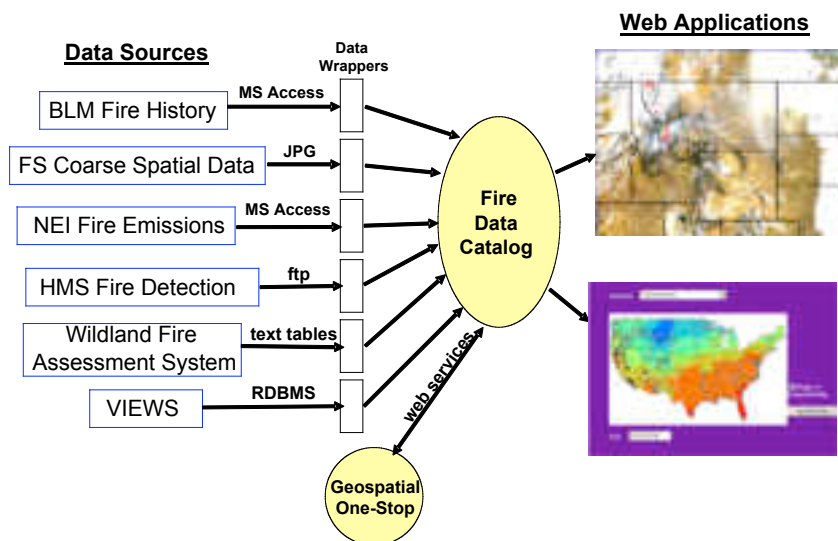


Figure 1. Data Flow and Component Diagram for fire and air quality data

### Next Steps

Close collaboration between ISI and Washington University will provide a testing environment for the developed mapping and data web service technologies and will allow a critical assessment of their value-adding capabilities in the integration of heterogeneous air quality data sources. We will register the California air quality data with DataFed.net and evaluate the capabilities for visualizing and exploring emissions data. This will allow the California data to be brought into context with national air emissions data. The diverse datasets available through the CAPITA catalog (including national air emissions inventories), provides a desirable setting for evaluating ISI’s automated mapping tool on larger datasets. A particular issue to be examined is the scalability of the automated integration approach as we move from small air quality management districts to larger districts, and ultimately to a regional area.