Uplink

November • December 2004
Volume 16, Number 6

Published bimonthly by the National Institute for Computer-Assisted Reporting
www.nicar.org

ENERGY

Drilling for lease deals

By David Pace
The Associated Press

Earlier this year The Associated Press decided to investigate oil and gas leasing on federal lands, as the Bush administration pushed to open more government land to exploration. We wanted to identify the major players, track the leasing process to find out if the rules were being followed and determine if the administration's friends in the oil and gas industry were benefiting disproportionately from their leasing decisions.

We began with the U.S. Bureau of Land Management's LR2000 database, which the agency's field offices use to record and track all actions related to oil and gas leasing on federal lands. The database can be queried from the BLM's Web site (*www.blm.gov/lr2000*) but using this proved too limited and time-consuming for our purposes. We needed the entire database so we could structure our own queries.

The database was easy to obtain but difficult to decipher.

continued on page 22

ANIMAL HEALTH

Diving deep into marine park problems

By John Maines, *South Florida Sun-Sentinel*

Brandie the sea lion died after plunging into her 27-foot-deep swimming pool at the Aquarium of the Pacific in Long Beach, Calif. Sadly for Brandie, the pool had been drained for cleaning.

Fanny and Brigitte, bottlenose dolphins and star performers at the Montreal aquarium, starved to death after their regular trainers joined a picket line in a strike. The animals refused to take food from anyone but their trainers, and both died after 38 days.

The official cause of death for Taffy the sea lion, dead at age 21 at the Morro Bay Aquarium in California, was succinct but not very clinical: "Love sick."

In May, the *Sun-Sentinel* in Fort Lauderdale gave readers a detailed look into the world of marine mammals in captivity in the five-part series "Marine Attractions: Below the Surface," by senior writer Sally Kestin. Although the billion-dollar U.S. marine park industry thrills millions of visitors each year with

continued on page 20

SPOTLIGHT: ROADS AND RAILS

Data helps probe death on the tracks

By Steve Orr, *Democrat and Chronicle (Rochester, N.Y.)*

The *Democrat and Chronicle's* coverage of rail-crossing safety began earlier this year after a retired couple were struck and killed by a CSX freight train as they drove onto a crossing in a Rochester suburb.

Within a few days, it emerged that CSX personnel had purposely deactivated circuitry that controls the gates and warning lights at that crossing after repeated complaints that the gates had become stuck in the down position.

The circuit was off-line for a full week without repairs made. During that time, crews were given orders to stop their

continued on page 12

SPOTLIGHT:

For more about roads and rails see:

Bits & Bytes

Upcoming training

Looking for intensive computer-assisted reporting training in the coming year?

There are four weeklong Boot Camps in Columbia, Mo., for journalists who want to learn how to acquire electronic information, use spreadsheets and databases to analyze the information, and to translate that information into high impact stories. IRE and NICAR provide follow-up help after participants return to their news organizations. The 2005 Boot Camps are Jan. 9-14, March 20-25, May 15-20 and Aug. 7-12.

Journalists interested in learning how to map data for news stories can take advantage of mini-Boot Camps in using geographic information systems (GIS). The training sessions are Jan. 14-16 and Aug. 19-21 in Columbia.

News managers interested in learning to better integrate CAR into their newsrooms can attend the April 8-10 mini-Boot Camp in Columbia.

For complete information about IRE and NICAR training visit *www.ire.org/training/*.

Scrape away

IRE and NICAR are seeking contributors to the Scraping Project, which was formed to address the concerns of journalists who need to quickly process public data on the Web. An increasing number of public databases are available online, but they are often stored on multiple pages accessible only by lengthy search strings. This

continued on page 4

INSIDE NICAR

Changes for the Database Library

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

The new staff at the IRE and NICAR Database Library is settling in, so it's a good time to introduce our talented roster of data analysts.

Brian Hamman is a first-year master's student focusing on new media and investigative journalism. He spent the previous two years at Washington University in St. Louis working on digital archives projects.

Catherine Rentz Pernot is a first-year master's student specializing in news reporting. She's currently reporting for the *Columbia Missourian*. Before joining IRE and NICAR, she analyzed business and government finances in Houston.

Megan Clarke returns to the Database Library as a second-year master's student. She also reports for KOMU-Columbia and produces special projects for Missouri Public Radio and KMOX radio in St. Louis.

Byung C. (Peter) Lee is a visiting journalist and scholar from South Korea, where he is working on a doctorate in journalism. He also is an award-winning reporter for the *Busan Daily News* in Busan and is interested in covering crime, poverty and the environment.

Matt Wynn is a junior and worked at a weekly newspaper in Omaha, where he covered municipal government and spot news. Last summer he worked at the *Omaha World-Herald*, where he worked on special sections, spot news and investigative stories.

We've also had some departures from the Database Library.

Master's student Christina Caron left at the end of the summer to take the Kaplan Fellowship, working with ABC News in Washington, D.C.

Jaimi Dowdell recently received her master's degree and, after filling in as acting director of the Resource Center, is continuing her work for IRE. Beth Kopine has joined IRE as director of the Resource Center.

Dowdell and recent master's recipient Aaron Kessler shared a byline with Mark Morris of *The Kansas City Star* for an article in the newspaper about a Columbia-based Islamic charity whose assets have been frozen by the U.S. government. Dowdell used social network analysis software to visualize the government's allegations of how the charity helped support Osama bin Laden and other terrorists. Kessler is a former IRE Resource Center staff member.

Holly Hacker received her master's degree and now is the higher education reporter for *The Dallas Morning News*.

Andrea Lorenz, departs the Database Library in January for *The Kansas City Star*, where she will work on her master's degree professional project.

Tim Ragones also received a master's degree and is reporting for WSAZ-Charleston, W. Va.

It's great to work with such enthusiastic and bright journalists. We hope you can get to know the new staff as they assist you.

• • • • • • • • • •

Our Annual Computer-Assisted Reporting Conference is coming up soon (March 16-20 in Hollywood, Calif.) and we'd like your ideas for panels, panelists and demonstrations. So please send any ideas to confideas@ire.org. Remember to include your contact information.

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

CHILDREN

Reservation deaths uncovered with data

By Brent Walth, *The Oregonian*

In Oregon, there is a place where children die at more than three times the statewide average. A place where basic protections, such as seat belt use for kids, are ignored. A place where children die anonymously in a secretive welfare system.

A state deaths database helped us find it.

But it took more than a year of reporting by our team, including reporters Kim Christensen and Julie Sullivan and photographer Rob Finch, to understand why the deaths continued with no public outcry, and why public officials were not being held accountable.

The place is the Warm Springs Reservation, where the U.S. government forced three tribes to relocate in the 19th century. Today, about 3,800 people live on the 1,000 square mile reservation in the central Oregon desert. About half the residents are age 19 or under.

As with many reservations, poverty is a big problem, and Warm Springs officials say their community struggles with the loss of history and culture and with a continuing tide of fractured families, alcohol abuse and domestic violence.

We soon found, however, that many people in Warm Springs did not see their circumstances as hopeless. Instead, they shared stories of tragedy and loss to show many child deaths could have been prevented and that tribal leaders had too often turned a blind eye to the problem.

By operating successful businesses, the Warm Springs tribes have been more prosperous than tribes at other reservations.

Many people pointed us to reforms tribal leaders have declined to take, even though they have proven to save children's lives elsewhere.

What's more, we found the secrecy surrounding tribal government – including a censored newspaper – meant that those leaders had never been held accountable.

Here is what we found:

• Traffic accidents killed kids more than any other cause, but only four of 10 Warm Springs infants and toddlers are properly belted in. Tribal police sporadically enforced seat belt and child restraint laws and ranked the problem as a low priority.

• Five children had died violently after the tribal welfare system allowed them to remain with dangerous adults. The cases included two unsolved child killings that Warm Springs residents had not been told about until our series appeared.

• The tribes' adolescent alcohol and drug programs fail at twice the rate of other programs in Oregon, and tribal leaders have ignored counselors' pleas for improvements that have been proven to work on other reservations.

• The tribal council has cut programs that help children even as it has continued to spend $147,000 annually on a luxury skybox at Portland's Rose Garden sports arena.

The response to the series, "A Place Where Children Die," was swift. (You can read the series online at *www.oregonlive.com/special/warmsprings).*

Shaken by our findings, tribal leaders launched a "top to bottom" review of tribal programs to reduce child death rates and are examining every child death case, something they had not done before. They also ordered police to follow a zero-tolerance policy for anyone allowing a child to ride unsecured in a car.

Meanwhile, the U.S. Indian Health Service took the extraordinary step of forming its own watchdog team to make sure tribal child-welfare officials acted quickly on neglect and abuse complaints.

The story began in the summer of 2000, when 4-year-old Andres Saragos died of heat exhaustion after his tribal court-appointed guardian locked him in a car for nine hours under a pounding July sun. *The Oregonian*'s investigation at the time, written by Courtenay Thompson, found the Warm Springs child welfare authorities sometimes ignored reports of neglect and abuse.

