# Rensselaer team shows how to analyze raw government data

November 15 2010

Who is the White House's most frequent visitor? Which White House staffer has the most visitors? How do smoking quit rates, state by state, relate to unemployment, taxes, and violent crimes? How do politics influence U.S. Supreme Court decisions? How many earthquakes occurred worldwide recently? Where and how strong were they? Which states have the cleanest air and water?

If you know how to look, the answers to all of these questions, and more, can be found in the treasure trove of government documents now available on Data.gov. In the interest of transparency, the Obama Administration has posted 272,000 or more sets of raw data from its departments, agencies, and offices to the World Wide Web. But, connecting the dots to derive meaning from the data is difficult.

"Data.gov mandates that all information is accessible from the same place, but the data is still in a hodgepodge of different formats using differing terms, and therefore challenging at best to analyze and take advantage of," explains James Hendler, the Tetherless World Research Constellation professor of computer and cognitive science at Rensselaer Polytechnic Institute. "We are developing techniques to help people mine, mix, and mash-up this treasure trove of data, letting them find meaningful information and interconnections.

"An unfathomable amount of data resides on the Web," Hendler continues. "We want to help people get as much mileage as possible out of that data and put it to work for all mankind."

# Mining Data.gov

The Rensselaer team has figured out how to find relationships among the literally billions of bits of government data, pulling pieces from different places on the Web, using technology that helps the computer and software understand the data, then combine it in new and imaginative ways as "mash-ups," which mix or mash data from two or more sources and present them in easy-to-use, visual forms.

By combining data from different sources, data mash-ups identify new, sometimes unexpected relationships. The approach makes it possible to put all that information buried on the Web to use and to answer myriad questions, such as the ones asked above. (Answers can be found on the Website http://data-gov.tw.rpi.edu/wiki/Demos).

"We think the ability to create these kinds of mash-ups will be invaluable for students, policy makers, journalists, and many others," says Deborah McGuinness, another constellation professor in Rensselaer's Tetherless World Research Constellation. "We're working on designing simple yet robust Web technologies that allow someone with absolutely no expertise in Web Science or semantic programming to pull together data sets from Data.gov and elsewhere and weave them together in a meaningful way."

While the Rensselaer approach makes government data more accessible and useful to the public, it also means government agencies can share information more readily.

"The inability of government agencies to exchange their data has been responsible for a lot of problems," says Hendler. "For example, the failure to detect and scuttle preparations for 9/11 and the 'underwear bomber' were both attributed in a large part to information-sharing failures."

The Web site (http://data-gov.tw.rpi.edu/wiki) developed by Hendler, McGuinness, and Peter Fox — the third professor in the Tetherless World Research Constellation — and students, provides stunning examples of what this approach can accomplish. It also has video presentations and step-by-step do-it-yourself tutorials for those who want to mine the treasure trove of government data for themselves.

Rensselaer offers the country's first undergraduate degree in Web Science and has one of the first academic research centers dedicated to the field. The White House has officially acknowledged Rensselaer's pioneering efforts in the field. Hendler has been named the "Internet Web Expert" by the White House, and the Web Science team at Rensselaer includes some of the world's top Web researchers.

"Rensselaer has pre-eminent expertise in what the Web is and what the Web future will be," says Hendler.

## Data.gov offers opportunity

Hendler started Rensselaer's Data-Gov project in June 2009, one month after the government launched Data.Gov, when he saw the new program as an opportunity to demonstrate the value of Semantic Web languages and tools. Hendler and McGuinness are both leaders in Semantic Web technologies, sometimes called Web 3.0, and were two of the first researchers working in that field.

Using Semantic Web representations, multiple data sets can be linked even when the underlying structure, or format, is different. Once data is converted from its format to use these representations, it becomes accessible to any number of standard web technologies.

One of the Rensselaer demonstrations deals with data from CASTNET, the Environmental Protection Agency's Clean Air Status and Trends

Network. CASTNET measures ground-level ozone and other pollutants at stations all over the country, but CASTNET doesn't give the location of the monitoring sites, only the readings from the sites.

