Uplink

September• October 2003
Volume 15, Number 5

Published bimonthly by the National Institute for Computer-Assisted Reporting

www.nicar.org

CHILDREN

Tracking quality of life

By Dave Davis
The (Cleveland) *Plain Dealer*

When Editor Doug Clifton called us into the conference room a year ago and said he wanted us to undertake an in-depth examination of the state of children in Northeast Ohio, we politely rolled our eyes.

It was clearly one of those dreaded editor-assigned projects, an important subject but one that had been more than a bit worn out in the press. It reminded me of the time a top editor here turned to me at a crowded projects brainstorming session and loudly blurted out, "Guns." She didn't add anything to the idea, but in her mind she was already laying out a five-part series that I was determined not to work on.

Clifton came to the table with more than simply the idea of "children" but not too much more. He wanted the stories to be data-driven, for us to take a report-card approach, compiling indicators that would

continued on page 22

SPOTLIGHT: POLLUTION

Taking the pulse of a body of water

By Lisa Stiffler, *Seattle Post-Intelligencer*

Located on the shores of the Puget Sound, our newsroom has a front-row seat to our region's primary water body. We watch the ferries, kids playing on the beaches, fishermen tromping out to the piers. But when colleague Robert McClure and I were charged with evaluating the health of the sound, we had to figure out how to check its vital signs.

We looked at different indicators: pollution – its sources past, present and future; populations of fish and other marine life; and contamination in seafood. Lise Olsen, then the computer-

continued on page 20

SPOTLIGHT:

FOR MORE ON POLLUTION SEE:

- Mapping contamination hot spots, p. 6
- Uncovering lead poisoning in Detroit, p. 8
- Finding pollution sources in Louisville, p. 10
- Following trends with air pollution data, p. 12
- The September/October 2003 *IRE Journal*

GOVERNMENT CONTRACTS

Database reporting shows no-bid deals

By Phil Williams, *WTVF-Nashville*

John Stamps was one heck of a talented guy with all sorts of unique experiences that were needed by Tennessee taxpayers – that is if you believed the bureaucrats who worked for Stamps' friend, former Tennessee Gov. Don Sundquist.

The longer I sat at my computer, examining state contracts data, the more incredible the insurance agent's talents – especially for tapping into his political connections to land state contracts – appeared. Ultimately, we discovered Stamps' involvement in more than $200 million in state contracts handed out by the Sundquist administration.

"In common parlance, it stinks!" said veteran state Rep. Frank Buck, as he reviewed the findings of our computer-assisted reporting project.

Our "Friends in High Places" investigation led to a federal-state criminal

continued on page 4

Bits & Bytes

Fresh migration, aircraft data

"San Diego County may have a stunning coastline, incomparable weather and a world-famous zoo, but all that has not been enough to stem the flow of residents leaving the region for fast-growing areas in the West where housing is far cheaper."

So began a recent story by the *San Diego Union-Tribune* that used Internal Revenue Service migration data, which is offered by the IRE and NICAR Database Library. The U.S. Census Bureau compiles the data, based on personal income tax return information, for the IRS.

The database allows journalists to track the movement of people and their incomes, in and out of counties, from 1992 through 2002. *USA Today* used the migration data in an October 2002 report about how migration of affluent Californians has led to the dominance of Republican politicians in the mountain west states.

Other recent updates include the U.S. Federal Aviation Administration Aircraft Registry, which lists all aircraft and aircraft owners registered in the United States (current through February 2003). Find out more at *www.ire.org/datalibrary/databases* or call 573-884-7711.

Breaking news help

Power blackouts. Nightclub fires. Wars. Plane crashes. When disasters and other big news occur, journalists can find

continued on page 4

INSIDE NICAR

Contribute your files to database library

By David Herzog, *NICAR and Missouri School of Journalism*

Journalists hungry for high-quality data have turned to the IRE and NICAR Database Library for more than a decade.

For those of you who are unfamiliar with it, the Database Library acquires data from government agencies and provides it in useful formats for journalists and journalism organizations. The library staff, directed by Jeff Porter, processes the data and often fixes problems in the data so customers can focus on analyzing it for news stories. The staff also creates documentation that helps journalists understand the data better.

The number of Database Library offerings has multiplied many times since the early days. Journalists can obtain data on a wide range of topics – transportation, public safety and business, to name a few – for low fees on a sliding scale. Smaller news organizations pay less than the bigger outfits.

When the Database Library started more than a decade ago, its initial offerings were databases that journalists had used for news stories and then donated. It was like a journalists' co-op for data.

We want to revive that practice and ask you to donate your databases – the raw files you obtained from government agencies – to IRE and NICAR after your stories run. We are mainly interested in national databases. Send us your data, along with any documentation and notes about quirks that you discovered. We'll take a look at the data to see if it's something we would like to add to our collection.

Jeff is leading this push to add more data. You can contact him at jeff@ire.org if you have data that you'd like to share.

Contact David Herzog by e-mail at dherzog@nicar.org.

Hands-on CAR training upcoming

IRE and NICAR have a number of training opportunities coming up for journalists who want to learn more about computer-assisted reporting.

Journalists can get intensive hands-on training in using spreadsheets and database managers to produce high-impact news stories at the CAR Boot Camps, which will be held four times next year in Columbia, Mo. The sessions are Jan. 4-9, March 21-26, May 16-21 and Aug. 1-6.

An Advanced Boot Camp on Mapping, scheduled for Jan. 9-11, will offer intensive hands-on training using mapping software. An Advanced Statistics Workshop, taught by Steve Doig of Arizona State University and Sarah Cohen of *The Washington Post*, will be Feb. 13-15 for reporters who want to move beyond basic CAR and use statistical analysis.

For more information, see the IRE training calendar at *www.ire.org/training/otr.html*.

FIRST VENTURE

State cars sit, waste tax money

By Morgan Loew, *KPHO-Phoenix*

Heading into the 2003 legislative session, the state of Arizona was facing a $1 billion budget shortfall. As in nearly every other state in the country, revenues had fallen, programs were getting slashed and lawmakers were desperately looking for ways to cut expenses.

Journalists were also under the gun to find wasteful spending and report it in print or on the air as fast as possible. We found it in an unusual place: state vehicles that were used so little, the state's own guidelines said they weren't worth the cost of owning them.

Compared to hard-core computer-assisted reporting projects that can take months of spreadsheet and database analysis, this story was pretty simple to do using the skills I learned in a NICAR Boot Camp in January.

Looking for efficiency

We found a recent audit of one state agency that referred to an "efficient use guideline" for state vehicles. It concluded state vehicles needed to be driven at least 10,000 miles per year to justify the cost of acquiring and maintaining the vehicles.

With that minimum mileage standard in mind, we sent a public records request to the Arizona Department of Administration asking for the entire state vehicle database. The department buys most state vehicles and leases them to individual agencies.

We specifically asked for the following: license plate numbers; agency the vehicles are leased to; year, make and model of the vehicle; mileage at the beginning of the last fiscal year; and mileage at the end of the last fiscal year.

As is the case with most requests, the data I had asked for had shortcomings. Some vehicles were bought in the middle of the year and some were retired in mid-year. The result would not have given me a clear picture of how many miles were driven over a full year. Fortunately, the public information officer I was dealing with helped me and we came up with a better request.

The department provided an Excel file at not cost...

Querying the data

The department provided an Excel file at no cost that had more than 2,000 rows and included a column that showed the average miles driven per month. It appeared the state had done most of the work for me. All I had to do was import the data into Microsoft Access and write a query that isolated vehicles driven fewer than 10,000 miles per year (834 miles per month.) It looked something like this in Structured Query Language:

```
SELECT *
FROM [veh list]
WHERE [FY03 YTD Average miles
driven per month] <= 834
```

The result was that nearly half the vehicles were underused, according to the state's own guidelines.

We listed the agencies with the most underused vehicles by using the following query:

```
SELECT Agency, count(*)
FROM [veh list]
WHERE [FY03 YTD Average Miles
Driven per Month]<= 834
GROUP BY Agency
ORDER BY count(*) DESC
```

So we had a story, but we needed a good way to tell it. I spent a couple of days driving around the state Capitol, studying parking lots that held state vehicles. My photographer and I performed surveillance next to one of the lots, and watched as several state vehicles just sat there day after day. We now had one good visual element to illustrate the issue.

Adding perspective

Next, I contacted the state Department of Education and got a list of relatively inexpensive state programs that were facing the budget axe. We ended up profiling a program for underprivileged preschoolers and another that helped adults get their high school diplomas.

The story boiled down to this: "State leaders say there is no money lying around to fund programs like preschool and adult education, but we found lots of extra tax dollars – literally 'parked' all over the state Capitol."

The day the story aired, the governor held a news conference to announce an "efficiency review." Part of that review dealt with state vehicles. The governor pledged to get rid of the vehicles that did not meet the minimum mileage standards. She said one of the beneficiaries of the savings would be education spending.

In the end, the database work made up only a small part of the story, but it provided the real meat. Without my limited background in CAR, I would have never thought to look for wasted tax dollars in a parking lot.

Contact Morgan Loew by e-mail at Morgan.Loew@meredith.com.

Bits & Bytes

continued from page 2

help in a hurry at IRE's "In the News," online at *www.ire.org/inthenews_archive*.

When power blackouts hit New York and other major cities in the U.S. and Canada, a special "In the News" section listed useful databases, publications, Internet sites and other resources. For instance, the site linked to several database and spreadsheet files from the Energy Department's Energy Information Administration. Also, the IRE and NICAR Database Library has data on federal contracts that includes the company, contact information, agency, type of work performed and place where the work was performed. Certainly there were companies with federal contracts involved in the blackout.

New Beat Book

From lobbying to loopholes, soft money to hard money, reporters will find useful information and inspiration in "Unstacking the Deck: A Reporter's Guide To Campaign Finance." Written by veteran journalists Michael A. Weber, Aron Pilhofer and Derek Willis, this is the latest publication in IRE's Beat Book series. It includes chapters on gathering documents and data and using campaign finance databases, such as those from the Federal Election Commission. There's also information on finding local and state data, and putting together your own database. The book costs $15 for IRE members and $25 for nonmembers. For details, see *www.ire.org/store/books/campaign.html* or call 573-882-3364.

Contracts

continued from page 1

probe, involving the FBI and at least three other agencies. Investigators raided at least two companies with ties to Stamps, and a federal grand jury subpoenaed hundreds of thousands of e-mail messages based on information unearthed by our reporting.

No charges have resulted so far, but the grand jury is continuing its probe.

Unlike a lot of investigations, this one didn't begin with a tip. Instead, it began from mere curiosity during a brainstorming session.

With Tennessee having just enacted the largest tax increase in state history, my photojournalist partner Bryan Staples and I wondered who was getting taxpayers' money. It was a question that couldn't be answered easily without a computer.

Getting records

First, using Tennessee's Public Records Law, I requested a database of state contracts. In fact, this request was the easiest part of the process. The state's Department of Finance and Administration quickly e-mailed the data in a .dbf file, which was then imported to Access. It contained hundreds of records, which required a targeted approach.

Among the fields of information was a description of the bidding process, including one that I immediately found of great interest: "noncompetitive" contracts. Those are contracts that are awarded without any sort of bidding process because the vendors are deemed to offer a unique service unavailable anywhere else.

Slicing out the no-bid contracts from the data, I then began to background the companies involved in some of the larger contracts. To accomplish that, I began with a basic Web search, then moved on to AutotrackXP's online public record service.

One contract, with a company called Workforce Strategists, set off immediate alarms.

The company had been awarded a $1.9 million no-bid contract to provide intensive counseling and coaching to help the unemployed get back to work. A "sole source memo" to justify the contract, also available in electronic form from budget officials, claimed Workforce Strategists was "the only company in Tennessee that has experience" for the job.

