November/December 2001

Uplink

SEPTEMBER 11

Airport Insecurity

By Jeff Porter
IRE/NICAR

Most *Uplink* readers are true believers, knowing that CAR is a necessity, not a simple luxury, for serious news organizations. That point certainly came home when newsrooms struggled with data after terrorism struck in the U.S. Sept. 11.

In the aftermath of hijackers seizing four airlines and causing thousands of deaths, reporters began digging into the database of Federal Aviation Administration enforcement actions to find stories about the security of their local airports. The database includes lapses and violations of every level – from the passengers who innocently carries a knife to more serious security violations by airlines and airports.

Do the data prove conclusively that airport security should have prevented the attacks? No. Do the data provide potential starting points for reporters covering the security status of American air travel? Absolutely.

The IRE and NICAR data library, which provides the data to journalists, was deluged with telephone calls and data orders. During the weeks after the attacks, 119 news organizations purchased the data. IRE set up delivery via its FTP server. The database was on the library's regular schedule to be updated in October, so IRE pushed to update it even sooner. During the week of the attacks, a series of text files was obtained from the FAA, downloaded from the agency's public FTP server, and turned into a series of dBase IV files so journalists could use the data in almost any database program.

Good timing

This turned out to be a good decision. After stories started appearing and reporters started asking questions, the FAA turned off the data spigot within a week after the attacks. The files could no longer be downloaded. It seems that the data was suddenly deemed "sensitive" by the agency that, since 1997, has provide copies of its Enforcement Information System database. At the time this was written, no one with the FAA would or could specify the concerns about the data, just the ominous message on its server: "The Enforcement Information System (EIS) is not available at this time due in part to security considerations. (14 CFR 191)" The federal regulation cited is fairly general, concerning "protection of sensitive security information."

Because of the FAA's refusal to answer questions, editors and reporters found interpreting the data daunting. Both the library and the NICAR-L e-mail discussion list agonized about trying to find answers about the complicated enforcement data. IRE members can look over that discussion at *http://www.ire.org/membership/listserv.html*. There are so many de-

Continued on page two

DATABASES

Dangerous drivers

By David S. Fallis
Washington Post

The initial idea was simple: Find out how drunken drivers fared in court.

The *Post* knew that Montgomery County, a District suburb of about 850,000 people, had a reputation among defense lawyers, police, prosecutors, traffic safety advocates and even drunken drivers as one of the most lenient in Maryland.

Ultimately, by layering paper records, electronic records and information gleaned in interviews, the *Post* was able to reveal a complete picture of the courts at work, finding that drunken drivers often have the upper hand.

Continued on page twelve

Inside Uplink

In addition to covering the Sept. 11 terrorist attacks, this issue of *Uplink* focuses on stories on transportation and the environment.

Learn more about a key database reporters used in the immediate aftermath of the terrorist attacks (see page one). Transportation stories include a look at tracking rail accidents (see page four) and rating airport convenience (see page six). On the environmental front, stories talk about radioactive contamination of workers (see page seven) and leaking underground fuel storage tanks (see page seventeen).

Census

John Perry of *The Daily Oklahoman* (Oklahoma City, Okla.) discusses how the paper used 2000 Census data to suggest that "school district boundaries may be a factor in residential segregation."

SEE PAGE EIGHT

Software, data

Aron Pilhofer, director of the IRE and NICAR Campaign Finance Information Center, lays out the pros and cons of open source software packages (see page sixteen). And training director Ron Nixon writes about the basics of building your own database (see page three).

Uplink

November/December 2001

Volume 13, Number 6
A newsletter of the National Institute for Computer-Assisted Reporting

EDITOR
Brant Houston

MANAGING EDITOR
Jeff Porter

ASSOCIATE EDITORS
Amy Sherrill
Mike Sherry

ART DIRECTOR
Lisa Triefenbach

NICAR is a joint effort of Investigative Reporters and Editors and the University of Missouri School of Journalism.

NICAR services include hands-on newsroom training in computer-assisted reporting, special academic and advanced training in data analysis.

DIRECTOR OF PUBLICATIONS
Len Bruzzese

SUBSCRIPTION ADMINISTRATOR
John Green

Uplink is published bimonthly by the National Institute for Computer-Assisted Reporting, 138 Neff Hall Annex Columbia, MO 65211. (573) 882-0684. Subscriptions are $40 for IRE members, $60 for nonmembers.

Postmaster: Please send address changes to NICAR. Send e-mail to jgreen@nicar.org

Continued from page one:

insecurity

cisions to make on how to use the data – from deciding which types of violations are important to which airports or airlines to compare. But for example purposes, let's follow the path of a fictional news organization and how it reached the decisions it reached.

First, a primer: The FAA enforcement database, dating back to 1962, includes four tables of information. The "main" table contains most of it, including the dates, often the violator's name and the location. A table called "security" lists lapses or violations, often with cryptic descriptions. Table "far4" cites the specific federal regulation violated, while a table called "actions" lists the outcome of cases, from warning letters to "proposed civil penalties" in dollars.

The FAA, of course, refused to explain the criteria of placing cases in the security table.

One technical problem – and the first decision to make – involves the duplicate records in the last three tables. The security table, for example, might list several violations involving the same case. Some news organizations wanted to simply count events, not necessarily adding up all the violations in the same case.

Using "distinct"

So, many CAR followers learned a technique using the DISTINCT function, an SQL statement that omits duplicated data. Since most of the questions came from Microsoft Access users, here's the Access query language that our fictional newsroom used to make a new table with one record per case identification number:

```
SELECT DISTINCT iid into cases
FROM SECURITY
```

The query simply created a new table with just id numbers, naming the new table "cases." An Access user could easily add a WHERE line to focus on certain security codes or descriptions, filtering for just some, not all, security cases.

But what to filter? Security matters are touched on in both the main and security tables. The main table includes a field called CAT_CODE, which specifies the type of violation, in broad terms, such as "security." The security table includes both descriptive and code fields that can be more specific about security violations. For example, security codes that begin with an "M" apparently deal with passengers or non-passengers carrying weapons. Security codes that begin with "C" often deal with failures of security personnel, often in connection with FAA inspections or surveillance.