Two years later, the newspaper decided to go back to see if tribal leaders' promises to improve child welfare had been fulfilled.

At the outset, I wanted to find out just how deadly a place Warm Springs was for kids.

The task seemed impossible. Tribal records are confidential and not subject to any state or federal disclosure laws. That meant the traditional reporting avenues, such as court files, police reports and budget records, were closed to us. We soon found this veil of secrecy isn't just for outsiders; the secrecy works against members of the Warm Springs tribes, too. Even parents who had lost kids to violent deaths had been kept in the dark by tribal police.

After a lot of phone calls, I discovered the state of Oregon kept a database of every death in the state going back more than 20 years. State officials had always refused the newspaper's requests for the data, citing confidentiality concerns.

I eventually won access to the data – about 700,000 records of Oregon deaths over a 20-year period – by negotiating a way to obscure the exact dates of death, allowing the decedents to remain anonymous. In turn we got key details, such as how old they were, how they died and where they lived, including county and ZIP code.

The ZIP code was key: The entire Warm

continued on page 19

Bits & Bytes

continued from page 2

project is an effort to help journalists share scripts that "scrape" data from Web sites for use in spreadsheet or database programs.

To begin the project, IRE and NICAR are collecting any scraping scripts, explanatory notes and databases, either simple or complex. The Scraping Project will make these scripts available for journalists to modify and use, provided they also share their modified scripts and documentation.

For more information about the scraping project see the October-November 2004 issue of *Uplink*. If you would like to submit a script, contact Jeff Porter at jeff@ire.org.

Extra! Extra!

For some inspiration, check out Extra! Extra!, an online guide to investigative and computer-assisted reporting stories. The Web page, at *www.ire.org/extraextra*, carries summaries of recent investigative pieces and links to the stories. Journalists can search for stories by category, from health and homeland security and politics to nonprofits.

With nearly 1,000 entries, it's more than likely that you will find stories related to your beat or whatever topic you're currently covering.

The story listing is just one more effort to help IRE members stay up to date on the latest in investigative stories and techniques. Readers are encouraged to submit stories and links by sending them to extraextra@ire.org.

FIRST VENTURE

Debunking some myths about speeders

By Holly Whisenhunt Stephen, *WOAI-San Antonio*

My first computer-assisted reporting story started shortly after I arrived at WOAI last year. The data involved more than 1.6 million municipal court records from Bexar County, Texas. The investigative producer I was replacing had already put the wheels in motion to get the data. It took some time to convince the court system what information was public. I'm told there was some grappling over obtaining the information, but it wasn't the biggest fight either. I picked up the data, which cost a couple hundred dollars, a few weeks after I started the new job.

During an IRE and NICAR Boot Camp, I learned to always make copies and never work with the original data. So, after loading the information, I put the disks in a safe place where they wouldn't get back into the mix. We received 10 years of data on a CD-ROM that had 12 comma-delimited text tables. I imported the data into Microsoft Access, making sure to bring all the fields in as text, and later changed the number fields in Access table design view.

When I imported the data, I selected the advanced options and created specifications to control the importing of the data. This was very helpful and saved lots of time when it came to importing the remaining tables, which had the same fields and same data types. The one point I cannot stress enough is to make sure the specifications are absolutely correct, or the resulting data table will be a mess, and you'll have to clean it up or repeat the import.

After importing each table, I checked to make sure all records were imported properly. Then I performed an append query to create tables with more than one year of data and joined all other tables. Before that I tallied the total number of records in the individual tables, so I could check whether the append query was successful. Each record represented a case in municipal court.

The municipal court data table contained 85 fields. The complaint description field was very helpful because it told us what the actual violation was. Other helpful fields included the violation date, time and location. The data also contained the officer's name and badge number, which made it easy to find out how many tickets an officer issued. We analyzed five years worth of data. The information contained several violations ranging from disorderly conduct to wasting water. We chose to focus on speeding, and ended up with more than 400,000 records that I merged into one table.

To separate the information, I queried the complaint description field. Because there are several different variations for speeding, I used the "*" wildcard in my Structured Query Language (SQL) WHERE statement to make sure all speeding violations were included in the query. After isolating the information, I performed a make-table query, so I would be working with a smaller table containing the data I needed.

This first project taught me the important steps of cleaning data, so it can be analyzed more easily. I performed several counts including how many tickets were written for speeding and speeding in a school zone. We wanted to find out where the most tickets were issued in the city. To accomplish this I needed the street address of the violation in one field, which was originally spread over four. I concatenated the contents of the four fields into a new one using SQL.

I must admit this is one of my favorite tricks because it seems like a huge task to accomplish, but with concatenating it's easy. Before doing this I made a new field called speeding1.fulladd to hold the new addresses.

Here is the SQL I used:

```
UPDATE speeding1 SET
speeding1.fulladd = trim
([speeding1].[new vioblk1] & " " &
[speeding1].[new vioblk2] & " " &
[speeding1].[vio street1] & " " &
[speeding1].[vio street2]);
```

After concatenating the addresses, I was able to get a better look at traffic-ticket patterns. I ran queries to group by address and count the number of times each appeared in the data. I chose not to change the names of all streets and highways. For instance, Highway 281 @ Bitters Road was also listed as Hwy 281 @ Bitters Rd. I decided that standardizing all of the addresses would be too time consuming.

Instead, I chose to use Access and an Excel spreadsheet to get an estimate of how many tickets were issued at specific locations. I queried the data in Access to show the streets and the count of accidents for each. I ordered the results in descending order so the streets with the highest number of accidents appeared at the top.

Then I noted the top streets and did another query to get the records for those streets. I put the records for each street into its own Excel worksheet and worked with the data there.

Although it was not scientific, we were able to estimate ticket locations and counts. We chose to publish the speeding sites where more than 1,000 tickets were issued in the San Antonio area.

Another thing we wanted to know was the average amount of time that passed from when a ticket was issued until it was resolved. I created a new field to hold the calculated times and ran an update query that inserted the number of days, as calculated using the violation and adjudication dates. Here is the SQL:

```
UPDATE speeding1 SET ADJTIME =
INT(([NEWADJ DATE]-[NEW VIO
DATE])/1)
WHERE [NEW VIO DATE] IS NOT NULL
OR [NEWADJ DATE] IS NOT NULL;
```

We were able to look at the data and determine some of the common "myths" linked with speeding tickets – for instance, that red vehicles are ticketed most often – weren't true in San Antonio.

Other fun facts we were able to find in this data:

• The officer who issued the most tickets and what time of day most tickets were issued. The prime time was 3 pm. A judge stated that's when people are picking up kids from school and maybe shift workers are leaving for home.

• A monthly breakdown when most tickets are issued. We found the dates with the most tickets were at the beginning of the month (the 3rd, 2nd, and 6th). The reason this was surprising is most drivers talk about and assume that officers strive to meet quotas at the end of the month and write more tickets then.

• The average age of drivers who get tickets: 31.

• The average speed over the posted limit that results in a ticket: 17 mph.

• The amount of tickets dismissed because officers did not show up for court or did not remember the traffic stop: 3 percent of all speeding tickets.

• The color vehicle received most tickets. We ran into some difficulties with this because not all records indicated the color of the vehicle. The Texas Department of Transportation does not record vehicle color on car titles because those colors can change. So we could only report on those who records included the data. White vehicles were the most-often ticketed.

Overall this was a great database to work on after Boot Camp. Many of the lessons I learned during the Boot Camp helped me to clean and analyze the data. The best advice was to be methodical and not overwhelmed by the volume of data. My advice to others is take your time, concentrate and understand what you are trying to accomplish before you start trudging through steps that might not be helpful.

As far as SQL goes, I now know it is impossible to learn everything you think you should know at Boot Camp. I started compiling a list of SQL statements and what they can accomplish. I find this to be very helpful when working on different data that has similar challenges. Another great resource is the NICAR-L listserv, where journalists doing CAR assist each other. Whenever I've had any questions (especially on deadline) I just post it online and usually receive several answers.

Contact Holly Whisenhunt Stephen by e-mail at HollyWhisenhunt@woai.com.

You could win $3,500.

Send in your entry for the Templeton Religion Story of the Year Award for the chance to take home $3,500.

Deadline to postmark entries: Feb. 1, 2005

For rules, entry forms and information on RNA's other reporting contests, click on the contest link at www.rna.org.

RNA Religion Newswriters Association
Helping journalists to achieve balance | insight | context in covering religion in the news
contests@rna.org | 614-891-9001, x2 | www.rna.org

MAPPING IT OUT

The latest uses of mapping in news reporting.