A quick check of the Tennessee Secretary of State's corporation information Web site revealed that the company had been incorporated just six days before that memo was written.

"Obviously, with six days, their experience is rather limited," Rep. Buck said.

In fact, the secretary of state's Web site also gave us the company's mailing address: Stamps' insurance office. At the time Stamps was completely unknown to us. But a quick check of newspaper articles through NewsLibrary newspaper archive revealed that he was a former business partner of the governor's.

And there was more.

A search of the state Web site revealed Stamps' role as a registered lobbyist for Education Networks of America (ENA), which had received more than $180 million in state contracts to provide Internet service to Tennessee schools.

I had remembered reading a note on ENA's Web site that its work for the state began as a sole source, or no-bid, contract. But the note was no longer posted there, and state officials denied it. However, a search of a Web archive produced the document I had remembered and gave us evidence to take back to state officials.

But we weren't finished asking questions with the computer.

Next, I returned to AutoTrackXP and began running every query that I could imagine. I ran a personal and business history of John Stamps, a history of all known businesses listed at the same address as Workforce Strategists, as well as a history of all other companies that shared the same registered agent as Workforce Strategists.

The result was the discovery of three sister companies:

One, Privatization Strategies, was part owner of Workforce Strategists. And it turned out that Privatization Strategies had been set up with the help of an attorney who was formerly a special assistant to Sundquist.

A second, Comprehensive Community Care (CCC), included then-Deputy Gov. Alex Fischer as one of its officers. A quick Google search led to the discovery that CCC was financed from the state's controversial TennCare health insurance program.

In addition, Stamps had also set up another corporation, ComTraining.Net, a company run by Fischer's wife and the wife of the state's economic development commissioner. We discovered that ComTraining.Net had provided training for Comprehensive Community Care, again using taxpayer money through TennCare.

Even though the Sundquist administration insisted there was nothing improper about such arrangements, we found information about TennCare contracts on the state's Web site and discovered that they prohibit contractors from paying money "directly or indirectly" to state officials.

On top of that, Stamps was also a lobbyist for two other state contractors who received tens of millions of dollars of taxpayer money.

Interestingly, our AutoTrackXP search also revealed that Stamps filed for bankruptcy just days after Sundquist had been sworn into office in January 1995. He had claimed $1.1 million in debt, with just $45,000 in annual income. A state database showed that, after he filed for bankruptcy, Stamps registered to become a Capitol Hill lobbyist.

And there was more. Despite Stamps' bankruptcy, state campaign finance databases revealed that he was handing out campaign contributions to help those who were helping him, including $1,000 to Sundquist's reelection campaign.

E-mail messages

But perhaps the most revealing information came from ENA e-mail messages obtained by the station.

Those messages revealed that ENA officials, including Stamps, had helped to write a sweeping reading initiative proposed by Sundquist.

In an e-mail labeled "extremely confidential and not to be discussed outside of ENA," company officials said that they had just negotiated a contract that would provide an extra $40 million a year.

That e-mail, which focused on the governor's reading initiative, was written before Sundquist had even presented his plan to state lawmakers.

"I think this would have sent up red flags – I know it would have," said state Rep. Gene Davidson, one of the House's leading education experts as he read the message for the first time. "My reaction would be that someone thought it was a done deal."

Lawmakers approved the plan, although it was never funded because of the state's budget crisis.

Other e-mail suggested the role of politics in those state contracts. In one message, ENA President Al Ganier had instructed his employees: "Our sales efforts are built on the concept that all of us are involved in relationship building for improvements in existing contracts."

About two months later, a team of federal and state agents searched the company, seizing records and copying computer hard drives.

Our investigation prompted legislators to pass a new law that requires more scrutiny of no-bid contracts. Tennessee's new governor, Phil Bredesen, invited us to privately witness his signing of the legislation. And he has imposed his own tighter controls.

"It's primarily an issue," the new governor said, "of having the public have some confidence that what we are doing is in their interests."

Contact Phil Williams by e-mail at pwilliams@newschannel5.com.

readme.txt

Phil Williams and Bryan Staples of WTVF-Nashville won a Certificate in the 2002 IRE Awards for "Friends in High Places."

To order a transcript of the stories, a tape or both, contact the IRE Resource Center at 573-882-3364 or rescntr@ire.org and ask for Story No.19649. IRE members can view a videostreamed excerpt of the report in the Broadcast Center area of the IRE Web site at *www.ire.org/broadcast.*

Read the "Friends in High Places" stories on the Web at *www.newschannel5.com/content/investigates/1133.asp.*

MAPPING IT OUT

The latest uses of mapping in news reporting.

Pinpointing polluters in Chicago area

By Jeff Porter, *IRE and NICAR*

Environmental reporting is a natural – pun intended – for journalists looking to use their analytic mapping skills and find information for news stories.

Environmental data that can be mapped from state and federal agencies is often a simple download away on the Internet. For example, go to the U.S. Environmental Protection Agency Web site at *www.epa.gov* and search for the phrase "underground storage tank." You'll get a list of state agencies that manage data about leaking tanks and, often, links to that data.

The environmental agencies often include the locations, reported in latitude and longitude, of regulated entities. For example, the Toxics Release Inventory data (*www.epa.gov/tri*) from the EPA includes the latitude and longitude. A journalist using a geographic information system can create a pin map based on those coordinates and, going further, use density mapping tools to show high concentrations.

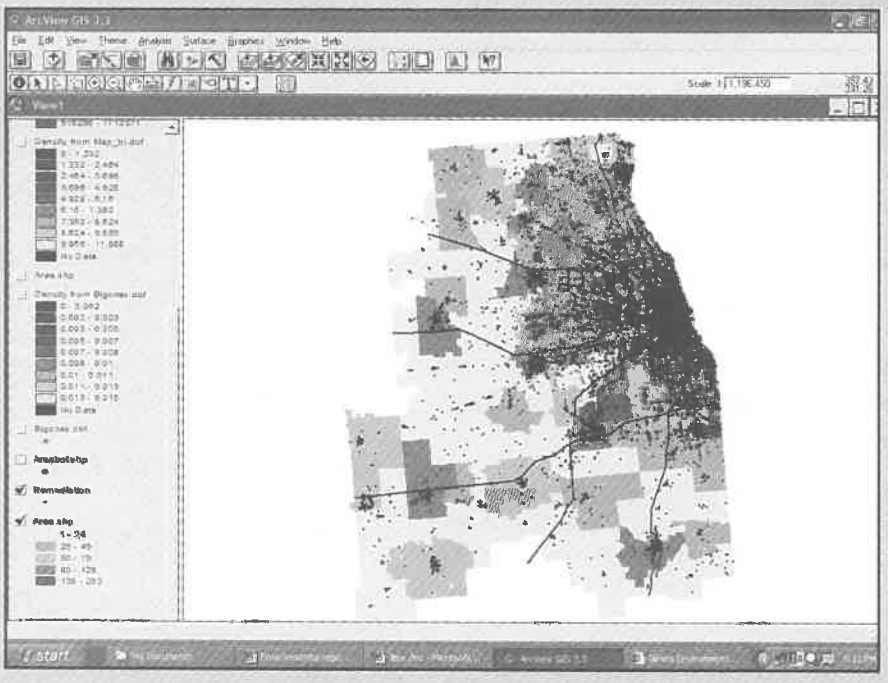
The Toxics Release Inventory and

other environmental datasets include even more geographic information that can be mapped: counties, cities and ZIP codes. While ZIP code areas may or may not have precise boundaries, most people know their own Zip codes.

The IRE and NICAR Database Library recently used those geographic elements when it assisted WMAQ-Chicago in pinpointing contaminated sites targeted for cleanup by the Illinois Environmental Protection Ageny. The data included hazardous waste sites and leaking underground storage tanks. The Database Library used ArcView 3.3 and the Spatial Analyst extension to conduct the analysis.

The data, obtained by station producer Michele Youngerman, came from the Illinois EPA in dBASE files, easily opened in ArcView after we cleaned up the data using Visual Fox Pro to convert nine-digit ZIP codes to five-digit codes and convert the coordinates from text to number.

(A note of caution about latitudes



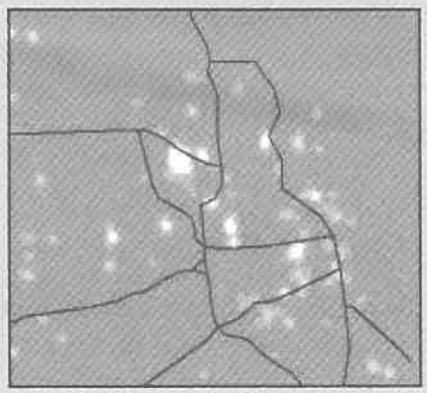
and longitudes: Make sure you have them reported properly. ArcView 3.x requires decimal degrees. Occasionally the coordinates are reported in degrees, minutes and seconds and will have to be converted.)

After we cleaned up the data, our analysis took several steps.

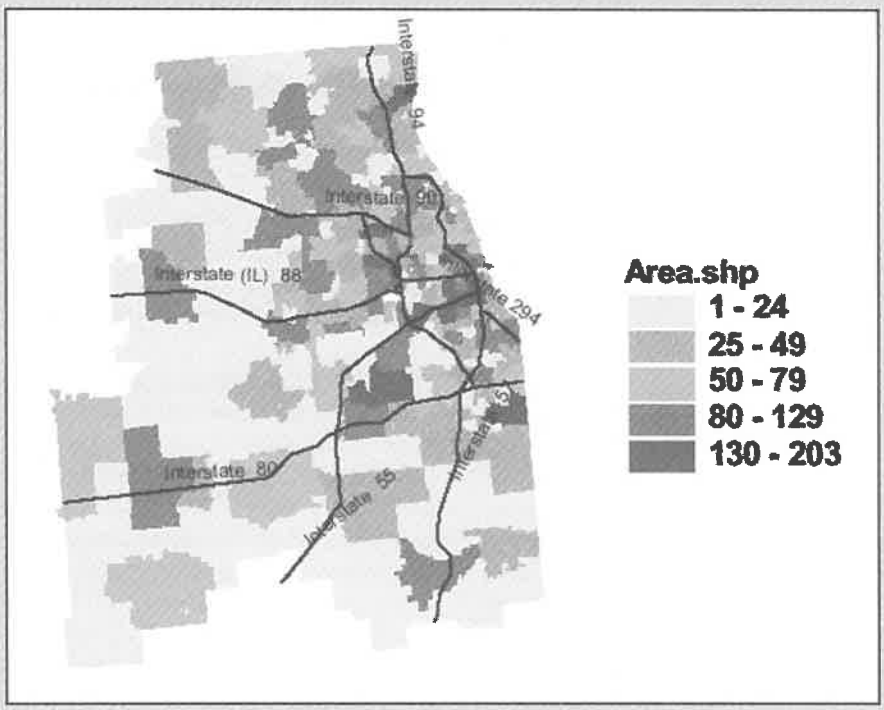
1. We created a point map using the numeric latitude and longitude coordinates via the View\Add Event Theme menu command.

2. The first map was too laden with dots; we wanted to restrict our results. So we used the definition query tool to include only larger polluters.

3. With some advice from GIS expert Matthew Waite of the *St. Petersburg Times,* we used the Spatial Analyst extension to identify "hot spots" of regulated sites. The resulting density maps weren't pretty, but they helped us focus on specific areas and confirmed some suspicions for the station.



4. We limited the maps to the specific ZIP codes provided by WMAQ that define its viewing area by using definition queries. Then we used ArcView's Geoprocessing wizard extension to clip base map layers – roads, cities and rivers – for just the viewing area.