Security problem

The problem is, though, the security table includes records that link to cases in the main table that aren't coded as "security" cases. Some hazmat cases, for example, were included in the table supposedly dealing with security lapses or violations. There's another CAT_CODE dealing with "cargo security" instead of just plain "security." The FAA, of course, refused to explain the criteria of placing cases in the security table. The NICAR-L e-mail discussion list debated which fields to rely on. Some suggested including all the security table records; others were comfortable using the CAT_CODE to filter for only "security" violations.

Beyond that very basic filter, though, newsrooms dealt with further filtering. Do they want to look at only violations when the airports or airlines are blamed? And if so, do they include only certain violations, such as failing to detect a firearm, or include a wider variety of violations? Should they factor in the passenger violations as well?

Again, editorial decisions had to be made.

Our made-up newsroom decided to look at only violations by airports or airlines, and filtering by the types of violations in the security table – beginning with the letter "C" in the security table's SEC_CODE field. So we re-created our "cases" table, filtering through SEC_CODE:

```
SELECT distinct iid into cases
FROM security
WHERE SEC_CODE like 'C*'
```

The work is partially done, though. Now, we could filter out all the cases in our main table with some additional filtering criteria:

Continued on page twenty-seven

Database tools

By Ron Nixon
IRE/NICAR

You've head the excuse from public information officers a thousand times. "We only have it on paper." But this time, they're telling the truth. All the documents you need are there and the agency is willing to give it to you. But what can you do with thousands of sheets of paper? How do you analyze it? You could give up and pull a few documents and do a largely anecdotal story.

Or you could do what hundreds of reporters have done: build your own databases. There have been numerous articles in *Uplink* showing reporters how to think about building a database – that is, think about what questions you want to ask the database, coming up with a structure and plan for inputting the data and establishing standards for data input. This article focuses on tools within database managers like Access and spreadsheets like Excel that will make data entry easier once you've come up with a plan for building the database.

Building blocks

• Forms—Entering data into a database can be a tedious exercise. And even the most loyal techno-junkie will get eyestrain trying to enter data into columns and rows. A form simplifies that process by allowing you to enter data into, well, a form. This removes lots of stress, especially when you're creating a database that has lots of fields.

• Input mask—These are useful functions in most databases and spreadsheets that allow you to specify exactly how you want data entered. Say you want all dates to be set up in the YYMMDD format. Using an input mask you can set your data field that way and no one can enter the dates any other way.

• Validation rules—Validation rules are similar to input masks. But unlike input masks, they allow you to limit the range of data entered into a field. Whereas the input mask will let you set the format of the date in the example above, validation rules allow you to limit the dates entered between two dates. For example you could only allow dates to be entered between 1990/01/01 and 2001/01/01.

• Look-Up wizard—The look-up wizard is a useful tool that can be used to cut down on entering repeated information like codes or city names. With the wizard, you can create a drop down menu and select the name or code you want. This cuts down on typing and mistakes from multiple data entries.

Entering data into a database can be a tedious exercise... A form simplifies that process.

• Auto number fields—Auto number fields are just that, a field that contains the number for each row. These are useful if you have data that does not have an obvious key field.

One more

• Indexes—If your database is going to include thousands of records, indexing the records might make sense. Indexes help speed up the searches and allow you to locate records quicker. Be warned however, that indexes themselves take up disk space and it may not be very beneficial to use indexes on smaller databases.

Building you own database can be a daunting task and even most CAR-jockeys would rather not, if they don't have too. But the tools listed above can make the task a lot easier. So the next time an agency says "We only have it on paper," don't despair; tools with database mangers and some spreadsheets can make the task a little less daunting.

Ron Nixon can be reached by e-mail at ron@ire.org

Tipsheets

More information on building a database is available from following tipsheets in the IRE Resource Center:

Tipsheet #806. Handout includes "Ten Commandments of Building Your Own Database" and advice from Jeffrey Meitrodt, Carolyn Tuft, and David Washburn. Also available on audiotape by calling (303) 649-1811 ($11).

Tipsheet #948. "Frugal and Proud of it...." Frank Krummer explains how to stretch a CAR budget by sharing data with organizations and building databases from Web sites or NICAR sources.

TRANSPORTATION

Tracking Rail Accidents

By Russ Clemings
Fresno Bee

Accident rates at railroad crossings are a mainstay of computer-assisted reporting, and for good reason — the data is freely available and easy to work with, and the subject is one of near-universal interest

But sometimes the most obvious story — the one that identifies the most dangerous crossings or lines — isn't the best story. After all, no matter how many safety measures are undertaken, there will always be a "most dangerous" crossing, and many times it will have earned that distinction simply because it's the busiest in terms of train or vehicle traffic or both.

Deaths at rail crossings, as it turned out, are relatively rare, at least when compared to ordinary traffic accidents, homicides and other violent causes.

Millions of dollars in federal, state and local funding has been spent in the past three decades to address the most serious crossing hazards, and by most accounts, progress has been made. Accidents and fatalities declined by two-thirds between 1976 and 1987, and since then have remained roughly level despite generally increased traffic.

Not every crossing that needs attention has received it, however, and those that were neglected became the basis for a *Fresno Bee* package on May 6, 2001.

Reports downloaded

First, the data: Since 1975, the Federal Railroad Administration has kept a database of rail crossing accidents. Railroads file reports, and an FRA contractor enters them into a computer, and the agency makes them available over the Internet at *http://safetydata.fra.dot.gov*. The same site also provides links to other FRA databases, including its national rail crossing inventory and its records on deaths and injuries among railway workers. The site allows users to query the data directly or download custom extracts in DBF (FoxPro), Access, Excel, Paradox and ACSII formats.

I downloaded 25 years of accident data for California in a couple of hours one afternoon last winter and easily loaded them into FoxPro. Then I started doing queries and found ... well, not a whole lot, actually.

Deaths at rail crossings, as it turned out, are relatively rare, at least when compared to ordinary traffic accidents, homicides and other violent causes. For the period from 1996 to 2000, the number of rail crossing fatalities in our six-county circulation area came to the not exactly eye-popping total of nine.

Beat the street

It was time to get out from behind the computer for a while and see if I could find a story that the data wasn't showing me.

Two major rail lines — the Union Pacific and the Burlington Northern Santa Fe — pass through most of the cities and towns of the San Joaquin Valley to link northern and southern California. I resolved to visit every one of their crossings in the heart of our region, Fresno and Madera counties. Maybe, I thought, an idea would strike me after I'd inspected a hundred or so crossings in person.