The Rensselaer team located a different data set that described the location of every site. By linking the two along with historic data from the sites, using RDF, a semantic Web language, the team generated a map that combines data from all the sets and makes them easily visible.

his data presentation, or mash-up, that pairs raw data on ozone and visibility readings from the EPA site with separate geographic data on where the readings were taken had never been done before. This demo and several others developed by the Rensselaer team are now available from the official US data.gov site: http://data.gov/semantic.

## Many examples on the Web

Other mash-up demos on the http://data-gov.tw.rpi.edu/wiki/Demos site include:

- The White House visitors list with biographical information taken from Wikipedia and Google (now also available in a mobile version through iTunes);
- U.S. and British information on aid to foreign nations;
- National wild fire statistics by year with budget information from the departments of Agriculture and Interior and facts on historic fires;
- A state-by-state comparison of smoking prevalence compared with smoking ban policies, cigarette tax rates, and price;

- The number of book volumes available per person per state from all public libraries;

- An integration of basic biographical information about Supreme Court Justices with their voting records from 1953 to 2008, with a motion chart that looks at justices' decisions over the years on issues such as crime and privacy rights.

The aim is not to create an endless procession of mash-ups, but to provide the tools and techniques that allow users to make their own mash-ups from different sources of data, the Rensselaer researchers say. To help make this happen, Rensselaer researchers have taught a short course showing government data providers how to learn to do it themselves, allowing them to do their own data visualizations to release to the public.

## Many potential users

The same Rensselaer techniques can be applied to data from other sources. For example, public safety data can show a user which local areas are safe, where crimes are most likely to occur, accident prone intersections, proximity to hospitals, and other information that may help a decision on where to shop, where to live, even areas to avoid at night. In an effort McGuinness is leading at Rensselaer along with collaborators at NIH, the team is exploring how to make medical information accessible to both the general public and policy makers to help explore policies and their potential impact on health. For example, one may want to explore taxation or smoking policies and smoking prevalence and related health costs.

The Semantic Web describes techniques that allow computers to understand the meaning, or "semantics," of information so that it can

find and combine information, and present it in usable form.

“Computers don’t understand; they just store and retrieve,” explains Hendler. “Our approach makes it possible to do a targeted search and make sense of the data, not just using keywords. This next version of the Web is smarter. We want to be sure electronic information is increasingly useful and available.”

“Also, we want to make the information transparent and accountable,” adds McGuinness. “Users should have access to the meta data – the data describing where the data came from and how and when it was derived — as well as the base information so that end users can make better informed decisions about when to rely on the information.”

The Rensselaer team has also been working to extend the technique beyond U.S. government data. They have recently developed new demos showing how this work can be used to integrate information from the U.S. and the U.K. on crime and foreign aid, to compare U.S. and Chinese financial information, to mashup government information with World Bank data, and to apply the techniques to health information, new media, and other Web resources.

**More information:** Some Mash-ups:

Clean Air Status and Trends Network (CastNet)
DEMO: data-gov.tw.rpi.edu/demo/exhib … t/demo-8-castnet.php
DESCRIPTION: data-gov.tw.rpi.edu/wiki/Demo: …
s_and_Trends_-_Ozone

US Global Foreign Aid from 1947-2008
DEMO: data-gov.tw.rpi.edu/demo/stabl … d2008/demo-1554.html
DESCRIPTION: data-gov.tw.rpi.edu/wiki/Demo: …
reign_Aid,_1947-2008

White House Visitor Search
DEMO: data-gov.tw.rpi.edu/demo/stabl … /top100-visitees.php
DESCRIPTION: data-gov.tw.rpi.edu/wiki/Demo: … House_Visitor_Search

Trends in Smoking Prevalence, Tobacco Policy Coverage and Tobacco Prices
Demo: logd.tw.rpi.edu/demo/trends_in … e_and_tobacco_prices
Description: logd.tw.rpi.edu/project/popscigrid

Provided by Rensselaer Polytechnic Institute