5. Using the clipped data, we summarized the number of regulated sites by ZIP code and mapped the results.

The resulting maps highlighted areas where large polluters — generating, hauling or managing hazardous materials – are also close to residential areas. The television station reported about thousands of contaminated sites near homes and playgrounds, although homeowners often don't know about the contamination. The analysis also helped give perspective for the story of three families facing serious illness. All three families believe the illnesses were caused by companies releasing toxic waste.

The stories can be found at: *www.nbc5.com/nbc5/2203768/detail.html*

Contact Jeff Porter by e-mail at jeff@ire.org.

Would you be willing to share a mapping example with fellow journalists? Send an electronic copy of the map along with details to David Herzog at dherzog@nicar.org

SPOTLIGHT: POLLUTION

Mapping soil tests shows legacy of lead

By Tina Lam and Wendy Wendland-Bowyer, *Detroit Free Press*

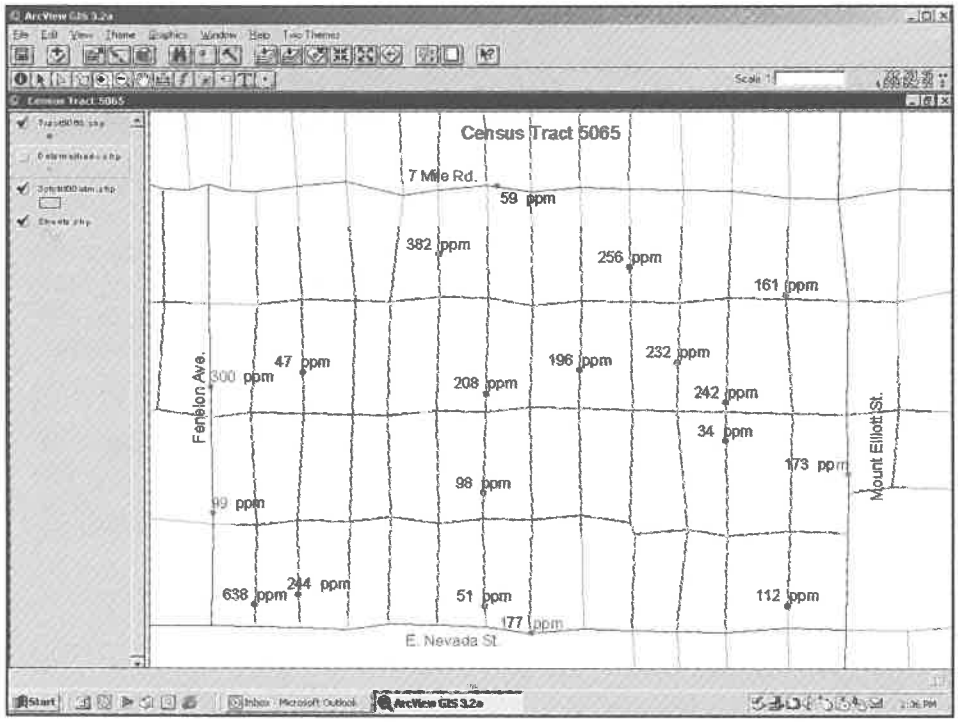
Lead poisoning is still a big problem in Detroit and other Michigan cities. Because many of our readers think it is old news, we wanted to bring a fresh perspective.

"We didn't want to write an anecdotal story," said Alison Young, then project editor and now an investigative reporter for Knight Ridder's Washington bureau. "We wanted to break news and give readers new information that would get their attention." Our early research showed little had been done to alert the public to the dangers of lead-contaminated soil and the legacy of long-forgotten lead smelters.

Eight reporters, a data analyst and three photographers spent seven months working on a series published in January. We interviewed 300 people, examined more than 10,000 pages of documents, filed a dozen public records requests and performed computer-assisted reporting. Since January, we've written follow-ups and three other major stories, including an analysis of a state database of children's lead test results.

We hired a soil expert who has studied lead left in soil from leaded gasoline – Howard Mielke, professor of environmental toxicology at Xavier University. His team took more than 400 soil samples, spread randomly along lines drawn through the Detroit metro area, from downtown radiating out into the suburbs. They used global positional satellite gear to determine where to dig each sample and then tested them in their lab in New Orleans for lead and other heavy metals.

When we got the soil test data, it arrived with coordinates of where each sample was taken. Victoria Turk, our data analyst, loaded the data into the ArcView 3.2 geographic information system (GIS) program to plot the locations of samples using the latitude and longitude coordinates. She then placed the sample point file over a streets layer and on printed maps. Turk also added a map layer containing the locations of schools hosting children from kindergarten through fifth grade to see how close they were to the contamination locations. Turk obtained the school data from the National Center for Education Statistics Web site (*http://nces.ed.gov*) and geocoded it in ArcView to create a point map.



Turk had to convert the schools data to match the projection of the soil sample layer.

Using the sample maps, reporters and photographers fanned out, interviewing people who lived near the spots we'd sampled. The research showed that lead was highest in the soil in the inner city, with levels dropping near and in the suburbs. Our story talked about similar findings in other cities and the fact that some scientists think contaminated soil is as great a contributor to lead poisoning among children as lead paint and dust.

One of the five days of the series was devoted to a defunct lead smelter in Detroit and the Environmental Protection Agency's failure to notify people nearby that lead might have blown into their homes and yards. We asked the Xavier University team to return to Detroit and sample areas around the smelter. They found elevated levels. The EPA had taken just seven samples before determining there was no neighborhood contamination. We did 100 and found that there was contamination, some of the levels astronomically high. Turk again used ArcView to map the results. She also used EPA's TRI Explorer on the Web (*www.epa.gov/triexplorer*) to find out how much lead and lead compounds had been emitted in the three counties near Detroit.

Our series included graphics showing what we found. It also had plenty of human tales, portraits of families all over the state who had been affected by lead poisoning. One story focused on how the city and state fumbled efforts to abate houses for lead paint and left

millions of dollars unspent or used on homes where no children lived.

During the project, we filed a FOIA request asking the Michigan Department of Community Health for its database of the results of all blood-lead tests for young children during the previous five years. The state requires that all clinics or physicians who do blood-lead tests report them to the state. After denying the database existed, the state eventually agreed to give it to us, but only on paper and only for $2,445. It was already too late for the series. But we sued the Department of Community Health in January in Oakland County Circuit Court and won our case – and $10,000 in attorney fees – in June.

The state provided nearly half a million records with names redacted. The main fields were a specimen ID (unique for each test, not each child), test date, result and the child's birth date, address to the block level, race and sex. We wanted to find the most-poisoned blocks in the state and see whether health officials were targeting them.

Megan Christensen, a former NICAR data analyst interning at the *Free Press*, joined us on the story. We used VEDIT to remove formatting glitches in the original file and then ran the addresses through ParseRat to separate a single field into components, such as block, street, direction and P.O. Box. We wanted a clean block and street number for each record. (For more about ParseRat, see p. 16 in this issue of *Uplink*).

The state had rounded the house numbers to the 100s. For example, 1500 Elm represented all houses from 1450 Elm to 1549 Elm, so that's how we defined our blocks. But we also had to separate Southwest Elm from Northeast Elm. We had to visit some of the streets and cross-reference others in street directories to make sure we had it right.

Once we separated the records by block, we had to throw out the duplicates. Some records were redundant because the same child was tested in multiple years. Because we had no names, we settled for a grouping of birth date, race, sex and address to identify each child. Unfortunately, we did not come up with a reliable programming solution to throw out duplicates. Instead, we used Microsoft Access to query the data and group results by block. Then we printed out block results and picked out the duplicates by hand based on matches of birth date, race and sex. Then we ran new block sums to determine how many separate children were poisoned on each block.

This forced us to focus on only the top 10-15 blocks because it would have been too time consuming for every block statewide. Had we been able to eliminate all the duplicates, the story would have produced more concrete numbers. But we were still able to focus on the blocks with the most poisonings and report on those families.

The analyst for the Michigan Department of Community Health admitted they'd never done this. He thought the data was too dirty to eliminate double counting, so had never attempted it. Our story showed that most health officials didn't know which were the worst blocks and were not targeting those areas. The state had even denied funding to one county with some of the worst blocks.

Our lead stories have triggered results. Embarrassed health officials have already started knocking on doors in the poisoned blocks that we found. Days after the January series ended, the EPA announced it would finish the cleanup of the abandoned lead smelter and sample the yards near it. The agency set up a satellite office in the neighborhood, got permission to test widely and discovered that 75 percent of the homes they sampled within one-quarter mile of the smelter had lead-contaminated yards. They're working on cleaning them up.

After another story we did on 16 other old smelters in Detroit that hadn't been investigated, the state sent environmental officials to examine every site. They've discovered that all but three were indeed lead smelters and could still be causing contamination. Soil sampling is in the works.

Our lead stories also led directly to legislation requiring inspectors to check for lead in day care centers. In August the governor introduced a 33-page plan to mount a new attack on lead poisoning, citing our work in the second paragraph. It should generate new legislation on several fronts, including requiring more lead testing and penalizing landlords who don't repair lead-contaminated apartments. We'll be watching.

Contact Tina Lam by e-mail at lam@freepress.com.

Contact Wendy Wendland-Bowyer by e-mail at wendland@freepress.com.

readme.txt

Here are some databases journalists can use when reporting on lead poisoning:

Child blood-lead test data – often contains a record for each lead poisoning test. Children who are lead poisoned often will receive follow-up tests so they can have more than one record. Data should include the test result, type of test (finger stick or blood draw, for example), date of test, gender, race and age of the child.

Environmental lead inspection data – contains detailed information about properties inspected for the presence of lead. Since this data identifies properties rather than poisoned kids, agencies may be more willing to provide a detailed street address. You can also get the name of the owner, test results, building type and inspection results.

Abatement loan data – contains information about loans granted by public agencies and housing authorities to homeowners for correcting lead paint hazards. The information includes loan recipient, home address, abatement address, loan amount and loan due date.

SPOTLIGHT: POLLUTION

Showing sources, effects of emmissions

By James Bruggers and Mark Schaver, *The (Louisville, Ky.) Courier-Journal*

Data analysis played a major role in two recent projects that showed the severity of dirty air in Louisville and identified the area's biggest polluters.

Stories published in May showed how the city's dirty air could cause a significant number of additional cancers and other health problems, and that the concentrations of certain chemicals were hundreds of times higher than levels considered safe by pollution regulators.

In June the newspaper followed up with a six-page special section that contained graphics pinpointing the sources of pollution and looked at emissions trends.

We used Microsoft Access, Excel, ArcView 8.3 geographic information system and Perl to wrestle with large quantities of data from local and federal air pollution regulators.

For the first project, environmental reporter James Bruggers, under the Kentucky open records law, obtained air-sampling results in an Excel spreadsheet from the Louisville Metro Air Pollution Control District. The district enforces the federal Clean Air Act in Louisville.

The spreadsheet contained, among other things, the average, median and 95-percent upper confidence limit values for 13 monitoring stations spread over the Louisville area, as well as the number of times an individual chemical had been detected over a 12-month period in 2000 and 2001. Roughly 100 of about 160 chemicals that were sampled had been detected at least once during that period.

To see which of those chemicals might be of concern, we wanted to compare the monitoring results with health-risk benchmarks developed by two regional offices (San Francisco and Philadelphia) of the U.S. Environmental Protection Agency. We downloaded them from the Web. The EPA lacks national benchmarks. These thresholds are sometimes called risk-based concentrations or preliminary remediation goals and were developed for screening potential toxic dumps.

We decided that we would focus on chemicals whose concentrations exceeded both benchmarks. We used Access to join the air sampling results table with the threshold tables. We joined the tables using the unique Chemical Abstracts Service number that identified each chemical.

Then we calculated a ratio showing which chemicals exceeded the thresholds. We exported the results into Excel and used conditional formatting to highlight in red the sites that exceeded the thresholds.