It didn't take quite that long. After about an hour of looking at crossings on the first of the lines, the BNSF, I was driving in an industrial area south of downtown Fresno. The street I was on paralleled the tracks; another ran parallel on the opposite side. Two other streets — California and Hamilton avenues — crossed both streets and the tracks within a few blocks of each other.

Unlike every other crossing I had seen until then, these two did not have crossing gates, only lights. Vehicles ranging

Continued on page five

Link

The Fresno Bee Web site has a map of rail accidents: *http://www.fresnobee.com/man/news/railroad.html*

Tipsheets:

More information on covering railroad crossing safety is available from the following tipsheet in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #1363. Clemings suggests sources on railroad crossing safety.

Tipsheet #1427. Madeleine Doubek discusses how to do local investigations with limited resources

To order call the Resource Center at 573-882-3364

Continued from page four: **accidents**

from tamale carts to tractor-trailers skittered over the tracks.

Suddenly the lights came on, a bell sounded, and a dark blue double-decked Amtrak California train lumbered into view from the north, blasting its horn. Some of the vehicles ignored the lights until the train was practically upon them.

I found two more such crossings in a poor residential neighborhood just north of downtown, and another on the UP line in Selma, a town of 20,000 about 15 miles southeast of Fresno. Later, when I got back to the office and looked in the FRA data, I saw that there had been accidents at each of those crossings. The California Avenue crossing had been the site of eight accidents in 15 years. None was fatal, probably owing to the fact that trains move slowly through that congested area.

Fatals galore

The Selma crossings on North Street was another story. In one six-month period, there had been three accidents, one fatal, and five years earlier there was a double fatal.

The story that developed from these discoveries focused on the failure of government agencies to address these hazards.

When I interviewed Selma officials, they admitted that they had never applied for money from the federally funded account (called "Section 130") that pays for crossing gates. But when I interviewed the state official in charged of distributing those funds, he told me that the city should have applied. With the poor safety record of that crossing, the application would be likely to get immediate approval and funding.

A county official suggested that I also look at a separate state-funded account that pays for grade separations — bridges and tunnels — to eliminate dangerous and congested crossings.

Anemic fund

That fund is notoriously anemic. The waiting list for appropriations is measured in years. But when I started reading the fund's annual reports, published by the state Public Utilities Commission and posted on its Web site, I came across a curious reference to two Fresno projects that had been stalled. Funds for them, the report said, would be taken back and spent elsewhere.

City officials somewhat sheepishly admitted in subsequent interviews that, yes, they had received authorization from the fund for bridges at two of the city's most-traveled rail crossings, but had to forfeit the money because they couldn't come up with matching funds.

The city's budget documents showed that millions of dollars in discretionary transportation funds were passing through its treasury each year.

The city's budget documents showed that millions of dollars in discretionary transportation funds were passing through its treasury each year. But instead of setting it aside for the rail projects, the city had been spending much of it on things like new curbs and sidewalks — even tree-trimming — in a handful of residential neighborhoods.

It may not have been the story that I set out to do. But for the families of victims like Ray Sartini Jr., who died when his car was struck by a train at one of those two Fresno crossings, or Jose Patino, the victim in the most recent Selma fatality, the fact that their accidents were relative rarities made their tragedies no less compelling.

Russ Clemings can be reached by e-mail at Clemings@cris.com

Stories

Other stories on railroads are available from the IRE Resource Center (*www.ire.org/resourcecenter*):

Story #16004.
***The Morning Call* (Allentown, Pa.) examines railroad crossings in the Lehigh Valley and finds that less than half of the 750 crossings found are marked with signs. From 1989 to 1998, there were 76 accidents and 23 deaths at these unmarked crossings.**

Story #14184.
***The Washington Monthly* finds that serious accidents involving the transport of hazardous materials on trucks and trains have become an almost daily occurrence. Hazardous material transporters are dangerously overworked and often not trained properly.**

To order call the Resource Center at 573-882-3364

The October 2000 *Consumer Reports* story, "Rating America's Airports," is available from the IRE Resource Center (Story #17254).

Tipsheets

More information related to transportation and magnetic tapes is available from the following tipsheets in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #448. Rose Ciotta gives pointers for using U.S. Department of Transportation sources, with a list of databases and story ideas.

Tipsheet #136. Brant Houston covers common questions on using data from nine-track tapes, from how to get the proper documentation to how to make it database compatible.

To order, call the Resource Center at (573) 882-3364.

TRANSPORTATION

Airport performance

By Jeff Blyskal
Consumer Reports

Airline ratings articles are a dime a dozen, in large part because of the easy availability of government airline performance databases and regular customer satisfaction surveys spoon-fed by J.D. Power & Associates.

Not so for airports.

Most airports are run by municipal authorities not geared for competition. Consequently, there are no airport performance compendiums a reporter can use to bang out a quick and comprehensive ratings. So when *Consumer Reports* set out on "Rating America's airports" (October, 2000), we knew we'd have to build the underlying database from the ground up.

Finding trends

Why rate airports, anyway? Preliminary reporting revealed that big airports dominated by major full-fare carriers were beginning to get their first competition from smaller secondary airfields, which had become metro market beachheads for low-fare Southwest Airlines. Nor was this phenomenon isolated; it was a trend taking shape in metropolitan areas from Boston to L.A., Miami to Chicago, New York to San Francisco.

Air travelers resigned to suffering long lines, long gate treks, and long delays at their same old airport may be surprised to learn that on some routes they can choose from a dozen different airports, if they take into account all their options at the departure, destination, and connecting cities. The choice, our ratings ultimately found, could make the difference between a tolerable, hellish or even a pleasant travel experience.

The benefits of building your own database are many: You get custom results that serve your specific goals. The drudgework of digging up all those numbers gives you greater depth of understanding and insight into your subject. Your original research is not dependent on available studies. Finally, you get exclusive data that your competitors will be hard-pressed to replicate.

Here are 10 tips for do-it-yourselfers based on our four-month airport project:

1. Define your standard of excellence. When *Consumer Reports* develops ratings, we aim to get to the heart of what consumers should be getting for their money. Airports are not a desired end per se, like a resort hotel; they're only a means to an end. But let's not be shy. Airports are an obstacle course between ground transport and what you're really there for: High-speed air travel.

Knowing this, we took our bold empirical stand: A good airport is one that gets travelers up and down from the sky as quickly, safely, and inexpensively as possible, with minimum hassle.

The benefits of building your own database are many: You get custom results that serve your specific goals.