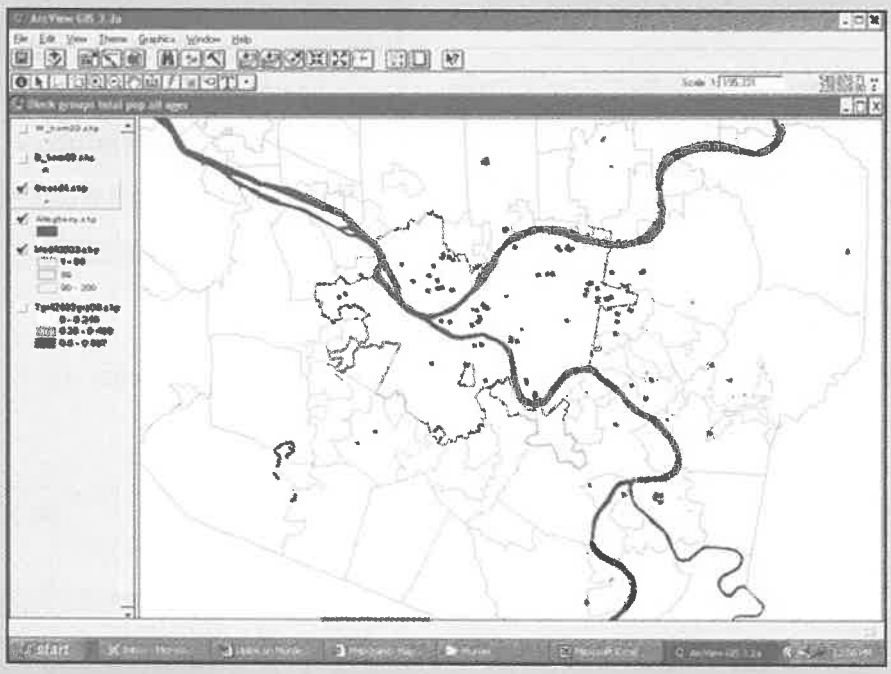
Unmasking murder myths

By Mark Houser
Tribune-Review (Pittsburgh, Pa.)

When Peguese, 26, was shot in his car in Pittsburgh three days after Christmas, his death helped set a record. By the time the coroner was done counting, Allegheny County had 125 homicides in 2003. Pittsburgh's murder rate of 21 per 100,000 residents last year was on par with Miami and Chicago. Suburbanites lamented they could no longer go downtown at night for a play or a concert without risking their lives.

I was assigned to do an investigative report on the rising murder rate, and I was skeptical about the conventional wisdom. From my shifts on the weekend rotation, I knew there was a good chance on any given Saturday that I would be chasing down a murder in the projects. Like the Peguese case, these usually involved a young black man shooting another young black man. Often, there were no witnesses. The incident would get a couple paragraphs the next day, and we moved on.

Although only 11 percent of Allegheny County residents are black, they accounted for three-quarters of homicide victims last year. Statistics also show that black victims were usually killed by blacks, while whites were killed by whites. Fewer than one in 10 homicides involved a white person killing a black person, or vice versa.



When I looked at homicides going back to 1990, I found another trend. Both in the early '90s and now, when the rate of killings rose, it was pushed by an upsurge of black victims. Police say this phenomenon tracks with the influx of first crack cocaine and now heroin. The casualty tally for whites hovered near 30 for each year through the decade, but for blacks it fluctuated from 36 in 1991 to 91 in 2003.

If we were going to run a big story about murder, then we needed to be clear about the problem. Downtown Pittsburgh isn't Disneyland, but neither is it a nest of armed muggers waiting in alleys to rob and kill their victims. The problem is localized in a few outlying neighborhoods most people don't venture into, and the law-abiding people who live there are highly unlikely to be victims. Residents of Homewood, the city neighborhood with the most shootings, repeatedly told me they felt safe there.

It concerned me that using these stark statistics could lead some to think I was playing up racist stereotypes. I needed to put the numbers in a way nobody could argue with, so readers could absorb this detail fully and move on to thinking about why it happens, what it means, what can be done, and why they should care. The best way to do that was with a map.

First, I supplemented a Microsoft Excel spreadsheet our crime reporter was keeping on each homicide with information from the coroner's log book. In the end, I had a database showing the address of where the body was found for every homicide in 2003, along with demographic information such as age, gender and race for each victim and for those arrested for the crimes.

I saved the Excel sheet in dBASE IV format, imported it into an ESRI

SPOTLIGHT: ROADS AND RAILS

NICAR has transportation data for journalists

By Jeff Porter, *IRE and NICAR*

The IRE and NICAR Database Library maintains several recently updated U.S. Department of Transportation databases relevant to the safety of U.S. roadways and the trucking industry. To order these and other databases, contact the Database Library at 573-884-7711 or 573-884-7332. Journalists have used these databases to identify dangerous roads and uncover problems with semitrailers on the highways.

The Department of Transportation's truck census contains records for each company that has commercial interstate vehicles weighing more than 10,000 pounds. This group includes buses, semitrailers and shippers of hazardous materials. Useful information includes the company's census number, name, address, number of vehicles owned or leased and the type of cargo it carries.

Although the DOT, citing national security concerns, has recently redacted much of the hazardous-material information, the IRE and NICAR Database Library offers an earlier version of the data that provided detailed hazardous-material information for carriers. See *www.nicar.org/data/trkcensus* for details.

The truck accidents database consists of records of accidents that involve commercial vehicles weighing more than 10,000 pounds. The vehicles included in this database are buses, semitrailers, moving trucks and rental trucks, among others. The data can be used as a starting point to see how many vehicles have been in accidents for different carriers. You can also link this data to the truck census database to find out details about the motor carrier. See *www.nicar.org/data/truck* for details.

The Hazmat database contains the incident reports of unintentional releases of hazardous materials for all modes of transportation (air, highway, rail and water). Journalists can analyze the data to learn more about chemical spills on highways, including which hazardous-material carriers have the most accidents and in which states. The chemicals most often released and accident details are also recorded. See *www.ire.org/datalibrary/databases/haz* for details.

The Fatality Analysis Reporting System database contains details about all fatal accidents, since 1975, on public U.S. roads. This data contains information about the location and conditions of each accident, the vehicles' make and model, the state that issued the driver's license and the status of any commercial license, prior speeding and DUI convictions, involvement in prior accidents, and more. See *www.nicar.org/data/fars* for more details.

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

ArcView 3.2 project and geocoded the crime scene addresses to create murder location points. The results showed the clustering of killings.

Because the crime addresses were often vague, without even a generalized street number, I had to plot about 50 of the 125 dots by hand. Then I used the Legend Editor to color the dots based on whether the victim was coded "B" for black or "W" for white in the database. (Unlike most U.S. cities, Pittsburgh has a minuscule Hispanic population and they didn't figure in the statistics.) Finally, I used Add Theme to pull in a Census 2000 map of Census block groups shaded by racial demographics to show the minority neighborhoods.

The final product was featured in "Murder in the Streets," a two-day series the newspaper published in July. (You can find it in the "Special Reports" section of our main page at *www.triblive.com*.)

The map made clearer than words could why the stories focused exclusively on black murders. Most of the voices I used were black, from the veteran homicide detective from the projects to the accused pusher worried about his safety on the streets. Or that of two teenage boys in a Pittsburgh ghetto and their uncles, one of whom sits in prison for killing the other in a gang shooting a decade ago.

The headline, "Black Death," was coupled with a moving photo of a mother in a funeral home caressing the chest of her dead son, a crack dealer who was shot off his bike. The map inside revealed the numbers, but the impact of the front page reinforced the stories' main purpose, which was to report the news we often overlook and remind readers that many victims are not innocents. Each was somebody's child.

Contact Mark Houser by e-mail at mhouser@tribweb.com.

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

Tracks

continued from page 1

trains before entering the crossing, dismount, make sure car traffic had stopped and then proceed. On the morning of the accident, the crew of a 110-car freight train failed to do so.

I was among several staff members asked to follow the story. As I did, I made frequent use of railroad safety data that is available online from the Federal Railroad Administration. While the data has the usual accuracy and timeliness limitations, it is as useful as any government cache I've seen for doing quick-and-dirty research.

It's much like the Federal Aviation Administration airplane accident and maintenance data that many journalists have used. The FRA online data can be tapped quickly to obtain background in the aftermath of a train derailment or highway crossing accident.

It also can be employed to craft follow-up stories or to take a more thorough look at railroad safety.

The data is available through the FRA Office of Safety Analysis Web site, *http://safetydata.fra.dot.gov*. In contrast to the agency's main Web site (*www.fra.dot.gov*), which is largely bereft of content useful to journalists, this site has nine separate sections rich with facts.

I don't profess to be an expert on this data, but here's some of what you can find: information on highway crossing accidents, as well as the crossings themselves; train accidents (derailments, collisions of trains, etc.); injuries and deaths of employees and non-employees; and FRA inspections of equipment and crossings.

In our case, we used the data to find other accidents on the rail line where the couple died, to look at the results of federal inspections of crossings in New York state and to pull out various nuggets of information about crossing accidents nationwide.