A week before we published the stories, the Philadelphia regional office posted a revised set of screening levels, which required us to perform our analysis again. Fortunately, there were only a few changes.

Our results were striking: Concentrations of at least 18 toxic chemicals and compounds measured in the air exceeded EPA benchmarks. Some were at levels hundreds of times higher than what the government considers acceptable. Concentrations of one chemical, the human carcinogen 1,3-butadiene, were 2,400 times higher than the EPA health threshold at one sampling site. With ArcView, we produced a map showing where the air samples were taken and used proportional symbols to show the relative quantity of butadiene emissions measured at each.

The number crunching was fairly straightforward; the real difficulty was making sure, through interviews with a variety of experts, that the newspaper's analysis was legitimate. So we talked to university-based toxicologists, environmental statisticians, former environmental regulators and physicians. We also relied heavily on federal government Web sites for information about the chemicals and compounds that we found in high concentrations.

After the stories ran, the newspaper obtained a just-completed environmental consulting firm's draft analysis of the air sampling results that confirmed our findings of a significant air quality problem.

Finding polluters

Next we wanted to analyze data to find pollution sources and examine trends. This is where it got a little complicated: There's more than one database with emissions information, and some of the data overlap. Some databases track emissions for which there are no national air standards, while others track emissions for which there are standards. And, because reporting rules have changed over the years, it can be difficult to identify yearly trends.

We started by analyzing the granddaddy of all pollution inventories: the EPA's Toxics Release Inventory, which tracks emissions of some 650 different chemicals. This data is available on the Web at *www.epa.gov/tri*, and can be easily sorted and manipulated using the EPA's Web-based interface; it can also be downloaded and put into Excel for additional analysis. We did both.

When we examined countywide trends, we limited our analysis to what the EPA calls its "1995 core chemicals," and "original industries." This gave us most of the chemicals and most of the industries, excluding power plants, which we analyzed separately. By slicing the data this way, we found that overall TRI

emissions in Louisville-Jefferson County had declined 39 percent between 1988 and 2001.

We had a spreadsheet for each year and combined the data into a single worksheet to make it easier to look at year-to-year trends. We used Excel pivot tables to quickly summarize data for each company.

When we examined the emissions trends from the top 25 TRI sources, we used the EPA's Envirofacts Warehouse, a very useful online resource, at *www.epa.gov/enviro*. Its emissions data is more current, and includes yearly releases for each facility on one screen. After a few points and clicks, a color-coded trend chart popped up on our computer screen.

TRI, however, fails to account for all emissions. Other categories of pollutants are reported by industry and compiled into different databases that can be obtained from the EPA, and sometimes, state and local environmental agencies. We obtained ours from the local air pollution control district.

The agency at first told us that it didn't keep the records in a form that our computers could understand. But they agreed to extract the data we were requested and gave it to us electronically. Some came in an Excel spreadsheet; others in text tables.

To get the data in a useful form, we used a Perl script to loop through the text files, pulling out the name of each plant and the pollutants associated with it.

With this data, we used Excel to rank each source of industrial pollution for each category of pollutant, and determined yearly trends. Some of what we found we expected, such as the major emissions from power plants. But we also found that a bourbon distillery was the largest source of volatile organic compounds, a contributor to summer smog. The alcohol vapors that escape from aging casks of Old Forester and Early Times are called "the angel's share" by industry insiders and are unregulated.

We ran our results by the local air pollution officials as well as some of the companies that came out on top. We then realized that one of the databases we had was faulty. The computer program that local air officials had used to extract the data on hazardous air pollutants had left some industries out. The local officials fixed the problem, and we ran our analysis again.

With stories, maps and charts, we presented a complete picture of industrial air pollution sources in Louisville. While much attention over the years had been focused on an area with 11 chemical plants, our analysis identified major sources far from there. We found that a single power plant in Indiana just across the Ohio River from Louisville produced more TRI emissions than two Louisville based power plants combined. We found that particulates, which can damage lungs and cause cancer, had increased by 76 percent since 1988, and that the hazardous air pollutants were on the rise since 1998.

At the same time, we showed that industrial emissions of volatile organic compounds had declined by 60 percent since 1988.

We used graphics and maps to help tell the story and keep most of the numbers there. Readers could identify the plants nearest them, and then see whether air emissions were going up or down at those facilities.

In ArcView, we made a map showing the location of the 50 counties in the U.S. with the highest air emissions based on the TRI data. We used ArcView's graduated symbols feature to show the relative size of each of those counties' emissions. We wanted to compare our county to others in the nation, and found Louisville-Jefferson County ranked 23rd.

We told the story through the people who work in these plants and those who live nearby. We also followed up the special section with a package on how University of Louisville medical experts had identified certain "asthma clusters" in Louisville, and that many people with asthma in Louisville blame poor air quality on making their medical condition worse.

Contact James Bruggers by e-mail at jbruggers@courier-journal.com.

Contact Mark Schaver by e-mail at mschaver@courier-journal.com.

readme.txt

Read the *Courier-Journal* air pollution stories on the Web at *www.courier-journal.com/cjextra/2003projects/toxicair/*.

The federal government provides information about pollutants and toxics on some of its Web sites. The Environmental Protection Agency offers fact sheets about the health effects of air pollutants at *www.epa.gov/ttn/atw/hapindex.html* and the Integrated Risk Information System database of information about possible human health effects from various substances at *www.epa.gov/iris*. The Agency for Toxic Substances and Disease Registry Web site contains toxicological profiles of chemicals at *www.atsdr.cdc.gov/toxpro2.html*.

You can find the benchmark data that the *Courier-Journal* used in its analysis of possible health risks at the EPA Philadelphia regional office Web site (*www.epa.gov/reg3hwmd/risk/index.htm*) and the San Francisco regional office Web site (*www.epa.gov/region09/waste/sfund/prg/index.htm*).

SPOTLIGHT: POLLUTION

Squeezing trends from air emissions data

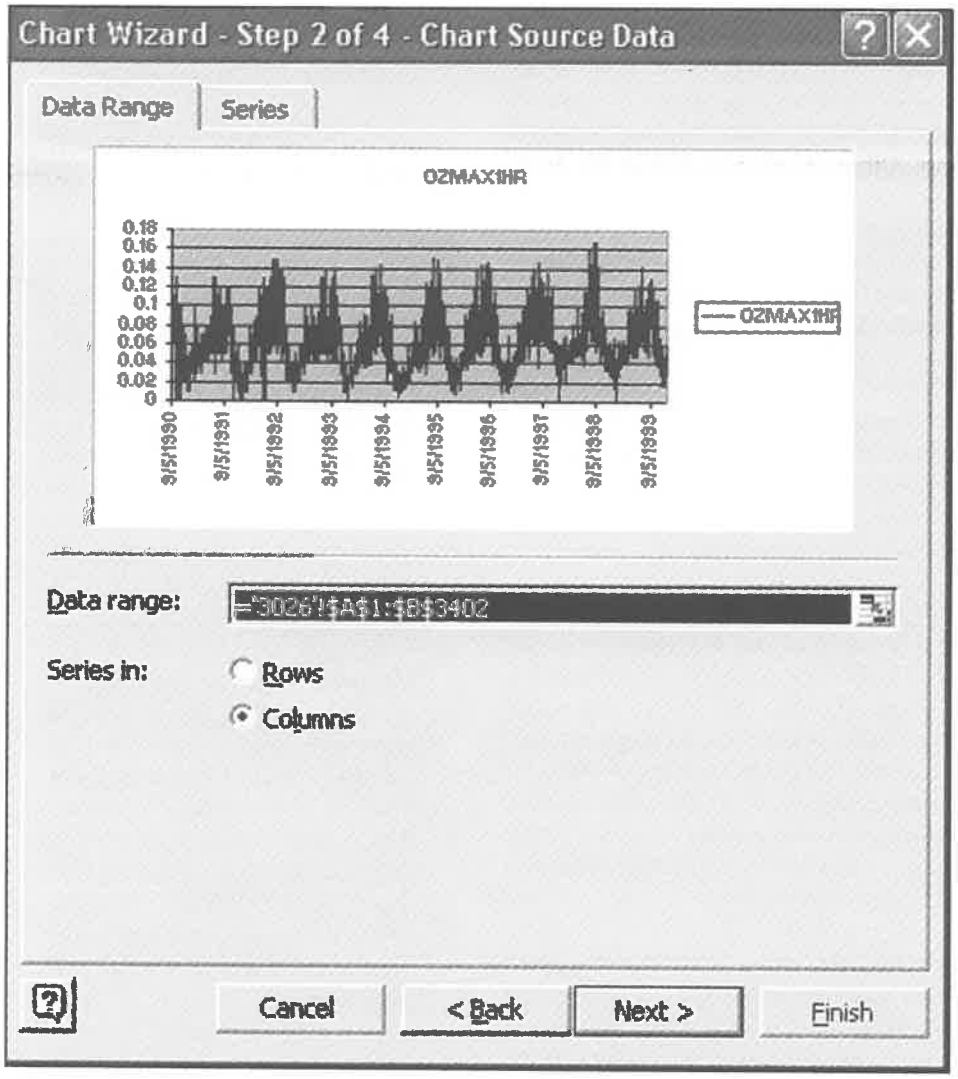
By Russell Clemings, *The Fresno Bee*

Even the experts can't agree on the best way to find long-term trends in air pollution data. But we had to figure something out for our "Last Gasp" project, which made the point that air pollution was worsening in parts of California's Central Valley even as it improved in Los Angeles.

Here's what we did. Our data came from California's Air Resources Board, which offers a free CD with data, maps and a nifty interactive program for displaying both. The board has a Web site (*www.arb.ca.gov/aqd/aqdcd/aqdcd.htm*) with downloadable hourly monitoring data back to 1980.

Most of the data is in dBASE files that can be imported into Microsoft Access with a couple of keystrokes. From there, you can construct queries to extract data for a single monitoring station (station ID numbers and locations are in a separate file) and export the resulting data to Excel.

The files have readings for a dozen pollutants, presented more than 20 different ways. Unfortunately, the raw data is just a jumble of numbers. Even if you scanned the entire file from top to bottom, it would be impossible to figure out the trend.



You can use the Excel Chart Wizard to create, for example, a line graph for the date and one-hour ozone levels, but even that doesn't help much. (see graphic on this page)

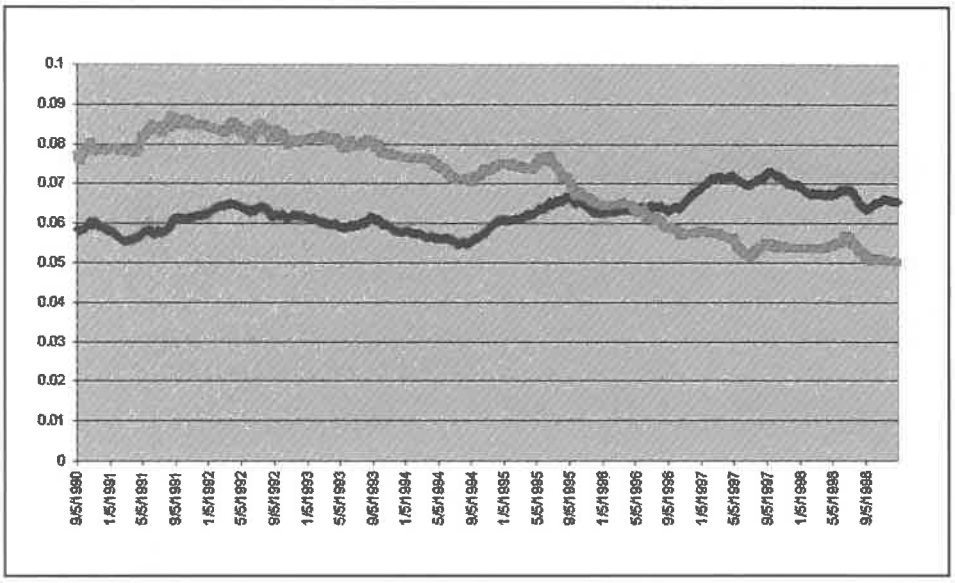
Because ozone is a summertime pollutant, the levels are high in the summer and approach zero in the winter. If you eyeball the graph, it looks as if the summertime peaks are edging upward, but it's hard to say for sure.