2. Define your measures. How do you measure how well an airport meets this standard of quality? A Nexis search turned up few ideas or tools, and Web research of airport and airline sites revealed spotty, incomplete information.

So we used some common sense and our own good and bad experiences getting through airports for guidance. What determines how quickly you get on your way? Flight delays are one good measure. What constitutes hassle? How about a panicked search for an empty parking space as departure nears; rental car shuttles creeping through traffic-choked terminal roads; assignment to a Mad magazinesque Gate A1– Yea! Short walk! – that's really at the end of a three-mile-long concourse.

The airports

3. Define the universe. Which airports? Drawing on the Annual Worldwide Airport Traffic Report published by the Airports Council International-North America trade group *(www.aci-na.org)* we zeroed in on the largest U.S. airports in the nation's most populous metro areas. We also included smaller alternative airports with jet service

Continued on page twenty-four

Radioactive contamination

By Pete Eisler
USA Today

The "Poisoned workers & poisoned places" series wasn't planned as a CAR project. With no databases to draw from, this story was supposed to be pure shoe-leather, an old-fashioned document dig through dusty archives spread across the country.

It took all of a month before we turned to the computer.

The series revealed that the U.S. government secretly hired more than 300 private companies during the 1940s and '50s to process thousands of tons of radioactive and toxic material for use in nuclear weapons. From small machine shops to big chemical plants, commercial manufacturing facilities across the nation were quietly converted to the risky job of handling uranium, thorium, polonium, beryllium and other dangerous material.

The role that CAR played in this story illustrates the assistance computers can provide on a big project, even when there's no real database analysis involved.

The newspaper showed that the government regularly documented severe health risks for employees at the contracting sites, but those studies were kept secret, leaving thousands of workers unaware of their increased odds for cancer and other illnesses. The contracting operations also contaminated the air, soil and water in dozens of communities, pumping out radioactive and toxic waste with little regard for the safety of neighboring residents.

The role that CAR played in this story illustrates the assistance computers can provide on a big project, even when there's no real database analysis involved. We used the computer less as a reporting tool than an organizational aid. But it nevertheless proved invaluable.

Virtually all of the documentation for the series came from a review of more than 100,000 pages of declassified federal records on the nuclear weapons program. They were stored in a half-dozen or so libraries and federal archives across the country, from Atlanta to Las Vegas, Chicago to Washington. Most had never been reviewed since their release by the government under a declassification program that has dumped millions of pages of files from the weapons program into the public domain over the past decade. All of them were paper records, many of them so faded that they were barely legible. There was no index. Only a tiny fraction had been organized into a few searchable databases.

Track and retrieve

Our goal was to go through all of these records, identify every company that secretly worked for the U.S. nuclear weapons program, document the type of work each firm did, and assess its impact on workers' health and the environment. With so many records to search, organizing our material in a way that would allow us to track and retrieve specific information and findings became critical. Some of the documents we found related to a single company; some included information that was pertinent to many of the contractors we were studying. Others didn't mention any companies at all, but contained general data on the health and environmental risks associated with particular operations that were being done at some of the facilities we were tracking.

Using Excel

So this was far more complicated than simply assigning a manila folder to each company and filling it up with whatever pertinent documents we happened to run across

We decided to build an Excel spreadsheet that we could use to track the records and relate them to the applicable contractors. We set up a field for each company we identified, with additional fields for the firm's

Continued on page twenty-five

The *USA Today* story, "Poisoned Workers & Poisoned Places," is available from the IRE Resource Center (Story #17197).

Stories

Other stories on contamination are available from the IRE Resource Center (*www.ire.org/resourcecenter*):

Story #16763. "Rocky Flats: From Cold War to Hot Property," *Westword* (Denver) looks back at the site of a 1950s atomic weapons plant. Decades of careless cleanup resulted in $18.5 million in fines.

Story #14236. "Toxic Burn/Toxic Burden," *The Nashville Tennessean* investigates a pattern of unexplained illnesses centered around East Tennessee's Oak Ridge Reservation, a longtime production site for nuclear fuel and components.

To order, call the Resource Center at (573) 882-3364.

The Daily Oklahoman (Oklahoma City) story on racial segregation is available from the IRE Resource Center (Story #18198).

Tipsheets

More information on segregation in housing is available from the following tipsheets in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #128. Frank Clifford and Anne C. Roark of *The Los Angeles Times* offer advice on how to use census data to investigate minority issues.

Stories

Other stories on segregation are available from the IRE Resource Center:

Story #17279. *The Columbus (Ohio) Dispatch* investigates the "uneven educational opportunities in the Columbus public schools."

CENSUS

Deciding boundaries

By John Perry
The Daily Oklahoman

I'll admit up front that our approach to the racial segregation story probably had as much to do with ego as with anything else.

With the release of the first detailed race data from the 2000 census in March, we knew just about every newspaper would be writing stories about what the data said about racial segregation. Also, in 1995 we had written a series based on 1980 and 1990 census data showing that Oklahoma City was slowly becoming more racially integrated, but more segregated by income class.

The map showed an island of white dots in the center, each representing 50 more minority kids, surrounded by a sea of darker dots, each representing 50 more white kids.

Most of our readers won't see the census stories in other newspapers, and probably no more than a handful have even vague memories of what we wrote six years ago. But still, we wanted to do something at least a little different from what we expected to see in other newspapers and from what we had already done.

No change in data

When we looked at the redistricting data released in March, we found that racial segregation hadn't changed since 1990 anyway. The dissimilarity index for blacks and whites was 58.9 in 1990 and 58.2 in 2000. For American Indians, Oklahoma's largest minority group, the index went from 35.9 in 1990 to 34.2 in 2000.

We also rationalized that there is limited benefit to readers from just describing segregation's existence. We wanted to look beyond that to a possible cause. And the No. 1 suspect was the Oklahoma City school district.

After a federal judge ordered busing to desegregate Oklahoma City schools in the 1970s, the district's attendance dropped from more than 70,000 to 35,000 in just a few years. The district's students are now mostly minority and poor, even despite a flurry of inner-city redevelopment in the last decade.

This suggested that families with children were deciding where to live based on school district boundaries. If that were true, we expected that families with children would more likely reside in racially homogeneous neighborhoods. That should make children more segregated than the general population. The data tended to support this theory.