Search and download

Using the online query forms, you can pull off records of accidents or derailments in a few minutes' time. You can search by railroad, by state and county, by particular crossing and so on. Some sections have 30 years' worth of data, though you'll find in some cases that you have to search a month at a time, which is highly annoying.

There are literally dozens of different searches you can perform. The results from these on-the-fly queries come mainly in the form of text documents or pre-made text tables and charts. It's easy enough to paste such a table into Microsoft Excel or Access, but the text documents require a great deal of manipulation or data entry to get them into a database. You can get some of this information by downloading entire databases, (see below).

The summary reports of individual accidents that these queries produce are vague and at least occasionally inaccurate. The online summary of the Feb. 3 accident in suburban Rochester, for instance, continues to say that the "crossing was protected by gates, cantilever flashing lights, bells (audible)," but doesn't note that those devices failed to activate in a timely fashion.

You can obtain an Adobe Portable Document File of the crossing-accident reports filed by the railroad by using the "Query and Generate Crossing Accident Reports" link. Having the ID number of the crossing in hand will help. You'll find that these more detailed reports also can be inaccurate. This summer, for example, I wrote about a 2001 fatal crossing accident that occurred in the Rochester area. The railroad report I downloaded indicated that the crossing signals worked as intended. It turned out that wasn't true, as demonstrated rather vividly by a security-camera video of the accident scene around which we built our story.

The same information that is included in the PDF accident reports is available in a full database that can be downloaded from the FRA Safety data Web site. Other databases of train accidents, casualties and highway crossings also can be obtained.

The databases are available in six formats, including Access, Excel and comma-delimited text. The files come compressed or uncompressed.

I used the Access 2000 format, and the download was flawless. I chose the uncompressed version; it took about a minute to get the 1.5 MB database onto my machine, using a high-speed Internet connection.

That 1.5 MB file, though, represented

readme.txt

The IRE Resource Center offers copies of stories by journalists who have use computer-assisted reporting for transportation stories. Here are some examples:

A *Kansas City Star* investigation found that the problem of fatigued driving among truck drivers is pervasive. The Star analyzed trucking databases for the series. (Story No. 11812).

An investigation by *The* (Cedar Rapids, Iowa) *Gazette* found semi-trailer trucks accounted for nearly half of all fatal accidents on Interstate 80 in Iowa. The *Gazette* analyzed the Fatality Analysis Reporting System (FARS) for the story. (No. 21020).

To order these stories and others, contact the Resource Center at 573-882-3364 or rescntr@ire.org and provide the story number.

a year of crossing accidents nationwide. That's the primary problem – you can download only one year at a time. I spent about an hour downloading and merging the tables to assemble a 22-year collection of highway-crossing accident data for New York state. That compilation has about 1,400 records in it; five years of crossing accidents nationwide has about 16,000. Each record contains 100 fields of data about a given accident, though there is a fair amount of repetition.

I worked mostly with the crossing-accident database, and found it helpful for several stories. You can sort not only for railroad or location or time frame, but also for the reported circumstances of the accident: car hits train or train hits car; type of crossing equipment; gender of driver; reported speed of train; and so on.

I used the crossing-accident data in several ways. I separated all the crossing accidents on the 20-mile-long rail spur where the Feb. 3 accident occurred, which was useful for a story about the claim that the spur had too many crossings.

Finding problems

For a story on CSX's maintenance record, I used data about FRA inspections of freight railroads in New York state. I found CSX had been cited for crossing defects during inspections much more often than its competitors. The inspection data isn't in a downloadable database so I did some on-the-fly queries, and pasted the results into a spreadsheet. I also pulled data on crossing accidents in New York, and reported that CSX had a higher number of accidents than other freight railroads.

For a longer story about CSX's crossing maintenance practices and how they contributed to the Feb. 3 double-fatal crash, I used the database to find how many accidents nationwide had occurred at crossings protected by gates and lights, and in how many of those accidents there had been allegations that the crossing equipment malfunctioned.

For a two-part story on the 2001 fatal accident, I was able to pull data showing how many accidents occurred when motor vehicles drove into the side of trains, and how many occurred at night. Both factors played a role in the accident I was chronicling, which occurred when the gates and lights didn't deploy promptly and the young motorist plowed into the side of a locomotive, possibly without ever seeing it.

There are literally dozens of different searches you can perform.

Understanding data

Before I worked with the data, I had to understand it by looking over the database file structure document, available for download from the Web site's download section.

As I went through that file structure, I found myself a bit confused on several key points, and it wasn't until I also downloaded a second document that they made sense. This document is a blank copy of the agency's "Highway-Rail Grade Crossing Accident/Incident Report" form, which can be found at the very bottom of the main query page under FRA Forms.

The blank form contains some of the codes used in the database, which are not replicated on the file structure sheets. One example is the type of crossing equipment. Another – and this is the one I was really after – is whether that equipment was working as intended when the accident occurred.

So how do you find this in the database? I'm getting a bit arcane, but it took me days to puzzle out on my own, and it might be worth filing it away just in case you need it.

In the main crossing-accident database are two fields entitled "Signal" and "Sigwarnx." If the crossing in question has automated equipment such as gates and lights, then the "Signal" field should include a digit from 1 to 7. But what do those digits represent?

If you consult the *instruction page* that comes with the accident/incident reporting form, you'll discover that those digits represent the reporting railroad's statement about whether the gates and lights worked properly. "1" indicates everything worked well; anything else suggests an alleged or confirmed problem.

Go one layer deeper and you'll find that if the railroad reported "5," "6" or "7," which stand for "confirmed" signal problems, they also must fill out the "Sigwarnx" field. This field, which uses letters instead of numbers, gives 19 possible explanations for the failure of the crossing equipment. The explanations include things such as "devices down for repair," "vandalism," or a train going too fast for the circuitry to keep up. (Here's a story tip for someone: Look in the database and see how often "insulated rail vehicle" is the explanation for a crossing-signal failure. It turns out that some railroad maintenance vehicles are made in such a way that they won't activate gates and lights.)

Admittedly, crossing activation problems are rare – or, if you believe the rail critics I spoke with, railroads only *report* such problems rarely.

But if you're writing about a crossing signal failure as I did in two local cases, it is immensely helpful to be able to pull out other accidents that might have occurred under similar circumstances.

Contact Steve Orr by e-mail at sorr@democratandchronicle.com.

Tech tip...

Automate common tasks with FoxPro programs

By Jeff Porter, *IRE and NICAR*

Automation can be a good thing. You can turn a burdensome chore into a quick-and-easy routine. You can update information more often. You can set up a scheme to do a task consistently.

To that end, the IRE and NICAR Database Library has automated quite a few tasks over the past year. We've set automated weekly updates of Federal Election Commission campaign finance data, Federal Aviation Administration accident and service difficulty reports databases, and monthly updates of databases of aircraft, airmen and Internal Revenue Service nonprofits.

Such automated tasks, via programs written for Visual FoxPro, keep data fresh and make downloads for users a snap. Not a FoxPro user? You can program automated tasks in almost any software.

One automated process, which we run almost daily, benefits lots of users of our databases, especially those with smaller news organizations seeking a state slice of data. Some state slices are particularly advantageous. Occupational Safety and Health Administration inspection data, for example, is voluminous for the entire U.S. but records for one state would likely fit on one CD

In years past, state-slicing was a manual process. From a copy of the data on CD, we queried the data, copied the results into a folder, then copied the documentation. The bigger the database – for example, the truck inspection database consumes seven CDs – the more laborious the process. The end result was, for the most part, that copying the entire country would be easier than slicing out a particular state.

So we set out on a project to automate state slices, and those lessons learned can be applied to many types of tasks that can be automated.

First, we inventoried what we wanted to offer by state. Our order form (*www.ire.org/datalibrary/orderform*) lists several databases available via state slice. Then we examined the databases in question. Some natural inconsistencies occur:

First, not all the databases have the same field names for states, and some have multiple fields listing states. For example, the federal contracts database contains three fields that might include a state – the location of the government office, the address of the contracting company, and the state where work was to be performed. The states in those fields might all be different.

Second, government agencies don't treat the state identifiers the same. For contractors, it's a postal abbreviation. For federal assistance, it's a two-digit Federal Information Processing Standards (FIPS) code. For the FBI, it's yet another series of two-digit codes.

Then there are the multiple-table databases. In some instances (OSHA, for example) the state code is included in every table. Others (truck accidents, for example) do not.

The bottom line: When you get ready to do some automation, get very, very familiar with the data structure and documentation.

Here's how we handled the inconsistencies: We created a table of databases – including a one-character code for each database, the full name of the database (strictly for human reference; our state-slicer program relies on the code, not the name) and a memo field that contains the SQL query language needed to query the data for a particular state.

We created a simple SQL-written program that has three steps:

1) It poses a series of questions to the user – On which local drive do you want to save the data? Which database do you want to slice? Which state do you want?