What we needed was a way to boil down these peaks and valleys to a long-term trend.

One of the ways that statisticians find patterns in data is to do a linear regression, and Excel fortunately has a built-in way to add a regression line to the chart we built. After you've finished the chart, just right-click on the data line and click Add Trendline, and you can lay down a nice, clean regression line over the data.

Doing that may show what appears to be a long-term rising trend in the data. There's just one thing wrong: As any competent statistician will tell you (and lucky for us, one did), a simple linear regression doesn't work with data that involves a time series like this one. The seasonal pattern in this data is so strong that it overwhelms any other pattern that a regression analysis might show you.

This problem is not unique to air pollution data. Economists run into it when they're studying things like retail sales, which rise and fall with the calendar. And there are tools available for dealing with it, such as the Census Bureau's X-ARIMA model. But even if you have the intellectual capacity to use something like that, you're still left with a significant problem – explaining it to your readers. Just saying, "Trust me, my computer says there's a trend here" probably isn't the best approach.



Here's another alternative: A 365-day (or annual) moving average. It's not too hard to explain: for each day, it's just the average of the past year's daily values. That's why it's called a "moving" average. If today is June 5, 2003, then the 365-day moving average for today is the average of all daily values from June 6, 2002 through today. Tomorrow's 365-day moving average, in turn, would be the average of daily values from June 7, 2002 through June 6, 2003.

Here's how you write a formula to get that from Excel. Suppose you have date labels in column A and data for several years in column B. Calculate the first annual average in cell C366 with this formula: =average(b2:b366).

Copy that formula down and graph the result, and you can say with some confidence that the 1990s saw one-hour daily ozone peaks decline edge upward or downward, or remain about the same, depending on the data. The above chart shows contrasting trends (using the 365-day moving average) at two stations — one in southern California that is declining and one in the San Joaquin Valley near Fresno that is rising somewhat.

At this point in our reporting, we are ready for quality control. When we did this in real life, we showed our graphs to various experts at the state and local air pollution control agencies. We also explained how we did the analysis and asked for their comments on it.

All agreed that the 365-day moving average was a valid way to show the long-term trend. But one expert went a step further. He encouraged us to throw out the winter values and look only at the summer data. His reasoning: Although no one knows for sure, it's considered likely that very low levels of ozone are not a health threat. The state's one-hour ozone standard of 0.09 parts per million is exceeded regularly in the summer. But in winter, typical levels are far less than that. And an increase from 0.02 to 0.03 is clearly not as important as an increase from 0.12 to 0.13. Yet our 365-day moving average would regard them both as important.

REQUIRED READING For Your Newsroom

Unstacking the Deck:
A Reporter's Guide to Campaign Finance
by Michael A. Weber, Aron Pilhofer and Derek Willis

This invaluable guide for pursuing stories about the impact of money on elections, political parties and candidates at the federal, state and local levels is packed with story tips, resources and Web sites.

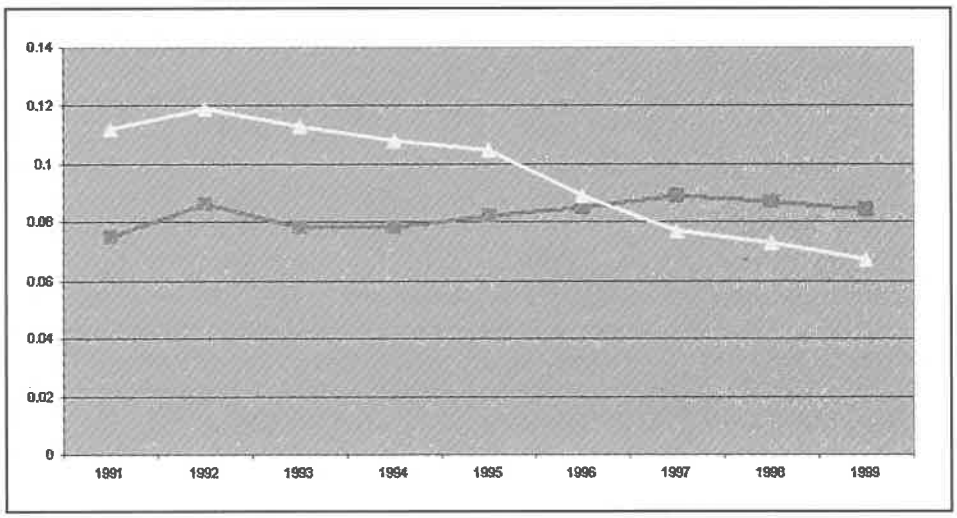
ORDER TODAY!
Call 573-882-3364 with your VISA or MasterCard
-OR-
Visit our Web site at **www.ire.org/store**
for online ordering or order form downloads

IRE MEMBERS: $15 each • NONMEMBERS: $25 each

The expert suggested a number of alternative approaches. We finally chose one that was doubly elegant, being both easy to calculate and easy to explain. For each year, we calculated and plotted just one value – the average daily peak for the summer smog season, May 1 through Oct. 31.(see chart below)

You can read the "Last Gasp" project on the Web at *www.valleyairquality.com.*

Contact Russell Clemings by e-mail at clemings@cris.com.



Tech tip...

Making sense of Navy aviation crash data

By Richard O'Reilly, *Los Angeles Times*

The Naval Aviation Safety Database was one of the research sources for the *Los Angeles Times'* Pulitzer Prize-winning series on the U.S. Marine Corps Harrier attack fighter jet. The four-part series, published Dec. 15-18, 2002, also won an IRE Certificate.

Reporters Alan C. Miller and Kevin Sack found the bulk of their material through numerous interviews and nearly 90 Judge Advocate General accident reports obtained under Freedom of Information Act requests. But an analysis of the Harrier through the Naval Aviation Safety data provided scope, context and nuance to the project.

Brief summaries from the database are available on the Navy Safety Center Web site, *www.safetycenter.navy.mil*. The full database must be requested via FOIA from Josie Bowden, 757-444-3520, ext. 7045 (JBowden@safety.center. navy.mil).

It took us months to get the database after filing the request. And when we received it, we found that data from 218 fields, about 27 percent of the 807 fields then in the database, were omitted on privacy and national security grounds. An appeal failed. What we got instead was the brief public summary of recent Harrier accidents supplied by e-mail. It was exactly the same data we could have copied from their Web site, which was useless for analytical purposes.

Virtually all of the extensive narrative fields contained in the database were excluded. That had the effect of removing all conclusionary information. Personnel data on injuries, causes and even the severity of injuries was withheld. Gender, date of birth, rank and unit were provided. But the field containing first and last initials was omitted.

Also missing were fields that identified recurring problems, identified instances in which planes were not in compliance with Marine standards and in which problem parts had been cannibalized from other aircraft. A field that provided a detailed identification of the component damage or failure was omitted. However, a field that identified materials or components involved in the incident was provided.

Data on Marine aircraft are contained in the Navy database because Marine funding is routed through the Navy, and the Navy is responsible for maintaining safety data for the Marines.

What the Navy eventually provided was a complex, well-documented, clean set of related data tables in tab-delimited format. There is a mix of data tables and code tables. Many of the fields in the data tables are coded. Code translation tables were provided for each of the coded fields we received. However, the translation tables for coded fields that were omitted also were omitted.

A large-format table relationship sheet was provided, as was a table list text file and a data dictionary text file. The table list (with the file name av_tablist.rtf) gave a simple list of table names, field names, types and sizes and whether the field was omitted or not. The dictionary file (av_dicty.txt) gave an English translation of field names plus an explanation. For example, the field "bank_degs_left" in the table "av_acft_at_esc" was described as follows: "LEFT BANK DEGREES~The left bank of the event aircraft in degrees at the time of escape."

Where fields were omitted in tables, the field place was held by a null value (hex 00) between the tab characters. That caused problems for me when reading the data into SAS. Therefore, I copied the original tables and then, using UltraEdit, did a search and replace across all the file copies in one operation to translate a tab-null-tab string into a tab-space-tab string.

Some of the code fields proved to contain hierarchical codes, which were not explicitly documented. An example is "incdt_phase_ops_c" in the table "av_incdt_occurn." It is described in the dictionary as "Significant Incident Phase Operation~The phase of operation in which the event aircraft was engaged, at the time the coded subject significant incident occurred, in the evolution of the reportable event."

An example entry in the three-character field is "33Z." The code translation table renders that as "No further breakdown," which is meaningless. In the same code translation table, however, you find that "3" translates as "Takeoff forward motion on run/catapult until attaining climb schedule." And "33" translates into "Aircraft breaks ground climb schedule not yet attained." In other words, that single three-character field can be decoded into three separate, related meanings.

There are other examples. The five-character field "imc_c" in the table "av_eng_cmpnt" is described in the dictionary as "Involved Material Component~Identifies the Naval

Safety Center code relating to the category of the event aircraft's component or engine (power plant)."

One such entry is "T1007", which is translated in the field code table as "Compressor pressure limiter section." The "T" alone translates into "Turbo-jet engine," and the first three characters, "T10," mean "Regulator-not detailed." Thus the code must be parsed to gain its full meaning.

My solution for all of the hierarchical codes (which were identified by inspecting the code table translations for partial string definitions) was to create new fields in the table to contain just the partial strings that had separate meanings. Thus I created an "imc_a" field that was only the first character of "imc_c," and an "imc_b" field containing the first three characters of "imc_c." This was easily done because all of the hierarchical fields contained the full complement of characters specified in the table list and dictionary files.

The Naval Aviation Safety Database contains records (more than 48,000 in the mid-2002 version we received) of "events" from Jan. 1, 1980, to within the last several months. An event can range from a no-damage, no-injury maintenance or operational problem that could have a safety impact and thus meets the criteria for reporting, to loss of an aircraft and its crew. The most dramatic events are those known as "Class A," with at least $1 million in damage or fatal injury.

There are three primary trajectories for analyzing the data — events, persons or aircraft — each of which utilizes its own key field for table joins. Within those trajectories there are subsidiary keys that join more detailed tables. As a result, you can end up with a lot of tables joined together, which can lead to a lot of rows being added as details embellish what started out as a single row for a single event. Much of the data will be constant and repeated on those rows, such as the total cost of the event, or the event category, or even the tail number of the aircraft. Other fields on the rows will contain unique data, as when several different components are involved, causing several rows from the component table to be joined to the aircraft table.

Ultimately, it is necessary to recode and collapse data to analyze it accurately. There is a lot of missing data, beyond what was withheld. The data collected has undergone a number of revisions since 1980. Later events have more complete data than earlier events. Sometimes data is missing because it is not relevant to a particular event.

To compute accurate percents for occurrences, the denominator has to be just the events for which the data exists. For instance, some reports for Class A events involving the Harrier did not have any data on the components involved. I created a new logical field called Parts, with "Yes" meaning there was parts data and "No" when it was missing. Then when calculating the percentage of jet engine problems, for instance, I divided by the count of Parts = Yes.

In some cases I forced empty records to sort at the bottom of a list instead of the top by recoding a blank field to "zzz."

If the summary being created involves a number of fields, it probably will be necessary to make several intermediate tables by doing record-by-record inspections and editing. That's because it probably will not be possible to sort across several fields and end up with the data desired from each field rising to the top of the sort.