Calculating block groups

The actual data crunching for the story wasn't that unusual. For our earlier series on segregation, we had used tract-level data to calculate the dissimilarity index. But when we visited inner-city neighborhoods, we found that a single tract can include separate and unique neighborhoods. So we decided to calculate segregation indexes based on block groups this time around.

When we calculated the dissimilarity index — most often described as the percent of a population that would have to move to evenly distribute them throughout an area — we found a small difference between segregation among children and among the total population. But the difference went in the predicted direction every time.

We also used ArcView 3.2 and the free add-on, Two Theme Analyst, to estimate from block-level data the population changes by school district The under-18 population inside the Oklahoma City district boundaries grew by 2,610. But that resulted from the loss of 8,309 white (non-Hispanic white alone, in census speak) kids and a gain of 10,919 minority kids. Near-

Continued on page ten

Pricing homes

By Steve Orr
Democrat and Chronicle

The data we work with usually are already compiled when we get it. With luck, everything is in nice, neat columns, and is seemingly perfect. We check carefully for outliers and other anomalies, but if everything seems to add up, we'll accept it as accurate and go from there.

It's a whole different story when you build a database from the ground up, by accumulating and keypunching your own data. On the one hand, it's reassuring to know you have total control over quality. On the other hand, though, it's disconcerting to realize that one misread digit or misdirected keystroke can turn your data into garbage.

We can track individual buyers or sellers, specific lenders, particular subdivisions – pretty much everything we need to.

The database in question was created to help with our ongoing coverage of an apparent mortgage-fraud ring led by Robert A. Amico and several family members, all of them homebuilders. New suburban homes were being sold and mortgaged for prices that seemed far too high, and many of them were quickly falling into foreclosure.

Jumbled piles

After initial investigation yielded huge, jumbled piles of deeds and mortgages, it became clear we had to systematically track every facet of the process – land acquisition, construction financing, initial sale and subsequent resales, purchase-price mortgages, second mortgages, foreclosures and so on.

First we created an Excel spreadsheet, which was fine for basic information, but we needed a database to look for relationships and play with the numbers, so we made one.

The hardest part, as noted above, was acquiring the data. We knew the names of the builders, their companies and their associates, so we were able to identify properties they had built or purchased. Some information was obtained online from the unusually robust Web site of the local county clerk *(www.clerk.co.monroe.ny.us)*, but because of the way the Web operation stores its records, we could not use it for most of our work. That later was accomplished the old-fashioned way – by standing at the counter in the clerk's office, wading through records on public-access computers and on paper, writing down the key information and later typing it into the database. Among the data we compiled were record-accession numbers used in the clerk's office; once we had these in hand, we could call up most of the records online from our office and print out copies of deeds and mortgages. (We found this feature particularly helpful as our lawyers asked for documentation as they vetted stories.)

Not recommended

Finding and comprehending all the data was not easy, and we would not recommend this approach to any reporter who does not possess a thorough understanding of how local property records are prepared and stored. Even though our team included someone with that knowledge, we found our data acquisition was a nightmare of shifting addresses and tax parcel numbers, of sales back and forth between interlocking companies and family members, of building loans for houses that were never built, and so on. Even though it became easier to find data after we had the first few subdivisions under our belt, the process took several months to complete. Truth be told, a few of those errors of transcription or entry were made and have subsequently been found, and there are a few data gaps that no degree of digging has been able to fill.

The Access database itself was relatively straightforward in its construction. For ease of use, it was broken into three tables, and

Continued on page ten

The Democrat and Chronicle story, "Mortgage Fraud in Suburbia," is availabe from the IRE Resource Center (Story #17289).

Data

NICAR offers the Home Mortgage Disclosure Act database. It holds demographic information on loan applications, such as race, gender and income, loan types and application outcomes. For more information, go to *http://www.nicar.org/data/frb/* or call (573) 884-7711.

Stories

Story #16134. "Flipping, Fraud and Fantastic Profits," WBAL-TV Baltimore, 1999. The series covers corrupt real estate deals involving up to 2,500 properties.

To order, call the Resource Center at (573) 882-3364.

The Democrat and Chronicle story, "Mortgage Fraud in Suburbia," is availabe from the IRE Resource Center (Story #17289)

Census

For more information on the U.S. Census, see the Census Bureau Web site at *www.census.gov/*.

Tipsheets

More information on property and mortgages is available from the following tipsheets in the IRE Resource Center:

Tipsheet #1350. "Ten things every reporter should know about property records." Joe Kaplan offers pointers from the 2001 Reporting Conference.

Tipsheet #526. Allen Pusey provides a reference guide to property searches with a glossary and hints about filings and agencies. To order an audiotape, call (303) 649-1811 and request #IRE96-54, $10.

To order, call the Resource Center at (573) 882-3364.

Continued from page nine:

homes

each suspect property was given a unique ID number. That number and the address and subdivision became the basis of the first table. By the time we finished, this one had 237 records – one for each home the accused folks had been involved with.

A second table was created that included a record for each property sale involving one of the homes. The key information here was the buyer and seller names and the sale price, along with a code to note what type of sale it was. (These codes, and similar ones in the mortgages table, proved the most confusing part of the database, as so many permutations were possible.) After the land purchase and the various resales and intra-family transfers were tracked down and entered, this one reached 581 records.

Mortgages

The third table tracks mortgages plus second mortgages that were taken out by some home-buyers/co-conspirators. Fields were created to record what happened to each loan - whether they were paid off or, as has happened at least 131 times so far, the borrower defaulted, prompting foreclosure proceedings. This table, which is still being amended as new foreclosures are filed, tracks 626 loans.

Tracking made easy

This arrangement allowed us to track the amount spent to build the homes in question, the amount for which they were sold (in many cases to straw buyers or family members), the amount borrowed, the number of foreclosures.

We can track individual buyers or sellers, specific lenders, particular subdivisions-pretty much everything we need to. The database proved invaluable in many stories, including those done after three family members and a handful of associates were indicted by a federal grand jury. Nine months after the indictment, we continue to check the database for tidbits to lard our follow-up stories, and continue to update it with new foreclosures.

Steve Orr can by reached by e-mail at sorr@DemocratandChronicle.com

Continued from page eight:

boundaries

suburban districts saw similar demographic shifts. Only more distant suburban districts had increases in white kid populations.

Map helps pinpoint

The most powerful proof turned out to be a metro-area map that used dots to show where minority kid populations and white kid populations had grown. The map showed an island of white dots in the center, each representing 50 more minority kids, surrounded by a sea of darker dots, each representing 50 more white kids.