2) It makes a copy of the table of databases, transfers the answers of the questions to the SQL language about that particular database, and copies the memo field material into a new program.

3) It then runs the just-created program, including the final step of copying the data, plus documentation, to the local drive the user chose during the first step.

To dissect each step: Since this is an internal tool at the moment, there's no fancy interface – a double-click on a desktop icon starts the process, and just four keystrokes later (that is, one for the drive designation, one for which database, two for the state), the state-slicing begins.

The first section sets the stage and offers some direction:

```
gcstatedir =
'\\compname\data\states\'
gcstoredir =
'\\compname\data\states\list\'

set directory to (gcstatedir)
```

Now, the series of questions begin:

```
? 'GET READY FOR A SERIES OF
QUESTIONS'
? ''
? 'YOU WILL BE ASKED:'
? '1. WHICH LOCAL DRIVE WILL
YOU SAVE THE DATA? TYPICALLY,
YOU WILL CHOOSE M OR N'
? '2. WHICH DATABASE ARE YOU
QUERYING?'
? '3. WHICH STATE ARE YOU
SLICING FOR?'
? ''
? ''
? 'YOU SHOULD HAVE IN FRONT OF
YOU THE STATE SLICE BINDER'
? ''
? ''
wait 'IF YOU ARE READY, HIT THE
SPACE BAR TO CONTINUE'

set alternate to (gcstatedir) +
'drive'
set alternate on

wait 'Which drive will choose
to save the data?'

close alternate
```

What's an "alternate?" Instead of just splashing questions and answers on the screen, the "alternate" command tells it to store that information into a file, in this case naming it "drive." The next few steps create a temporary table, append records from the just-created file named "drive" and parse it into records so that we can record the local drive's letter.

```
create cursor tempdrive;
(junk c(50),;
drive c(1))

append from drive type sdf for
not empty(junk)

replace all drive with
upper(substr(junk,at("?",junk)+1,50))

goto top

store alltrim(drive) to
curdrive
```

We repeat the process for the other questions/answers. The "store" command takes needed pieces of information – drive, database, state – and stores that information temporarily, only long enough to complete the state-slicing task. Those temporary, arbitrary, names are "curdrive," "curdbase" and "curstate."

For the sake of example, let's say we choose the "N" drive on our machine, the government database of firearms licensees (the code on that would be "F"), and the state of Alabama (the code in this database would be postal: "AL").

The next step is to use the table of databases, named "tables_stored" and make a copy of the table. The program names the copy as the word "table," plus the one-letter code of the database in question and the state code. So our copy of the table would be named "tableFAL." Then, opening up that copy, we start plugging in the stored pieces of information into the memo field content of the record that relates to the specific database.

```
replace all sql with
strtran(sql,'XX',(curstate))
for id = (curdbase)

replace all sql with
strtran(sql,'~',(curdrive)) for
id = (curdbase)
```

The field called "sql" is the memo field that contains the SQL program language for each individual database to slice. The function "strtran" is a find-and-replace tool with three arguments – the field, the character to find and the material to replace it with. For the firearms licensee data, here's some key SQL language:

```
gcstatedir =
'\\compname\data\states\'
gcpickupdir =
'\\compname\data\states\pickup\ATF_FFL\'
gcstate = 'XX'
gcyears = '04'
gcdrivedir = '~:\ATF_XX\'
```

Notice the "XX" and "~" that are serving as placeholders for the actual state code and the drive name where we want to store the data and documentation. The "gcpickupdir" folder already contains the documentation that goes along with the data.

The query for the data, then copying the results, uses the "gcstate" variable, with the "XX" value replaced by "AL" for our example of Alabama:

```
select *;
from 'Ffl'+(gcyears);
where premst = (gcstate) or
mailst = (gcstate);
nowait

copy to (gcpickupdir) + 'Ffl' +
(gcyears) type foxplus
```

```
The last step is to make some
copies of the data and documen-
tation to a folder named
"N:\ATF_AL."

command = 'xcopy /e /v ' +
(gcpickupdir) + '*.* ' +
(gcdrivedir)
run &command
```

This is, of course, a work in progress: The next step is to add another question and allow the user to automatically place it on our FTP server if someone needs the data right away. After that, we'll start adding our archival data in addition to our current copies of databases we provide year-by-year. All that can be accomplished in some not-too-difficult Visual FoxPro language.

We'd love to see how you've used your favorite applications to automate tasks; if you've got one you're willing to share, please e-mail me.

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

OPEN SOURCE

Building a no-cost, open newsroom I

By Aron Pilhofer, *The Center for Public Integrity*

Over the past two years, I've reviewed nearly a dozen free and open-source applications I have found useful in a newsroom setting. Starting with this issue of *Uplink*, I'm taking a different tack: I'm going to show you how to put this software to work in your newsroom.

This series is intended to enhance rather than replace or repeat existing documentation. I'll point out some of the more important things to think about, as well as how to avoid pitfalls. The goal is to provide the road map for a totally free, fully functional computer-assisted reporting platform to rival commercial products costing tens of thousands of dollars.

The heart of computer-assisted reporting is data, and the best way to handle large quantities of data is with a high-end database server. That job goes to the subject of this article: MySQL, which boasts an ideal combination of speed, power and features.

In future articles, I'll cover a number of other software packages, including the Apache Web server (the most popular server in the world) and PostNuke, a powerful but simple content management system that will drive our intranet and deliver data to our newsroom users.

I'll also discuss Subversion, a software package designed for programmers that I have found invaluable as a repository for important files, scripts and other documents. Subversion allows sharing of documents across a network. It tracks changes to the documents, so users are able to revert to a previously saved version at any time.

And finally I will talk about how all this software works together with the software most journalists are familiar with: Microsoft Office and SQL Server (yes, they play well together).

Basic needs

For the purposes of simplicity and brevity, I'm going to assume users have adequate hardware. In this case, a basic system would be a Pentium 4 desktop or server, 512 megabytes or more of memory and at least 40 gigabytes of hard disk space. One of the benefits of open-source software is that it tends to have fewer bells and whistles than other commercial software. Although I wouldn't recommend doing so, it would most likely run just fine on something with less power.

MySQL ... is fast, stable and free.

The server should have a tape drive or separate hard disks to use for backups. If you are going to put critical data on the server, you should have a system in place to ensure that it is safe.

In terms of operating systems, I am assuming the server will be running some flavor of Windows, preferably 2000 or XP. Although I have never personally tried to install these software packages on a version of Windows server, I believe they will work just fine.

In all cases, you will need local administrator privileges to install and configure this software, which is something you may need to negotiate with your IT staff.

I do not recommend running Windows 9x/ME. It is too unstable and some of the tools simply will not work. Linux is an excellent alternative to Windows, and will run faster on older hardware, but the configuration is a bit trickier.

Starting with MySQL

MySQL is the world's most popular open-source database server for a number of reasons: It's fast, stable and free. MySQL AB, the Swedish company developing their namesake product, has made remarkable strides in just the past five years building a software package that is second only to Oracle in speed and power, and it is closing the gap with SQL Server in terms of features.

Just in the past few months, MySQL AB has been rolling out some new tools to make installing and administering a MySQL Web server easier. But let's start by grabbing the files we need from the MySQL Web site (*http://dev.mysql.com*).

The server itself comes in several flavors. The 4.0.x line is considered the "production" version, although I recommend the 4.1.x (4.1.5 as of this writing) variant even though it is still technically in a testing stage. MySQL AB is extraordinarily conservative, meaning versions can be in testing for months or years before they are declared officially stable.

The 4.1.x line has been in development for nearly two years and is rock-solid stable. I have never seen it crash. Plus, it has several performance and feature advantages that make it the version of choice right now.

Version 5.0.x is in early testing right now and will add the last gaping feature hole in MySQL: stored procedures. I have experimented with it without trouble, but believe it is still too early to use in a production environment.

On the download page, you'll notice there are three versions of the MySQL server for Windows. The one to get is labeled "Windows Essentials." Also a relatively new innovation for MySQL, this version includes not only an installer but also a configuration wizard to make the process easier.

Download the installer and accept the default settings (you really can't make a mistake). The application will be installed into the following directory: C:\Program

Files\MySQL\MySQL Server 4.1\. When asked to select a server configuration, "typical" will suffice.

Initial configuration

Once the software is installed, you'll be prompted to launch the configuration wizard. Here, your selections do matter a bit more. At the initial screen, choose "detailed configuration" to begin setting up your new server settings.

The second page will ask whether the server will be used as a development machine, a multipurpose server or as a dedicated server just for MySQL. Unless you intend to run MySQL on your personal desktop computer, choose "Server machine." This will allow MySQL to use a portion, but not all, of your system resources when under heavy loads.

The third page is a bit more complicated, and delves into the various types of storage engines MySQL offers. In short, the server can store data in one of three primary formats, each of which has certain advantages for certain applications. Since we are optimizing our server for speed, select the first (default) setting, "multifunctional database."