The most important advice I can give is to be skeptical of every result you derive and try to figure out an alternate way to get it. If you come up with the same result two different ways, it is probably reliable. But if you get different results, you'll need to figure out why and fix it.

Contact Richard O'Reilly at richard.oreilly@latimes.com.

readme.txt

"The Vertical Vision" received a certificate in the large newspaper division of the 2002 IRE Awards. To order the stories and the contest entry form, which contains detailed information about the reporting of the project, contact the IRE Resource Center at 573-882-3364 or rescntr@ire.org and ask for Story No. 19777.

IRE offers a detailed tipsheet (No. 1850) about analyzing the Navy crash data. IRE members can download the tipsheet from the IRE Resource Center Web page (*www.ire.org/resourcecenter*).

For more about the reporting of the series, read "Harrier jet bedeviled with mechanical ills" in the May/June 2003 *IRE Journal*.

The series and a related multimedia package are available on the Web at *www.latimes.com/news/specials/harrier*.

CAR TOOL

Using ParseRat to carve up names

By James E. Wilkerson, *The (Allentown, Pa.) Morning Call*

If you've ever sat in on a session about data cleaning at one of the annual IRE and NICAR computer-assisted reporting conferences, you've heard one question that comes up again and again: "What about names?"

Many of us have gazed longingly at a huge table containing jumbled names, wondering what sort of havoc we could wreak if we could just break them up into more manageable fields.

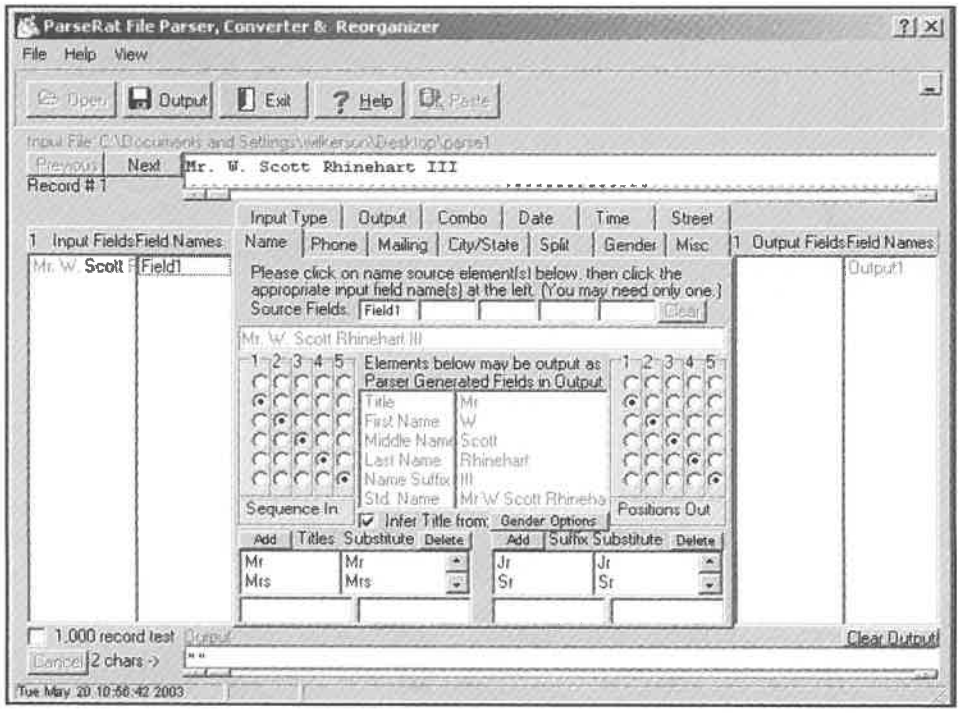
A $50 program called ParseRat can do that by flashing through thousands of names in a few seconds to produce a nice, clean list parsed however you like.

While I've primarily used ParseRat (*www.guysoftware.com/ParseRat.htm*) to carve names, the program is capable of doing much more. The compact program file takes only a few minutes to download but can handle very large files in a variety of formats, from the antiquated EBCDIC data exported from mainframes to data-heavy Web pages using modern XML tagging.

ParseRat steps

For example, you can easily grab large tables from Web pages and, with the help of scripts included with the program, modify them for importing into a spreadsheet or database manager. Or you can use the program to parse data from a simple .dbf or complicated COBOL file. In fact, you can grab text in just about any format: delimited, fixed-width, "tagged" with HTML, XML or other coding, among others.

There are routines for converting unusual date and time formats. The program can even convert from one measurement system to another, such as metric to U.S. standard, and allows you to write new conversion rules for measurements not included.



rules for measurements not included.

ParseRat will even take a stab at determining gender from names, and will generate Soundex codes useful for searching for names phonetically. But it's obvious from the start that ParseRat was created with mailing lists in mind and there are a bunch of built-in options to parse names, addresses and phone numbers.

The program's major drawback: The interface is somewhat complicated, with tons of options available through small buttons in an often-jumbled package. The help files have a homemade feel that can be confusing at times. But after spending an hour or so with the documentation, I was able to navigate around the program with relative ease.

First, preparation

You should also be ready to do some prep work on your data before running it through ParseRat. I use a powerful but inexpensive text editor, UltraEdit, available for $35 at *www.ultraedit.com*. There are a number of text editors out there and having one is a must, because they can crank through very large text files that would choke a conventional desktop word processor such as Microsoft Word or the trusty Windows Notepad.

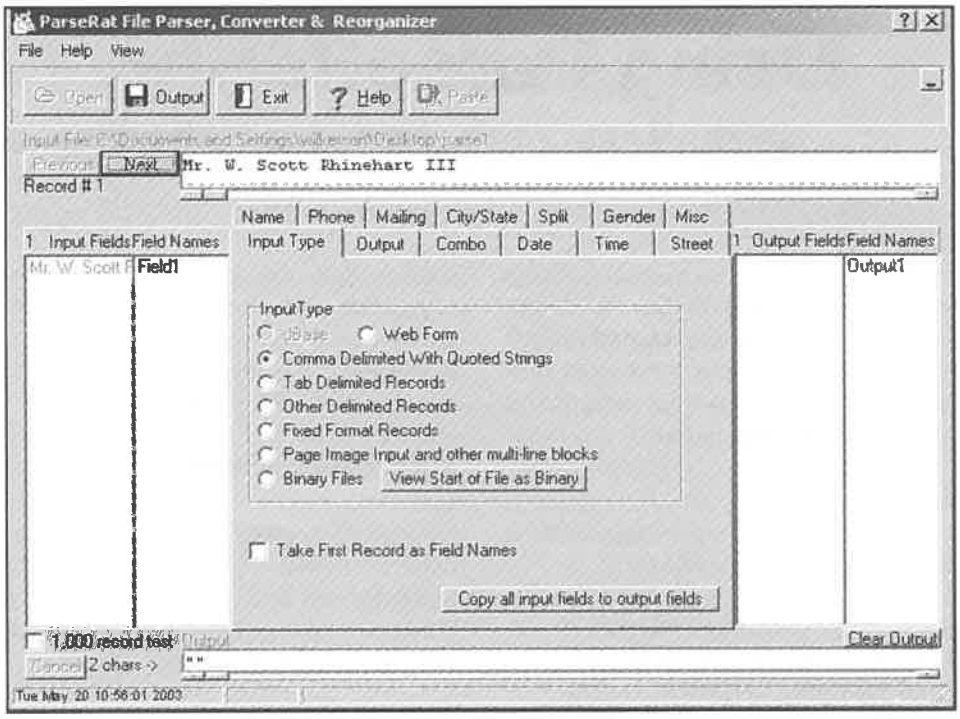
Before using ParseRat, make sure all header information and extraneous material has been stripped from the top of the text because the program will use the first line of data to make judgment calls about parsing the entire file. The cleaner the file, the better the output, so take some time to ensure everything's as uniform and as clean as possible. I recommend isolating only those fields on which you want to work to keep things simple and speed up the process.

Here's a walk-through on parsing names, which I do regularly.

First, use your text editor to choose the most complicated name you can find and move it to the very top of the

list. By first showing ParseRat a name with the most elements – "Mr. James E. Wilkerson III," for example – you prime the program to look for all of those elements in each record. And, because you only get to see the first record when making parsing decisions, it's easier to do with a name that contains all of the possible elements.

Sometimes it's easier to make up that first name than dig through your data to find the perfect model. If you make up a name just remember to delete it later.



Software interface

After that, fire up ParseRat and choose Open\Input file and browse for the text file. The next screen will ask you to confirm the type of data – delimited, fixed width, etc.

The interface gets pretty clunky here, but it's easy enough to navigate once you get the basics down. We will focus on the Input Type, Name and Output tabs.

After confirming the text type, click on the Name tab. This is where you specify the field (or fields) containing the name you want to clean, and tell ParseRat what you want the output to look like. To the left of the tabs, you'll see an input selection to the left and an output indicator to the right. (You should have at least one field visible in the input section – if not, go back to the Input Type and try another type until you do.)

Near the top of the dialog box, click on the first box available in the source fields input. Then, click on the appropriate source field to the left. The name of that field will appear in the source fields box, and the first name in your text will appear in the elements boxes underneath.

ParseRat takes a stab at determining the order in which the elements of the name appear, and ranks them from one to five using the radio buttons to the left of the sample name, starting with "Title" and ending with the full name. You can change that order by clicking on the buttons. You can change the output order by using the similar feature to the right of the sample name.

After setting up the input and output order on the name tab, you need to specify the fields to output. Click on the Output tab to get a lot of options. The most important is the parser-generated fields, which show you the output that ParseRat will generate. Scroll down and take a look at the results to make sure they match your expectations.

To the right of parser-generated fields is a small button labeled "Add Field." Clicking on that button will add a blank output field to the right of the dialog box. Count the number of fields you want to output and create that many blank fields with that button.

Choices, choices

Click on the first output field at the far right and choose the parser-generated fields name that you want to place there by clicking on it. Click on the next output field name and do the same, repeat until you have all of the output fields you need in the order you want.

If you want to keep the original record, place it in one of the output fields by clicking on it at the far right of the dialog box.

Make sure you have selected the appropriate output format – I usually use comma-delimited text – and click the output button in the top tool bar. Then sit back and wait while ParseRat generates your cleaned names.

It should be obvious when you start poking around on it that the program is capable of much more than I touched on here. Take some time and read the documentation, and you'll soon be flying through parsing jobs you might have thought impossible.

Contact James E. Wilkerson by e-mail at james.wilkerson@mcall.com.

readme.txt

Have a favorite CAR tool that *Uplink* readers should know about? Drop us a note at uplink@nicar.org and tell us about it.

OPEN SOURCE

New Linux offering great for serving data

By Aron Pilhofer, *Center for Public Integrity*

It used to be that all the tools a journalist doing computer-assisted reporting needed were a spreadsheet, database manager and a decent SQL book. But in these days of networked newsrooms with a PC on every desk, the CAR specialist's job has morphed to include such diverse skills as Web design, server-side scripting, database administration and data warehousing.

An increasing number of reporters and editors want access to data directly. They want the ability to do quick queries on multiple databases through an easy-to-use interface, such as an intranet Web form. The more tech-savvy journalists want to do their own data analysis using more advanced tools, such as Microsoft Access database manager.

Stand-alone database managers (Access, FoxPro and the like) aren't well suited for the task of sharing data. For that, you need a central server capable of making data available over a network.

Which brings us to Red Hat Linux 9.0, the latest version of the most popular Linux distribution on the market. After more than two months working with Red Hat 9.0, I can say that it is by far the best distribution of Linux ever released and an ideal solution for any newsroom looking to centralize its data operations at a fraction of the cost of other alternatives.

Assuming you already have the hardware on hand, a very basic Microsoft Windows/Internet Information Server/SQL Server installation will range from $2,500 to more than $10,000 (not including support) depending on features desired and number of users.