The story stayed conservative and said only that, "Child population shifts in the Oklahoma City area suggest that school district boundaries may be a factor in residential segregation."

Given more time, we would have talked with people who recently moved into or out of different school districts. If the qualitative data from the interviews supported what we saw in the quantitative census data, we would have been more confident saying this is a cause of segregation.

But this is an issue we hope to revisit next year when we get the long-form census data that will include income and occupation information. And something we noticed makes us eager to do so. Even in the suburban districts where the number of white kids grew, the minority populations grew faster. This suggests that concerns about class, as much or more than about race, may be driving families from the inner city.

John Perry can be reached by e-mail at jperry@oklahoman.com

Family offense

By Susan L. Oppat
The Ann Arbor News

I've spent most of my 21 years as a reporter working cops and courts, crime and punishment. It's been painfully clear that when it comes to the littlest victims, the children, there's plenty of crime, and not much punishment.

Largely, punishment relies on already-traumatized children having the courage to tell strangers they were sexually assaulted, often at the hands of their own family members. But even when they do come forward, children often aren't believed.

And, of course, there are the children who are too young to report. They are too young to speak – but not too young to be victimized.

Goals

I wanted to prove what I knew from experience that it happens, that the offender is only rarely a stranger, that it happens even in the best families, that the children are boys and girls, and every age.

I also wanted to speak with conviction about what happens to the people who sexually abuse children.

Self-trained on Microsoft Access, I created a database with 35 different categories of information, the analysis of which would be the basis for all my reporting. I included:

- The name of the suspect, age, gender, address, ZIP, previous criminal record, if known.
- The name of the child, age, gender, address, ZIP code.
- The location of the assault, date(s) of assault.
- The relationship, if available, between the defendant and the victim.
- Details on the nature of the assault, if available.
- Which police agency investigated.
- The exact charges, and how many were filed.
- The name of the judge, prosecutor, defense attorney.
- Whether the outcome was achieved by jury trial, bench trial, or plea agreement.
- The charges on which the defendant was convicted, if any.
- The sentence, and whether the defendant was sentenced to jail, prison or probation.

I took a laptop computer to the county felony court, and spent more than a week combing through 435 cases of sexual assault on children under the age of 17 filed in the previous 61/2 years. I had to limit myself to that time period because that's how long the cases have been listed on a computer database, and child sex assaults can be separated from all other sex assaults by an identifier on the case number.

I was not charged a fee to sit in the file room and read through the hard-copy files. Court records here are listed in a computer database, but the information contained in the files is not. So it all had to be done by hand. I was a rookie with databases, but two things became immediately apparent:

Consistency of input data is paramount.

- Consistency of input data is paramount. You must use a limited range of entries to get useful data. For example: Always refer to the police agency by acronym, or always spell it out. If you sometimes enter the acronym, and sometimes spell it out, your data won't be compiled accurately, and you'll be doing a lot of arithmetic by hand, always the quickest way to inaccurate data and ruined credibility.
- The other thing seems like a no-brainer but it's crucial. Back up your data. Often. At LEAST daily, if not more. Keep it on diskettes AND the hard drive.

You REALLY don't want to re-enter all that data.

Time to analyze

Then it was time to analyze the data, over a period of days:

How many of the defendants were men, how many women? Where did most of the defendants live in the county? What was the average age of the children? How many were related to the defendant, or at least knew the defendant? How many, really, were strangers? How many of the children were assaulted in their own homes? How many defendants were charged with the most serious offenses? How many of those pleaded

Continued on page twenty-six

The *Ann Arbor News* story, "Sex Assaults on Children, the Littlest Victims," is available from the IRE Resource Center (Story #17295).

Data:

County court records on 435 cases covering 6 1/2 years.

Software:

Microsoft Access

Tipsheets

More information on covering sexual abuse, and crime in general, is available from the following tipsheets in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #562. From a panel from the 1996 IRE National Conference. This tipsheet includes two sensitive, compelling **stories on sexual abuse.**

Tipsheet #1305. Geoff Dougherty explains what kind of data is relevant for CAR investigations on violence and crime.

To order, call the Resource **Center at (573) 882-3364.**

Tipsheets

More information on covering drunken driving is available from the following tipsheets in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #868. Rob Gebeloff provides a list of useful Web sites for researching a transportation story.

Tipsheet #973. Kathleen Johnston offers a list of tips for using CAR in a broadcast story.

Data

- **Drunken driving citations from Montgomery County police**
- **Drinking Driver Monitor Program database from state Office of Probation and Parole**
- **National Highway Traffic Safety Administration's Fatality Analysis Reporting System (or FARS, available from NICAR)**
- **Circuit court records, collected from dial-up access and parsed using Monarch**

Continued from page one: **drivers**

Those who had killed people in alcohol-related collisions often didn't spend a day behind bars. Drivers with arrests in the double digits often got lighter sentences with each new conviction. Prosecutors sometimes didn't know drivers' full records, leaving those who had killed others to be sentenced as first-time offenders.

In examining who was responsible, the *Post* found prosecutors ignored repeat offender laws and expedited cases through generous plea bargains. Judges were unaware of mandatory jail laws for habitual drunken drivers and failed to distinguish between second-offenders and those with more than 10 arrests. The *Post* also found that key lawmakers, some of them attorneys who defend drunken drivers, were causing the state to lose millions in federal highway funds every year because of their refusal to crack down on their clients.

The reporting began with a request for district court officials to provide us electronic court data, which they refused to do. This led to a decision to sample cases in Montgomery County.

Those who had killed people in alcohol-related collisions often didn't spend a day behind bars.

County police generated for us a list of all drunken driving citations from 1997, giving the *Post* a starting point. Among the information contained was the defendant's name, date of arrest, date of birth, home address and charge, but nothing about what happened beyond the point of arrest. The data – about 5,000 cases – were keyed into Excel.

At that point, we decided it would be unwieldy to research each of these 5,000 cases, so a random sample of roughly 500 cases was extracted and isolated in a new spreadsheet.

We began what eventually became a countless number of trips to the courthouse to research the outcomes of each case in the sample, adding that information to our database. Files had to be culled by hand for case details because only limited information was available from district court computer terminals and through a dial-up account. Ultimately, among the particulars we tracked: manner in which the case was disposed, judge involved, type of sentence, fines, community service, probation, any treatment required, type of plea, defense attorney, time from arrest to disposition, etc. On key cases, we requested audiotapes of court proceedings, so we could listen to the judges' and prosecutors' remarks and other details not in paper files.