The next page asks where to store data for one of the three types of tables used in MySQL (InnoDB). Because we will not use this type of table, accept the defaults.

Next, the wizard will prompt you to indicate the number of users you expect to hit your database server at any one time. In a small- to medium-sized newsroom, chances are you can be relatively conservative and select the lowest number you feel comfortable with by using the "manual setting" selection and pull down.

The next page asks how the server should be configured to communicate with users, and you're going to have to bear with me here. This can be confusing, but it's important.

By default, MySQL is set up to operate over TCP/IP, the standard addressing system for most networks – including the Internet. Under this scheme, each computer on a network – whether it is a server or desktop – is assigned a unique IP address that allows them to talk to one another.

You can see your machine's IP address on your network by going to the command line, and typing "ipconfig" at the DOS prompt. You should see a number that looks something like "192.168.1.34".

This address is critical to communicating over a network, especially between servers and clients. Without this address, for example, an e-mail server wouldn't have a clue where to send your e-mail. The same applies to MySQL. If you don't know what your MySQL server's IP address is, neither you nor your users will be able to log in over a network. (That's not entirely true because there is another protocol for finding the server called a "named pipe," but that's quite a bit more complicated.)

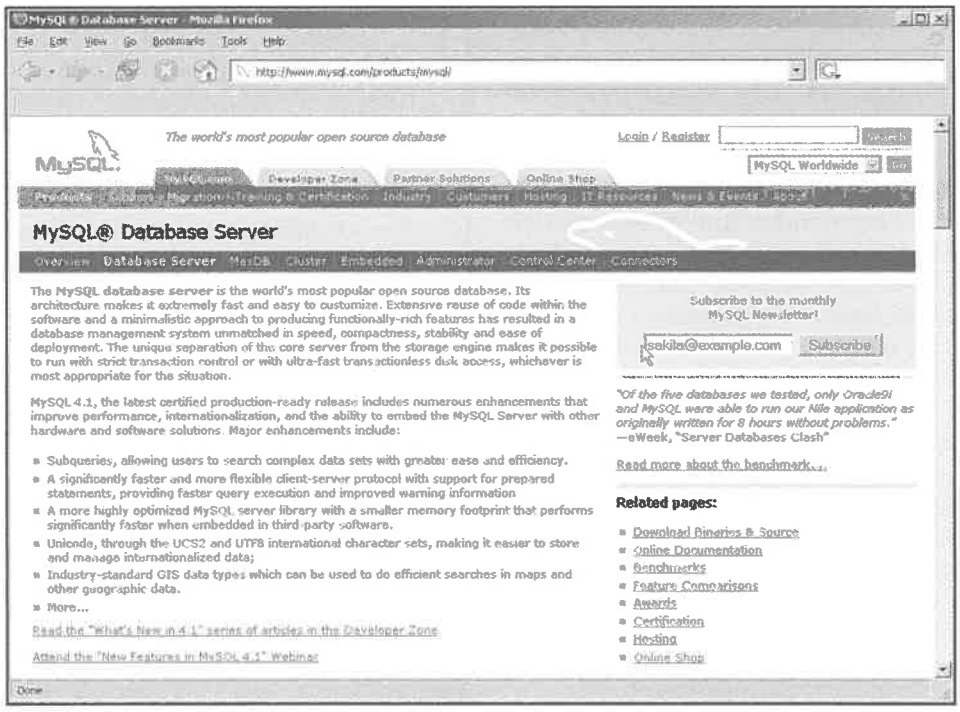
In addition to the IP address, it's also important to know which port your server software communicates over. Think of the port as a channel on your radio dial. If all radio stations broadcast on the same channel, the results would be a mess of confusing and conflicting signals. The same is true for a server, which couldn't keep all the communications straight if they came in over the same channel.

Thus, server software is configured to run on different default ports – 3306 in MySQL's case. To get back to the configuration wizard, this page is prompting users to indicate whether or not they wish MySQL to use TCP/IP (we do), and, if so, what port it should use. We can accept the default values in this case. (That may seem a bit anticlimactic after the above explanation. But it's a critical concept to understand, as you will see later).

On the next two pages, regarding character sets and installing MySQL as a service, select the default settings. Although if you intend to store text characters not found in English or most Western European languages, you might wish to choose the "best support for multilingualism" option on the character set page. On the second page, the default settings will install MySQL as a Windows service where it will run in the background automatically upon booting the computer.

The next page deals with security, specifically setting a password for the default administrator account (called "root"

continued on page 18



Building

continued from page 17

in MySQL parlance). This is critical, and should not be skipped. By default, MySQL ships with a blank password for the root user – which is a very bad thing. Root access should be granted only to the database administrator and highly trustworthy individuals who absolutely need it, because a root user has the power to alter or delete anything and everything – a fairly gaping security hole to say the least.

Choose a good, secure password that you won't forget, enter them in the textboxes and click next. There's no need to change anything else on this page.

On the final page, click the execute button and, if all goes well, the configuration will complete successfully. If you receive an error message when it tries to install MySQL as a service and you have had a previous version of MySQL installed before, you may have to back up to the configuration page and change the name of the service to something like "MySQL1." That has solved the problem for me.

For more information about the new MySQL installer and configuration wizard, see this article on the MySQL Web site: *http://dev.mysql.com/tech-resources/articles/4.1/installer.html*.

Final words

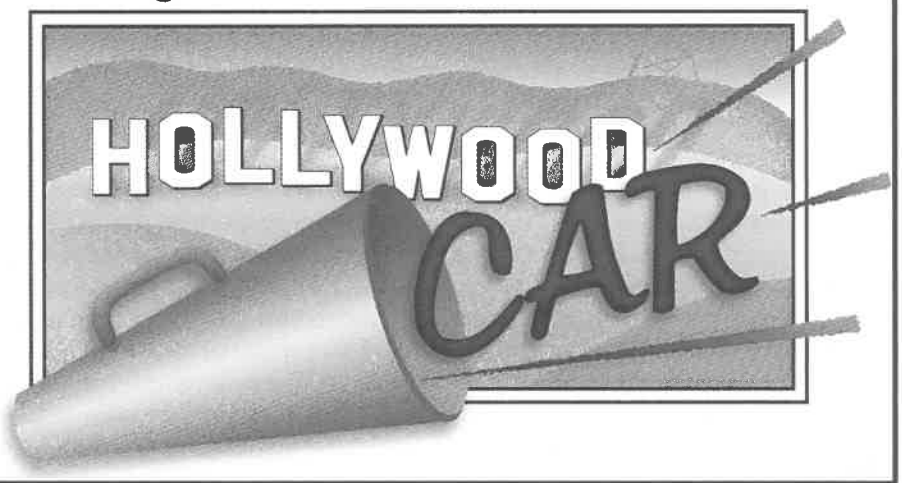
You now should have MySQL up and running on your server. To confirm, check the Processes tab in the Windows task manager and you should see either mysqld.exe or myswqld-nt.exe listed there.

In the next article, I will cover the basics of communicating with MySQL, (including how to connect via Microsoft Access/Open Database Connectivity), importing data and making that data available to your newsroom.

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

2005 Annual Computer-Assisted Reporting Conference

Hollywood, Calif. • March 17-20



HOTEL DEADLINE is Feb. 18

Don't miss out!

Panels are planned on every newsroom beat and topic in the news. Key topic areas include local government, business, crime, education, health, environment and more. Issues will range from getting the most out of census numbers to following up campaign finance data to understanding federal contracts and international data. Panels will also focus on homeland security, military and infrastructure issues.

PLUS!

Exhibitors, networking, software demonstrations and special sessions for beginners, educators, broadcasters and advanced users.

For CAR beginners, select panels will be coordinated with specific hands-on classes. The combination will give particiapnts a mini-boot camp experience with plenty of opportunity to apply what they learn.

Hands on classes will cover basic to advanced skills in spreadsheets, database managers, using the Internet, building your own database, mapping, statistics, social network analysis, SQL and more.

Host:
KNBC-Los Angeles

Conference Hotel:

Renaissance Hollywood Hotel • 1755 N. Highland Ave. • Hollywood, CA 90028

Hotel reservations: **Call 800-468-3571 by Friday, Feb. 18**, to get the discounted room rate of $164 plus tax. When making your reservation please ask for the IRE/2005 Annual CAR Conference room block.

Presented by Investigative Reporters and Editors, Inc. and the National Institute for Computer-Assisted Reporting

For more information, registration and the latest schedule, visit
www.ire.org/training/hollywood05/

Deaths

continued from page 3

Springs reservation has a unique ZIP. So we could calculate with precision, how many Warm Springs children had died and what had killed them.

But were the death rates high?

I used Microsoft Access and Microsoft Excel to calculate and compare mortality rates for Warm Springs kids ages 19 and under to those across Oregon. I did it the same way state and federal officials set mortality rates: by calculating the number of deaths per 100,000 residents.