Red Hat 9.0 can be downloaded for free. A boxed version costs $39.95 and comes with printed manuals and 30 days of Web-based support. Red Hat also sells higher end enterprise editions (the ES series) that range from $350 to $800 and come packaged with up to a year of live telephone support.

Like all open-source software, Red Hat's license allows administrators freedom to install it on an unlimited number of servers accessible to any number of networked users. Microsoft generally requires a separate license for each user or system on which its software will be installed.

As for support, both Microsoft and Red Hat charge for it. But while some kind of formal support is a nice safety net, it isn't absolutely necessary for Linux any longer because the installation, configuration and administration tools have improved so much in the past few years.

And among the new breed of simpler, easier Linux releases, Red Hat 9.0 is without a doubt the best.

Installation

The boxed version can be purchased at any major computer store, but the three installation disks can be downloaded from the Red Hat Web site, *www.redhat.com/download/products.html*. Be aware that downloading will take several hours even with a very fast Internet connection. The boxed version might be worthwhile for those new to Linux because of the printed documentation and support.

Once you have the disks, installation should be a simple matter of booting from the CD and following on-screen instructions. I installed Red Hat 9.0 on two desktop configurations and a laptop without a hint of trouble. Users will most likely be able to simply select all the default settings and continue.

Although there may be a few speed bumps along the way – especially for new users – the Red Hat Web site and printed documentation should cover these topics adequately. If not, there are several excellent Web sites that cover Linux basics, including the Linux Documentation Project (*www.tldp.org*) and Linux Planet (*www.linuxplanet.com*). It might help to have an information technology person nearby just in case.

Red Hat has made selecting the right mix of software a snap. Users can initially select from several broad installation categories such as server, workstation, desktop, and the appropriate software for that function will be installed. Users can also customize their installation by selecting among the hundreds of individual applications that come packaged with Red Hat.

Once the software has been selected, the actual installation should go automatically. Reboot and you should have a working Linux machine.

Software

Red Hat 9.0 includes the two critical applications any newsroom data center would need: a Web server and a high-end database application.

Apache 2.x is the standard Web server packaged with Red Hat 9.0, and it's outstanding. It is by far the most popular Web server in the world. According to the most recent statistics, roughly two of every three Web sites run Apache.

It's popular because it is secure, stable and easy to set up. The default configuration probably will work just fine for internal newsroom use but, if users need to tweak any settings, that can be done by editing a master configuration file.

Red Hat 9.0 offers several database management solutions. The best is MySQL, which is optimized for speed, stability, reliability and power, and ideal for serving data. (For details see "My SQL primed for heavy-duty action" in the March-April 2003 *Uplink*).

Goodbye stand-alone Access databases and random .dbf files floating around the

network. No more problems with corrupted data on deadline. So long to worries about users accidentally deleting or altering data. MySQL, like any high-end database server, allows administrators to give users just the amount of access they need and no more, or the ability to query a database but not to update or delete records.

Red Hat 9.0 comes packaged with a number of scripting languages – PHP, Perl, Python, Ruby – and development tools that allow you to combine the power of Apache and MySQL to create simple Web-based database search pages.

Although there are many ways of installing software under Red Hat, including compiling from source code, the simplest way is to use prepackaged applications called RPMs. Just as Windows has its .exe, the software packaged for Red Hat machines carries the .rpm extension. Several years ago, the company created a special program to handle RPMs called the Red Hat Package Manager, which works a little bit like the add/remove programs control panel in Windows.

RPM did make installing software easier, but it was far from perfect. As long as the user stuck to programs bundled with the distribution, things were fine. But users sometimes want to install more bleeding edge versions of programs, or applications that were not packaged with the system – and that could cause problems.

Red Hat 9.0 solves this problem by simply removing it. The new package manager will not install anything but the applications that come with the Red Hat distribution. Users can install other RPMs, but that must be done from the command line, not through a graphical interface – an excellent compromise for newer users and power users alike.

Maintaining Red Hat

In addition to being easy to install, Red Hat also is simple to maintain because of the Red Hat Network. All users (even those who downloaded the software) are given access to it. RHN is similar to the Windows "update" feature but much more powerful because it updates not only the Linux operating system but any standard installed software.

Red Hat is quite good about posting bug fixes and updates to the Red Hat Network almost as quickly as they are released, thus ensuring that your system is as up to date as possible.

Managing your server is easy and secure with Webmin (*www.webmin.com*), a Web-based administration tool that is bar none the best available. Administrators can administer a Linux server over a network or even over the Internet, although that is not recommended for security reasons.

Webmin is simple to install, and very intuitive. It is also open source, meaning it is totally free. There are modules available to help manage everything from a firewall to user accounts to disk partitions. Webmin, unfortunately, is not among the standard packages that come with Red Hat 9.0, but its excellent documentation can be downloaded from the Webmin Web site.

The major problem with Red Hat 9.0 is its slightly out-of-date version of MySQL. The 4.x release, which adds Union queries and full support for full-text indexes, is now considered the production version. But Red Hat still inexplicably packages the older 3.2x version of MySQL.

Upgrading to the 4.x series is not as easy as installing Webmin and, if you have an existing installation of MySQL on your machine, it can cause problems. There is excellent documentation on the MySQL Web site (*www.mysql.com*) that covers everything you'll need to know to install the latest version of MySQL.

Conclusion

Red Hat 9.0 ranks as the single most painless installation I have ever seen in my 20 years of computing. It won't cost you a cent to download, but offers the power and features of systems costing tens of thousands of dollars.

Contact Aron Pilhofer by e-mail at apilhofer@publicintegrity.org.

CALL FOR ENTRIES

2003 IRE AWARDS

The annual contest of Investigative Reporters and Editors, Inc.

CATEGORIES

Newspaper

Television

Online

Other Media:
Magazine/Speciality Publication
Book
Radio

Special Categories:
Renner (crime reporting)
FOI
Student

Postmark DEADLINE
January 12, 2004

The contest recognizes the best investigative reporting in print, broadcast and online media, and helps identify techniques and resources used by entrants.

For entry forms and detailed information, visit our Web site at
www.ire.org/contest

Pulse

continued from page 1

assisted reporting specialist at the newspaper, and I sifted through state and federal databases to identify who was still harming the sound and checked whether regulators adequately protected the public and environment. The results were surprising.

When it comes to measuring water pollution, the sediments tell the story. We found data from plugs of mud that were sampled and analyzed to look at historic pollution trends. Changes in the pollution levels nicely followed the growth of Seattle, the shift from coal burning to electricity and even testing at the Hanford Nuclear Reservation in eastern Washington.

We looked also at current sampling data, which was collected by numerous agencies, including the National Oceanic and Atmospheric Administration (NOAA), U.S. Geological Survey, the Army Corps of Engineers and state ecology and wildlife agencies. Primarily, we worked with NOAA and state Ecology Department data, identified in scientific papers and provided to us by the authors for free upon request, or found online in association with research conferences.

Then the challenge was making parts per million meaningful. We compared the readings to sediment cleanup standards for Superfund and state programs (*http://epa.gov/waterscience/cs* and *www.ecy.wa.gov/programs/tcp/smu/sediment.html*). The standards helped us distinguish between severe pollution and relatively benign contamination. Then we checked out some risk information. NOAA has a Web site including a chart showing the risk posed by pollutants at different concentrations for fresh and salt water at *http://response.restoration.noaa.gov/cpr/sediment/sediment.html*.

We didn't find much that was shocking, but were able to calculate the number of square miles of contamination in the sound. We found that the pollution was getting better overall and that the main backsliding came from pollution associated with cars, not industry.

These results led us in search of the pollution sources – where did this stuff come from, and was it still pouring in? We used databases of cleanup sites to identify old polluters and places where contaminants could be leaching from tainted sediment. Some of those locations had been placed under the federal Superfund program, others under state regulation. The U.S. Environmental Protection Agency has a searchable site for some federal cleanups (*www.epa.gov/superfund/sites/npl/locate.htm*), but that didn't provide the data we needed for our analysis.

For that, we went again to the Ecology Department. We downloaded from its Web site an Excel file called the Confirmed and Suspected Contaminated Sites (CSCS) Report. It had loads of information including what the pollutant was; if it was in the land, air or water; the progress of the cleanup and, roughly, how long the contamination had been known about.

While still in Excel, we sorted out the counties in the Puget Sound watershed. Then we imported the spreadsheet into Access, creating a new table so we could run queries selecting the chemicals of greatest concern – PCBs, heavy metals, dioxins, polycyclic aromatic hydrocarbons (PAHs) and phenols – then tallied how many sites were poisoned with these most worrisome chemicals and where they were in the cleanup process. A query showed that cleanup had not started at more than 200 of 1,400 sites. We could also look at the sources of some long-banned chemicals, particularly PCBs.

For current contamination sources, we looked to the federal National Pollutant Discharge Elimination System (NPDES) program, which regulates water pollution released from industrial and municipal facilities, and storm water pollution from local governments, construction sites, industry and sand and gravel operations.

The amount of pollution released is supposed to get ratcheted down with each five-year permit renewal to achieve the Clean Water Act's goal of eliminating discharge into waterways. When renewals are delayed – we found some decades-old permits – pollution remains high.

The Ecology Department provided the data at no charge, though Olsen did yeoman's work getting it in a format that we could use – first dealing with an Oracle file in a version we couldn't use, then just getting it as a text file that she turned into an Access table. From that, we pulled out the locations for rivers and streams in the sound's watershed, as well as those that dumped directly into it. That left us with more than 170,000 records tracking every release from 1999 to mid-2002, including the pollution limits, violations and cause of the violations. The data also had latitude and longitude coordinates, so we could plot the sites in ArcView 3.x to show how widespread the polluters were and that even upland locations miles from the sound could hurt its health. We also mapped the most frequent violators.

The data was fraught with errors, mostly from the incorrect use of codes that indicate why violations occurred, giving us numerous false violations. We ended up compiling our list of top offenders and passing it over to the Ecology Department to check against paper records. We also discovered – happily before publication – that the department tracked the industrial sites separately and had failed to include them in our request, basically telling us they'd forgotten it.

It turned out that the biggest water quality violators were not industrial facilities, but public sewage treatment plants. In fact, by counting in Access the number of permit violations per facility, we found that a pristine park that

is a popular tourist destination was operating an outdated sewage plant, placing it among the top violators. By simply using the "find" function, we then searched an Excel file of penalties the Ecology Department had levied in recent years and discovered that the department went easy on the government entities that were top violators.

One shortcoming we found with the permits data was that pollution was tracked by concentration, not total amount, so it was impossible to tally the volume of pollutants released. EPA's Toxic Release Inventory totaled pollution annually by pound, but the data includes only select categories of industrial polluters and an abbreviated list of chemicals that changes over time. Despite these shortcomings, we created a graphic showing the number of pounds of pollutants released by the largest facilities, noting that even with this limited data, more than 1 million pounds of chemicals was dumped into the sound annually.

Storm water discharge, perhaps the greatest water pollution threat, is one of the trickiest to measure. Most NPDES storm water permits lack discharge limits or pollution monitoring requirements. Some cities and counties will conduct their own tests, though the most comprehensive study we could find was a 1995 report from a local public utility that listed what was in the water and whether it violated federal standards.

We also checked for damaged waterways that have Total Maximum Daily Load (TMDL) restrictions limiting the pollutants legally dumped into them. These so-called 303(d) list waterways can help illustrate long-term effects of pollution. The EPA maintains a list of the sites (*www.epa.gov/OWOW/tmdl*), but it would be worth checking with them or the state agency managing the list to make sure the online version is up to date.

After figuring out what we could about pollution in the environment and where it was coming from, we wanted to know what that meant for the critters soaking in it, plus the people eating the local seafood. We also were curious about the aquatic life population numbers.