Creating a picture

The idea was to create as full a picture as possible of each case in the sample, although at that time there was little sense of where this would lead.

One of the reporters involved in the project left the paper, and the initial round of reporting aged for about eight months before it was again actively pursued. At this point, court cases had to be re-researched, as many of the pending cases, had since been disposed. As this process got underway, it became apparent that some of these drivers had been arrested up to three times since the initial research on the 500 sample cases. This forced us to re-evaluate the situation: Repeat offenders were returning quickly to the streets, with little or no punishment or treatment. We now knew recidivism was a major part of the story, and that we wanted to focus on chronic offenders, not one-time drunken drivers. We wanted to more thoroughly build a comprehensive picture of drivers in the sample, and all of their arrests for drunken driving, before and after the sample case.

Were there other resources we had overlooked?

Road blocks

Driving records from Maryland Motor Vehicles Administration were of little help, as they were limited to the past three years of activity and drivers often privatized their records entirely. Computerized court indexes used to look up individual cases were incomplete: Well into the project, we learned that the District Courts strip (for storage space) traffic cases from the computer indexes and transfer them to microfiche generally unavailable to the public. Arrests that had appeared in the records a year ago, could no longer be

Continued on page thirteen

Continued from page twelve **drivers**

found in the main system.

In the course of reporting, we learned that the state office of probation and parole maintained a database of drunken drivers ordered into its "Drinking Driver Monitor Program." After an open records request and some negotiating, we obtained a copy of their database.

The database (a .dbf file) was provided to us on a tape, which we transferred onto a CD. There were about 15 years of records, each one representing a defendant's pass through the program. After scrutinizing about 275,000 records (using FoxPro) for consistency and integrity, it became clear that the data had weaknesses: sentencing information usually only included the amount of probation, but not actual time in jail; information on fines, etc. was limited. Again, we could not draw statewide conclusions, nor compare jurisdictions. But, the data did give us a quick way to ferret out other drivers from Montgomery County who had as many as 15 drunken driving arrests.

And, equally as important, it helped us pinpoint arrests for some of the drivers in our sample that we had been unable to find in court or motor vehicle records.

Refused again

About this time, we also began pursuing a second track of the story: How did drunken drivers who killed or maimed others in traffic crashes fare in court?

Again, we turned to the courts, asking for a list of all who had been charged with some sort of drunken driving injury or fatality accident, and again, they refused to cooperate. Police and prosecutors also were of no help: They said they were unable to provide us names of defendants involved in injury and fatality, alcohol-related crashes. And Department of Corrections prison or jail records, an option in other states, can only be accessed in Maryland when the defendant is still in the system. Once gone, a prisoner's records are generally private.

We also sought records from the state medical examiner regarding all drivers who died automobile crashes and had alcohol in their systems at the time. That office refused our request. We sued, and later won in court, but too late for use in the series.

We then resorted to the National Highway Traffic Safety Administration's Fatality Analysis Reporting System to identify fatal, drunken driving crashes in Maryland. The data contains a broad range of information about fatal crashes, including whether alcohol was suspected and the date and location of the crash, but not the names of those involved. Extracting a list of candidate crashes by date and location, we then turned to newsclips and police sources to get names of involved parties. This helped us identify some drivers, but not enough. FARS data was incomplete or in error and many crashes went unreported.

We began what eventually became a countless number of trips to the courthouse to research the outcomes of each case in the sample, adding that information to our database.

Eventually, we triangulated resources. Using our dial-up access to an indexing of Circuit Court records, Dan Keating was able to generate – through a script that captured screen displays which were parsed using Monarch – a list of everyone charged with killing someone while driving drunk in the county since 1982. This finally provided us the universe of defendants in crashes in which the drunken driver had survived.

Stepping back

It quickly became apparent sentences for surviving drivers in these fatality crashes were not much different: Some drivers received probation after fatal crashes. And, some of these deadly drivers went on to be arrested again for drunken driving, but prosecutors, judges and police were unaware of their prior involvement in alcohol-related fatalities.

Stepping back, we began to focus on the cases and drivers who best demonstrated the

Continued on page twenty

Stories

Other stories on drunken driving are available from the IRE Resource Center (*www.ire.org/resourcecenter*):

Story #16229. "Case Dismissed," WTHR-TV (Indianapolis). Records show thousands of cases of DUI's dismissed because police witnesses fail to show up for court dates.

Story #14312. "Serving justice: DWI and district court, Fayetteville (N.C.) *Observer-Times*. The paper analyzes district court with the lowest DWI conviction rate in the state and finds inflated conviction rates and judges interpreting laws differently.

To order, call the Resource Center at (573) 882-3364.

COURTS

Dragging justice

By Polly Saltonstall
Freelance writer

For years New Bedford, Mass., residents questioned why police seemed unable to make inroads on the city's escalating drug woes. They complained that dealers once arrested would show up back on the streets selling again within weeks.

My editors asked me to take a look at how drug cases fared in the court system and to determine if the courts were part of the problem.

The resulting series "Busted Justice: Drugs, the Law and Bristol County" did just that. We found that despite tough state drug laws designed to minimize disparities, the courts were churning out an unpredictable, crazy-quilt pattern of sentences. In some cases, defendants who appeared to be low-level dealers or drug couriers were sentenced to as much time, sometimes more, than people with long records. Some defendants with previous drug convictions avoided jail completely, while others charged with their first offense ended up behind bars for years.

Recidivism high

And as a glut of cases bogged down the court system, delaying trials in some cases for as much as two years, recidivism among defendants was high. Many defendants were charged with new drug crimes while awaiting trial for previous arrests. But many of these people facing multiple cases were able to lump them together and serve one sentence for all charges. Most of the most serious cases – such as drug trafficking where prison terms are mandatory upon conviction – were either dismissed or reduced, as prosecutors used the threat of the mandatory sentences to extract guilty pleas on lesser charges for much less jail time.

Doing this as a CAR project was a big deal for the *Standard Times*, since no one on the staff knew anything about data analysis. I knew a little Excel but that was it. We had limited computer capacity and an extremely limited budget.