(For exact methods on calculating mortality rates and making sure they are statistically sound, I used formulas published each year by the state of Oregon's Center for Health Statistics and Vital Records. You can find a straightforward explanation of how the formulas work in the appendix of the center's annual reports, located at *www.dhs.state.or.us/publichealth/chs/vol2.cfm.*)

To calculate a death rate, you need the number of deaths within a certain age group, and population for the same age group, for a specific period of time.

The database gave us the number of deaths.

I obtained population data from the U.S. Census Bureau's American FactFinder (*http://factfinder.census.gov*) Web site. Using Excel, I lined up the number of Oregon children's deaths by ZIP codes and counties with the populations.

When I ran the numbers, no place in Oregon had a child death rate that came close to that in Warm Springs.

We had to take special care, because the U.S. Census Bureau typically undercounts residents of Indian reservations. Warm Springs officials said their own counts of the reservation population showed that the U.S. Census usually missed between 10 percent and 15 percent of residents. To be safe, we increased the Census tally of Warm Springs by 25 percent – twice the estimated undercount.

This adjustment lowered the mortality rate for Warm Springs. But it didn't matter: Even with this conservative population estimate, we found that Warm Springs children die at a rate 3.4 times those in Oregon.

We could now call Warm Springs the deadliest place for kids in Oregon.

But my reporting had shown that, tragically, mortality rates for Native American children are higher than those for the United States as a whole. I worked with Indian Health Service statisticians for a year and found Warm Springs' child mortality rates were nearly double those for Native Americans nationwide.

I ran my calculations past state vital records officials and IHS statisticians to make sure they were statistically sound.

The numbers checked out.

To understand why child death rates were so high at Warm Springs, we catalogued 20 years' worth of the tribal newspaper, the Spilyay Tymoo, for key players, major events and deaths. Again, I used Excel to organize the information by name, age of death, cause of death and other key information we gleaned from obituaries.

We also collected documents wherever we could: the tribes' grant proposals to; state and federal agencies; state alcohol treatment reports on Warm Springs; and state police highway death investigations. We obtained the approval of parents to release children's medical records, which often included police and autopsy reports.

The most important reporting we did was spending time with the people of Warm Springs. We conducted more than 300 interviews with people in Warm Springs, including teachers, physicians, fire and rescue crews, police officers, and tribal judges and prosecutors. Rather than finding a closed society, we found people eager to talk about their lives and their struggles to keep kids safe. Many shared stories of their lives and tragedies faced by their families. Warm Springs residents allowed us to observe elk hunts, sacred ceremonies, school activities and funerals, and even take part in a traditional sweat lodge ritual.

We found there was no substitute for the time we devoted: The people at Warm Springs opened up when they saw we wanted to shed light on the solutions that had been ignored by tribal leaders.

Their stories – not the data – became the heart of our series.

Contact Brent Walth by e-mail at brentwalth@news.oregonian.com.

readme.txt

For more information using data to report on dangers and neglect to children see:

"Tracking quality of life," by Dave Davis of *The* (Cleveland) *Plain Dealer* in the September-October 2003 *Uplink*. Davis tells how journalists at the newspaper analyzed more than a million records from a number of sources, including the U.S. Census, and found that children in Cleveland were worse off than their counterparts in many comparable big cities.

Tipsheet No. 1822 by Jason Method of the *Asbury Park* (N.J.) *Press* lists resources for reporting on children's issues. The tipsheet contains details about using the Adoption and Foster Care Reporting System database.

To order the tipsheet contact the IRE Resource Center at 573-882-3364 or rescntr@ire.org.

Diving

continued from page 1

slick choreographed performances by dolphins, seals, sea lions, whales and other animals, our behind-the-curtain look showed a far less pretty picture of life in captivity.

Our primary research tool was the Marine Mammal Inventory Report (MMIR), a database kept since 1972 by the U.S. government's National Marine Fisheries Service. The MMIR was designed as a tracking and research tool to follow every major event in the life of a marine mammal in captivity – from its birth or capture, to its transfer between marine parks, its death or release into the ocean.

Among our findings: More than 3,850 marine mammals have died under human care, many of them young. Of nearly 3,000 whose ages could be determined, a quarter died within a year of birth, and half were dead by age 7. Of about 2,400 deaths in which a specific cause is listed, one in five marine mammals died of uniquely human hazards or seemingly avoidable causes including capture shock, stress during transit and poisoning.

The tip that led to the series came from a local animal rights activist, who contacted Kestin urging her to look into the death of an orphaned dolphin calf found stranded near Kennedy Space Center and named Rocketman by rescuers.

What intrigued Kestin in the federal records about the young dolphin was not what the activist saw – that the dolphin had died as a result of a bad decision by the federal government. Kestin instead focused on the reasons the government had ruled out several Florida marine attractions that wanted Rocketman.

One had a herpes outbreak among its dolphins. Another had a "history of losing calves, maybe due to a viral disease." Another had inexperienced staff and problems with veterinary care. The conditions described in those federal documents did not seem to match up with the happy image portrayed by marine attractions. Florida, the birthplace of the marine park industry, seemed the perfect place to launch an investigation of the industry's origins and history.

Antiquated database

The creators of the MMIR database some 32 years ago probably imagined they were creating a reporting system that by 2004 would be a wealth of information for animal biologists and zoo caretakers. Instead, we found a creaky old database that wasn't even really understood by the agency that keeps it. In fact, it had never even analyzed the information in the MMIR.

More than 3,850 marine mammals have died under human care.

The fisheries service headquarters in Silver Spring, Md., is based out of the headquarters of the National Oceanographic and Atmospheric Administration (NOAA). There, fisheries service officials at first refused to honor Kestin's Freedom of Information Act request seeking the database. They said they wouldn't give it to us because they didn't know how.

They could only give us printouts, which translated to more than 800 pages of legal-sized paper for the 9,678 records in the MMIR.

When the *Sun-Sentinel* threatened to sue, fisheries service officials continued to plea that they had no way of getting the database out of their computers. It was a DOS program, they said, and they barely knew how to use it beyond simple data entry and retrieval. They said they had no instruction manual, and the company that designed the database interface in 1995 was no longer in business. All this was hard to believe coming from an agency that's part of NOAA, the same people who run our weather satellites and use supercomputers to study global climate change.

We tried walking the fisheries service folks through the process of using different variations of the DOS copy-to-file command. They said they tried it, and it didn't work, and their bosses didn't want them spending time messing with it. We hammered at them for weeks, and it looked like we were headed for court.

Rosemary Armao, the investigations editor at the paper, came up with the solution – she proposed sending me to Silver Spring to look at the database and try to retrieve it. The cost was a bargain compared to hiring lawyers: $140 for a round-trip ticket on JetBlue, and $85 for a room at a run-down motel in Silver Spring. We also expected the fisheries service to balk at the idea of a stranger messing with their computers. But they liked the idea. When government officials say they can't make the computer work, they're not necessarily lying.

For the trip to Maryland, I took a laptop loaded with several hexadecimal editor and disassembler tools such as Hackman, expecting that I'd need to try to crack into PPIMS, the proprietary program that ran the MMIR.

But I never had to. It turned out that the PPIMS interface was linked to 104 dBASE files on a fisheries service server, and agency technicians knew where those files were located. The only challenge was to find which ones I really needed to re-create the MMIR reports, which include vital statistics such as name, age, species, birth or capture date, date and cause of death, and the current and previous locations of the animal.

I used a three-step process to find the tables I needed. Of the 104 possible

choices, many were junk files. These appeared to be remnants of long-forgotten research: queries that had been run years earlier but were still trapped in the system

Weeding out the tables I needed turned out to be simple. First I picked the largest ones. Then I selected the ones with names that sounded important. Finally, I asked the fisheries service people if they recognized any of the table names from their use of PPIMS. It only took three hours to identify the correct tables (datashet.dbf, inventory.dbf, species.dbf, and phf.dbf), import them into Microsoft Access and join them into a working copy of the MMIR.

Data gaps

For anyone who wants to take a shot at the MMIR, a warning: The tattered old database is filled with problems that make analysis difficult. There's missing data for many animals (the entries go back to the 1950s, apparently typed in from old paper records), but there are also many surprises. For example, of the 7,121 animals in the database, more than 900 vanish without any explanation. Usually, this happens in older records from the 1970s and 80s, and occurs when the animals are transferred out of a facility. They are never shown as arriving anywhere. They're just gone.

Another headache – and this applies to any research on zoo animals – is the issue of reporting how long animals live in captivity. Animal rights activists often look at only the average age at death of captive-born animals, which can be very low because of the high mortality rate in the first year of life. For example, the MMIR tells us that the average age at death for bottlenose dolphins that are born in captivity is only 2.5 years. But if the analysis is done a different way, looking at only animals that survive at least a year after birth, the average age at death climbs to 6.3 years.

