

Automating the Integration of Heterogeneous Databases

Eduard Hovy, Andrew Philpot

Digital Government Research Center
Information Sciences Institute
University of Southern California
Marina del Rey

Stefan Falke

Center for Air Pollution Impact and
Trend Analysis
Washington University, St. Louis
St. Louis

This new and ambitious project addresses a pervasive problem in database management: the alignment of different databases with nearly identical content. We focus on data about air quality, which is collected by local agencies and then sequentially gathered and integrated ‘upward’ by the chain of respective state, regional, national, and sometimes international parent organizations. This (annual) systematic integration by parent organizations of data from their children is an extremely complex exercise, since the data formats at each level are maintained and developed somewhat independently and are constantly being changed. At present, the data mappings are created by hand; the whole data delivery process takes about a year. Our challenge is to (semi-)automate this mapping process.

Researchers 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) are collaborating with the US EPA, the State of California Air Resources Board (CARB), and (at present) the Santa Barbara County Air Pollution Control District (SBAPCD), one of some 35 Air Quality Management Districts in California.

Our research aims at developing a suite of tools to map, transform, and re-aggregate data from one schema to another (ISI), and to develop sophisticated data visualization tools (CAPITA). At ISI, to induce data mappings, we are treating the problem as analogous to learning the correspondences between two human languages, which is a novel and rather unusual approach. Machine translation (MT) research over the past decade has had remarkable success using statistical techniques to induce lexical and structural mappings from one language to another, given parallel corpora as training material. To date, we have replicated as ‘source’ and ‘target’ the 2001 CEIDARS databases from SBAPCD and CARB, and developed several statistical algorithms that search for cell-by-cell, column-by-column, and column-within-annotated-row regularities, using the Expectation Maximization (EM) algorithm to massage the probability distributions. The complexity of the databases and somewhat small size of the data compared to MT data has slowed progress; while our algorithms do identify some mappings successfully, there is still a considerable way to go.

At Washington University in St. Louis, we have been researching the dynamic integration of heterogeneous data used in forest fire emissions research and management. We have developed classes of data wrappers that homogenize data formats and record details of data access requirements. Data types include monitoring network data, gridded model output, emissions inventory databases, and satellite imagery. The data classes are used to register datasets in a data catalog with standard metadata as well as specific data access instructions. The classes of data types and registration forms will allow outside air quality researchers and managers to register their data sets. The registered access instructions are interpreted for browsing and visualization by the CAPITA Voyager Services engine. The Voyager Services includes interfaces for rendering data “views” including maps, time series, and tables. The views are each created with their own web services thereby allowing them to be configured into custom applications using standard web programming languages (such as JavaScript and ASP).

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 into the Voyager Services system and evaluate the capabilities for visualizing and exploring the data's properties. 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.