We tracked down data collected by our state Fish and Wildlife Department on mercury and PCB contamination from around the sound for a variety of fish. This data was provided after many phone calls and with much trepidation. We promised to discuss how we were using the numbers, making sure we were comparing measurements appropriately, and finally they sent us an Excel file with more than 100 records. The data included latitude and longitude coordinates, so we mapped the locations where samples were collected.

As we expected, the map showed that fish from industrial areas were more contaminated than their country counterparts. But we found a surprise: not only did the long-lived, bottom-feeding fish in industrial areas show higher levels of contamination – the salmon did too.

Finally, after looking at pollution in the sediment and in the fish, we wanted to know how sea life fared in numbers. Using state catch figures, we plotted tons of fish caught over recent decades using Excel. We created fever charts that showed, in species after species, populations in deep decline.

Putting all of these measures together we were able to present a holistic view of a water body in a precarious position, one whose fish population was edging to the point of no return unless changes were made quickly to nurse the sound back to health. Since the series ran, the governor has pledged to spend $100,000 on orca recovery efforts, Ecology Department officials have proposed stronger water quality and shoreline development regulations, and funding has been secured for a rescue tug to protect against shipwrecks.

Contact Lisa Stiffler by e-mail at lisastiffler@seattlepi.com.

Computer - Assisted Reporting Boot Camps

These unique seminars give journalists a jumpstart in computer-assisted reporting techniques. Participants come to Columbia, Mo., where they are trained in how to acquire electronic information, use spreadsheets and databases to analyze the information and to translate that information into high-impact stories. The National Institute of Computer-Assisted Reporting provides follow-up help when participants return to their news organizations.

- Jan. 4-9, 2004
- March 21-26, 2004
- May 16-21, 2004
- Aug. 1-6, 2004

What participants have said about IRE and NICAR training:

"The workshop and the conference have convinced me that the investigative reporting approach and techniques can be easily applied to beat reporting and daily journalism."
– Afi-Odelia Scruggs, Professor of Journalism at Ohio Wesleyan University

"The training and workshops are first-rate, to be sure. ... Our newspaper, and hence our city, is the better for it in innumerable ways, both great and small."
– Willy Stern, *Nashville Scene*

"Thanks for your inspiration and guidance. I'm hooked on data!"
– Glenn Henderson, *The Palm Beach Post*

More information is available at www.ire.org/training

Children

continued from page 1

somehow shine light on children in Cleveland and allow us to compare the condition of their lives to that of kids elsewhere.

We spent the next few months searching for data. We wanted to touch on all the important areas: health, education, crime, housing. We wanted figures that could be compiled on a city-by-city basis. And for big cities, we wanted to be able to examine what was going on in individual neighborhoods.

Along the way, we found a couple of new wrinkles to this much-written story. (Much of our effort went into reporting on environmental health issues, such as identifying the 1.5 million youngsters statewide who either live next door to a toxic waste site or in poor housing). We also came up with some creative uses of Census 2000 data.

The result was a five-day series, "Children Left Behind," published in late July that showed how kids in Cleveland are so much worse off than those in other places in Ohio or in comparable big cities across the country.

We found that Cleveland students are twice as likely to drop out of high school than those in Philadelphia or Milwaukee. A larger share of our kids lives in poverty, are lead poisoned and become teenage parents.

We also found that nearly half of Cleveland's children are living in risky housing, meaning that they are at risk for injuries, lead poisoning, and potential asthma triggers like cockroaches and rats.

"We've got Third World numbers for kids," one community leader told us. And our data analysis showed he was right.

By geocoding the addresses in five years of Ohio death records obtained from the Ohio Health Department with ArcView 3.3, we identified some city neighborhoods where infants are dying in rates that rival Guatemala, Guyana and Romania.

This wasn't tricky, really. We used the age field in the death records to find babies less than 1 year old. Then we used ArcView's StreetMap geocoding extension to create a point map by matching the home addresses in the death records against a streets map. Later we used ArcView to attach a Census tract number to each death record, so that we could then compile infant deaths by neighborhood.

Using Ohio's birth records, which show the mother's age and baby's weight and gestational age, we were able to determine that premature and low-birth weight babies born to teenage mothers cost state taxpayers and hospitals an estimated $161 million from 1997 through 2001, the latest year for which we had records.

All told, we analyzed more than a million records...

And, not surprisingly, these crummy conditions hit poor and minority children much harder than others. For example, our analysis of five years of death records revealed that black children accounted for 25 percent of all child deaths, but only 14 percent of the child population. Among infants in Cuyahoga and Hamilton counties, Cleveland and Cincinnati, respectively, African-American babies comprised more than half of all deaths, yet less than one-quarter of all infants.

We found the death data easy to work with. We paid $100 for a year's worth of records, and the state provided them on CD in dBASE files. The five files held roughly a half-million death records. We used Visual FoxPro for the heavy number crunching.

Ultimately, we settled on eight indicators: the percent of children in poverty (Census 2000), the percent of children with a single parent (Census 2000), the rate of premature babies per 1,000 births (Ohio Health Department), the birth rate per 1,000 teenage girls (Ohio Health Department), the infant mortality rate (Ohio Health Department), the number of children sent to state detention centers (Ohio Department of Youth Services), the percent of children living next to a toxic waste site (Ohio Environmental Protection Agency), the percent of children living in risky housing (Census 2000), and the percent of children with a disability (Census 2000).

Aside from the birth and death records, all of the data was free. And the agencies seemed happy to give it to us.

All told, we analyzed more than a million records gathered from Ohio agencies, other states and the 2000 Census to produce our indicators and rank our cities.

To make comparisons between cities fair, we used Census data to group places into eight categories based on population, median household income and the percent of residents living in urban, rural or suburban setting.

Then we used a simple rank order approach to rank cities within each category. Here's how we did it: We ranked all of the cities in each category eight different times, once for each indicator. We summed the ranks for each city, divided by eight, and re-ranked the cities on the final number.

We gave each indicator equal weight. But we didn't include an indicator if it was based on fewer than 10 incidents. For example, we excluded infant mortality rates that were based on fewer than 10 deaths.

There were more than 1,600 places in our rankings.

The most controversial part of our analysis was our estimate of kids who lived in risky housing. (To be upfront with readers, we published the figures for all eight indicators for all 210 cities and towns in our circulation area).

We created a risky-housing index using four Census figures on the block level. They were the percent of children in poverty, the percent of homes that were built before 1970, the percent of residents who are renters, and a ratio of the value of homes in the block group compared to the average value in the city on the whole.

Each of the four parts of the equation was expressed as a percentage. We then averaged the percentages and applied this figure against the number of children living in the block group to get our estimate of those living in risky housing.

The idea was that low-income children in neighborhoods with older homes, a high number of renters and homes valued at substantially less than the rest of the city were likely to be living in risky housing. Our firsthand inspections in many neighborhoods appeared to back this up.

Still, that didn't slow the angry calls, especially from the older, so-called "inner ring" suburbs where we indicated that in some of them up to one-third of the children were put at risk by their homes.

Many callers felt this was a misuse of Census data, but I'll let you be the judge. You can find the stories, graphics and photographs at *www.cleveland.com/children.*

You might also want to check out the Annie E. Casey Foundation's ongoing "Kids Count" project (*www.aecf.org/kidscount*), which examines the well being of children in all 50 states and the District of Columbia.

Contact Dave Davis by e-mail at ddavis@plaind.com.

IRE and NICAR Services

Investigative Reporters and Editors, Inc. is a grassroots nonprofit organization dedicated to improving the quality of investigative reporting within the field of journalism. IRE was formed in 1975 with the intent of creating a networking tool and a forum in which journalists from across the country could raise questions and exchange ideas. IRE provides educational services to reporters, editors and others interested in investigative reporting and works to maintain high professional standards.

Programs and Services

IRE Resource Center: A rich reserve of print and broadcast stories, tipsheets and guides to help you start and complete the best work of your career. This unique library is the starting point of any piece you're working on. You can search through abstracts of more than 19,000 investigative reporting stories through our Web site.
Contact: Carolyn Edds, carolyn@ire.org, 573-882-3364

Database Library: Administered by IRE and the National Institute for Computer-Assisted Reporting. The library has copies of many government databases, and makes them available to news organizations at or below actual cost. Analysis services are available on these databases, as is help in deciphering records you obtain yourself.
Contact: Jeff Porter, jeff@ire.org, 573-882-1982

Campaign Finance Information Center: Administered by IRE and the National Institute for Computer-Assisted Reporting. It's dedicated to helping journalists uncover the campaign money trail. State campaign finance data is collected from across the nation, cleaned and made available to journalists. A search engine allows reporters to track political cash flow across several states in federal and state races.
Contact: Brant Houston, brant@ire.org, 573-882-2042

On-the-Road Training: As a top promoter of journalism education, IRE offers loads of training opportunities throughout the year. Possibilities range from national conferences and regional workshops to weeklong boot camps and on-site newsroom training. Costs are on a sliding scale and fellowships are available to many of the events.
Contact: Brant Houston, brant@nicar.org, 573-882-2042

Publications

The IRE Journal: Published six times a year. Contains journalist profiles, how-to stories, reviews, investigative ideas and backgrounding tips. *The Journal* also provides members with the latest news on upcoming events and training opportunities from IRE and NICAR.
Contact: Len Bruzzese, len@ire.org, 573-882-2042

Uplink: Bimonthly newsletter by IRE and NICAR on computer-assisted reporting. Often, *Uplink* stories are written after reporters have had particular success using data to investigate stories. The columns include valuable information on advanced database techniques as well as success stories written by newly trained CAR reporters.
Contact: David Herzog, dherzog@nicar.org, 573-882-2127

Reporter.org: A collection of Web-based resources for journalists, journalism educators and others. Discounted Web hosting and services such as mailing list management and site development are provided to other nonprofit journalism organizations.
Contact: Ted Peterson, ted@nicar.org, 573-884-7321

For information on:

Advertising: Pia Christensen, pia@ire.org, 573-884-2175

Membership and subscriptions: John Green, jgreen@ire.org, 573-882-2772

Conferences and Boot Camps: Ev Ruch-Graham, ev@ire.org, 573-882-8969

Listservs: Ted Peterson, ted@nicar.org, 573-884-7321

Mailing Address: IRE, 138 Neff Annex, Missouri School of Journalism, Columbia, MO 65211

Investigative Reporters and Editors, Inc.
138 Neff Annex
Missouri School of Journalism
Columbia, MO 65211

NON-PROFIT ORG.
U.S. POSTAGE
PAID
Jefferson City, MO.
Permit NO. 89

Ira Chinoy
Philip Merrill College of Journalism
1117 Journalism BLDG
University of Maryland
College Park MD 20742

Uplink

Uplink Info

A newsletter of the National Institute for Computer-Assisted Reporting

Editor
Brant Houston
brant@ire.org

Managing Editor
David Herzog
dherzog@nicar.org

Asst. Managing Editor
Jeff Porter
jeff@nicar.org

Art Director
Lisa Triefenbach

Copy Editor
Pia Christensen

Contributing Editors
Jaimi Dowdell
Holly Hacker

Director of Publications
Len Bruzzese
len@ire.org

Advertising Coordinator
Pia Christensen
pia@ire.org

Subscription Administrator
John Green
jgreen@ire.org

Subscriptions
IRE members $40, nonmembers $60

Uplink Address:
IRE-NICAR, 138 Neff Annex
Missouri School of Journalism
Columbia, MO 65211

NICAR is a joint program of Investigative Reporters and Editors Inc. and the Missouri School of Journalism.

NICAR services include supplying journalists with government databases, training programs, tipsheets and data analysis.

Editorial
573-884-7711

Subscriptions
573-882-2772

Postmaster: Please send address changes to IRE-NICAR.