The marine park industry cringes at either formula. They argue that average-age-at-death calculations are unfair because they do not include the age of animals that are now living. They prefer the use of "life tables," similar to actuarial tables used by insurance companies to predict human life expectancy. The American Zoo and Aquarium Association's life-table analysis provided to us for the series said that its own analysis of the MMIR shows that bottlenose dolphins now in captivity have a life expectancy of 20 years. Of the 875 dolphins that have died in captivity with a known birth date, only 17 percent have lived to age 20 and beyond. And 20 years is a long way from the 2.5 years and 6.3 years we found in the MMIR, and nothing close to the life expectancies reported by some marine parks at shows and on their Web sites – up to 30, 37 and 45 years. The association says it will take another 30 years of data collection before they have an accurate picture of how long bottlenose dolphins live in captivity.

Despite checking with experts around the country, the best we could do for our series was point out the different sets of numbers, which unfortunately didn't clarify the lifespan question.

The MMIR also suffers from underreporting, particularly in the past decade. In 1994 at the urging of the industry, the marine fisheries service was stripped of enforcement power over marine parks. The job was turned over to the U.S. Department of Agriculture, which keeps inspection records of facilities only on paper and in Adobe Portable Document Files, and only for three years before they're tossed.

In the years before the 1994 agency switch, the MMIR shows that 30 percent of marine mammals born in captivity died within a year. But after 1994, the rate drops to 15 percent. This remarkable improvement is not from advances in veterinary care. Marine parks simply stopped reporting stillborn deaths or the deaths of newborns to the government. To demonstrate this, we used LexisNexis to find newspaper accounts of the deaths of newborn animals, which are often publicized because the birth of a calf or pup is a big public relations event. Many deaths were not reported in MMIR. In a few hours, library researchers found 26 deaths in recent years that had never made it into the database.

And the MMIR might not have much of a future. In Washington, a bill passed by the House Committee on Resources in the fall of 2003 and an amendment introduced in April would cut back on the parks' reporting responsibilities. Opponents fear it would lead to the elimination of the MMIR, and could take away the public's ability to find out what happens to their sea stars.

The MMIR has data on every marine park in the country, so it's possible to do analysis on local parks. The contact person at the marine fisheries service is Jennifer Skidmore, at jennifer.skidmore@noaa.gov or 301-713-2289.

Contact John Maines by e-mail at jmaines@sun-sentinel.com.

GET THE RECOGNITION YOU DESERVE!

IRE AWARDS

Deadline for entries is Jan. 10, 2005

so be sure to enter your best work now.

For entry forms and more information, visit

www.ire.org/contest

Drilling

continued from page 1

While journalists have used LR2000 infrequently, oil and gas industry insiders rely on it for information that helps them prepare for oil and gas lease auctions. As a result, the BLM's Denver office, which administers LR2000, pulls a complete extract of the database each month and makes it available to the public for a nominal fee. We used two extracts in our analysis, one pulled from the database on Feb. 15, and the other on March 15. Each cost about $45 and covered BLM actions since 1982. Each extract consisted of 11 compressed files.

The extract totals more than 11 gigabytes and is organized by state, with each state file (the Eastern states are combined into one file) composed of a series of compressed tables. There are 11 case recordation tables in each state file, although most of the essential information can be found in four of them: the case table, listing the essential details about each lease; the action table, detailing all actions taken by BLM on each lease; the customer table, listing details of each lease owner; and the land table, providing details on the location of each lease. We wanted to look at leasing decisions on all federal lands, so we unzipped the tables and uploaded the data from each state in Microsoft SQL Server, a process that took the better part of a day. We used Enterprise Manager to manage the tables. The largest, which listed all BLM actions on various leases, had more than 9 million records.

The next task was cracking the BLM's complicated coding system. The Denver office provided a database schema, a spreadsheet with definitions of the hundreds of action codes the agency uses for oil and gas leases and a dictionary explaining how each of the data elements is used, both now and historically. The documentation helped, but it took many telephone calls to BLM database administrators before the data began to make sense.

Finding leaseholders

We set out initially to identify those individuals or companies with the most federal acres leased for oil and gas development. That proved to be quite a task, since the BLM database identifies unique leaseholders only within each state. And the agency makes no effort to associate legal subsidiaries with their corporate parent, even though the law limiting the number of federal acres that can be leased by a company in any state makes it clear that a corporation's holdings include those of its subsidiaries. To get around this problem, we created a table of the top leaseholders in each state and manually assigned codes to each that would enable us to calculate total holdings by the same company and its subsidiaries across the country.

Using Microsoft Access, we created a master table of companies by assigning the same code to the various names used in the state tables to identify the same company and its subsidiaries. This involved researching Securities and Exchange Commission filings to verify company subsidiary names. Then we joined that master table to each state table on the company name field and summed up the acres owned by each company using the assigned code field to group by. The same technique was used to calculate each company's total holdings across the country.

In state after state, we noticed that one address in Artesia, N.M., kept appearing among the top leaseholders. The one thing they all had in common was the name Yates. A little digging uncovered the fact that the Yates family of Artesia, through nearly three dozen companies, individuals and trusts operating out of the same building in Artesia, controls 2.7 million acres of oil and gas leases on public lands, far more than any other company or individual. Because leases owned by affiliated companies or individuals aren't counted together under BLM rules, unless the companies are legal subsidiaries, the Yates family has been able to accumulate far more leases than any of the big oil companies.

Our analysis found that the Yates family and a dozen large oil companies – including one formerly headed by Commerce Secretary Don Evans – now control a quarter of all federal lands leased for oil and gas development outside Alaska, despite the acreage cap law that was supposed to prevent such concentration.

Exploration data

After identifying the largest leaseholders, we began looking at how the leases were being used. LR2000 proved less useful here, because the BLM tracks most oil and gas exploration activity on federal lands in a separate database, the Automated Fluid Minerals Support System, or AFMSS. But LR2000 does include coding that enabled us to differentiate between leases that are producing oil and gas and those that aren't. We found that nearly three-fourths of the 40 million acres of federal lands leased for oil and gas development in the lower 48 states aren't producing any oil or gas.

Using the Freedom of Information Act, we obtained an extract from the AFMSS database that detailed drilling activities on federal leases for $41.

We had negotiated the format of the AFMSS data with BLM by including in the FOIA request a detailed listing of the types of information we were seeking. After the request was filed, we answered questions posed by AFMSS data managers who were writing the queries to respond to the FOIA request. Ultimately, we received four tables in delimited text format – one detailing information on all applications for drilling permits, one on closed inspections, one on wells started or drilled and one on well inspections. The data covered the same time frame as the LR2000 data, from 1982. The well inspections table was the largest, with 1.9 million records.

Using SQL Server to match that data against the LR2000 files of non-producing leases, we found that 98 percent had never had a single well drilled

and 97 percent had never had a single application filed for a permit to drill. We joined the tables on the lease identification number.

While industry officials complained that regulatory delays were the main cause for the lack of exploration, the data showed that it took only 61 days on average during 2000 for the BLM to make a final decision on each drilling permit application. The average decision time has declined steadily since the Bush administration came into office.

In opposing new federal oil and gas leases in sensitive areas, environmentalists had long argued that industry was not exploring many of the leases it already had. They claimed that these undeveloped leases, available for annual rents of $2 an acre or less, were being used more to boost oil company share prices than to increase oil and gas production. Our analysis was the first to use the BLM's own data to show that two-thirds of existing oil and gas leases have never been explored. Industry officials said large inventories of undeveloped leases are both normal and necessary to protect their investment in energy exploration. They blamed the lack of exploration on a shortage of drilling equipment and skilled personnel, regulatory delays and pipeline infrastructure deficiencies.

Both the LR2000 and AFMSS databases offer a wealth of information for further analysis. AFMSS, for example, tracks the inspection history of all oil and gas wells on federal lands and details the violations found and penalties paid. It also includes the decision timeline for key stages in the processing of drilling permit applications. LR2000 includes data on environmental reviews required for most federal oil and gas leases, rental payments on leases and appeals of BLM lease decisions, among other things.

Contact David Pace by e-mail at dpace@ap.org.

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 20,000 investigative reporting stories through our Web site.
Contact: Jaimi Dowdell, jaimi@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: David Donald, ddonald@ire.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@ire.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: Matthew Dickinson, mat@ire.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: Matthew Dickinson, mat@ire.org, 573-884-7321

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

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

Senior Editors
Sarah Cohen
Stephen K. Doig

Art Director
Lisa Triefenbach

Copy Editor
Pia Christensen

Contributing Editors
Megan Clarke
Andrea Lorenz
Brian M. Hamman
Catherine Rentz Pernot

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

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

Postmaster: Please send address changes to IRE-NICAR.

20742-0001

Ira Chinoy
Philip Merrill College of Journalism
1117 JOURNALISM BLDG
UNIVERSITY OF MARYLAND
COLLEGE PARK MD 20742-0001

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