I started my data collection with a computer printout of all the people arrested in 1999 on drug-related charges (cops told me they could not provide this on disk and I would not have known how to handle it anyway if they had). A newsclerk typed that list into Excel, and I then took it on my laptop computer up to district court where I looked through paper files for matches with my arrests.

We found that despite tough state drug laws designed to minimize disparities, the courts were churning out an unpredictable, crazy-quilt pattern of sentences.

The court's computerized files included minimal information, such as names and dispositions. But I wanted to know more about the process, such as whether the case was resolved through a trial and whether charges had been reduced or dismissed as a result of plea bargains. So I spent three weeks going through more than 2,500 files (all the court's cases for one year) looking for matches. For each case, I noted: the arresting officer, the date of arrest, the defendant's name, address and birthdate, the applicant's place of employment and birth, the judge, the prosecutor, the original charge and whether that charge was reduced as a result of a plea bargain, and the type and date of disposition.

Complex database

The big problems began when I tried to analyze my database, which ended up being about 2,200 lines with 30 fields. All I knew was Excel and it quickly became apparent that my questions were too com-

Continued on page fifteen

Saltonstall's series, "Busted Justice: Drugs, the Law and Bristol County," is available from the IRE Resource Center (Story #17118).

Stories:
Other stories about drugs are available from the IRE Resource Center (*www.ire.org/resourcecenter*):

Story #15057. *The Charleston (W.V.) Gazette* examined the role of confidential informants in the war on drugs. It found that people who help police arrest suspected drug dealers can earn thousands of dollars and avoid criminal charges and lengthy prison sentences of their own.

Story #12458. A series by *The Boston Globe* exposed a pattern of injustice in law enforcement's pursuit of **drug crimes in Massachusetts.**

To order, call the Resource Center at (573) 882-3364.

Continued from page fourteen: **justice**

plex for that program. I called the folks at IRE who told me to use Access. But neither I nor anyone in my newsroom knew how to use Access. So I signed up for the IRE Bootcamp and took my database with me to Missouri. At the bootcamp, Tom McGinty showed me how to import my database into Access, but then realized I had lumped too much data into individual cells. For example, I had put complete sentencing information in one column. This included whether a defendant had been sentenced to jail time and how much, as well as length of probation. I had to separate the data in separate fields in order to answer such questions as what percentage of people found guilty had been sentenced to jail and then to figure their average jail time.

Cleaning the data and crunching the numbers took another two months because I was still learning the program as I went, and also because I became sidetracked on another story involving potential police corruption and/or misconduct. My queries included finding the percentage of convictions, percentage of guilty findings where charges were reduced as a result of plea, type of drugs involved, average ages of defendants, number of dealers versus users, and the length of time between arrests and sentencing.

Startling finding

One startling finding was that very few people charged with dealing drugs in a school zone, a charge that carries a minimum of two years in jail under Massachusetts law, were actually sentenced on that charge. Prosecutors told me this was because the school zone law was flawed. Still these same prosecutors adamantly opposed a legislative proposal to loosen that particular mandatory sentence.

The next wrinkle was finding a way to compare my data with other states and/or regions. Although I obtained national drug conviction statistics from the Department of Justice, I was reluctant to compare them with the local data or rely too heavily on those comparisons without knowing whether I was comparing apples with apples.

Statewide sentencing data was based on final dispositions and final charges. It did not consider the impact of plea bargains and to what degree charges were reduced as a result of pleas. The most conclusive information came from records obtained from the prosecutors in a neighboring county that showed the local prosecutor was making more deals than his other colleagues.

Hitting the pavement

Once I had a set of raw numbers and conclusions, with some help from two other reporters, John Doherty and Aaron Nicodamus, I hit the pavement and began reporting. This entailed interviews with defense attorneys, prosecutors, judges, police, policy makers, including members of a state task force convened to analyze state sentencing trends.

All I knew was Excel and it quickly became apparent that my questions were too complex for that program.

In an attempt to give the numbers faces, I used a public records request to obtain criminal background reports on almost 40 defendants and interviewed many of these same defendants or their attorneys – for those in prison that proved time consuming because corrections officials required inmates to write me a letter agreeing to interviews and not all wanted to talk. Massachusetts charges a $25 fee, set by state law, for an individual's criminal record. I found the information provided by the state corrections system did not always correspond to information in the local prosecutor's files, but that's another story. I sat through dozens of court trials

Continued on page twenty-six

Tipsheets

More information on covering drug trends is available from the following tipsheets in the IRE Resource Center:

Tipsheet #1249. Rose Ciotta offers advice on using data sources to track illegal drug trends.

Tipsheet #1411. Karen Dillon explains how "law enforcement agencies across the country were evading state laws to keep millions of dollars in seized drug money with the help of the U.S. Department of Justice."

To order, call the Resource Center at (573) 882-3364.

More information about the listserv is available on the NICAR Web site, *www.nicar.org*

Tipsheets

More information on software is available from the following tipsheets in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #1303. Justin Mayo explains how to use Monarch, which allows users to "easily extract and analyze data from report files downloaded from any mainframe or mini-computer." Monarch can take print image files and convert it into a database or spreadsheet.

Tipsheet #282. If you're trying to get an editor to spring for some software or hardware, Rose Ciotta offers some advice on selling CAR to your superiors.

Tipsheet #160. Penny Loeb offers some basic advice on hardware and software needs for CAR reporters. The tipsheet also includes several CAR stories by Loeb.

To order, call the Resource Center at (573) 882-3364.

SOFTWARE

Affordable CAR

By Aron Pilhofer
Director, IRE/NICAR
Campaign Finance Information Center

I was well into a major campaign finance project last year when I realized I had a problem. I needed to analyze a database of two million records, and I knew that was a job I could not coax Microsoft Access into handling.

I needed a high-octane database manager that wouldn't break the *News Journal's* modest CAR budget. That ruled out SQL Server, Oracle and every other commercial database application I could think of.

After a few days of searching for a workaround, a colleague suggested I look into something called open source software. I had heard the term before, but assumed it meant just another flavor of low-end freeware or shareware.

I needed a high-octane database manager that wouldn't break the *News Journal's* modest CAR budget.

I was skeptical, but even more desperate, so I gave it a shot. An hour after downloading and installing a database application called MySQL, I knew how misplaced my concern was. MySQL was not only the answer to my specific problem, but I was able to find other open source applications that allowed me to extend my newsroom's CAR program well beyond what I thought was possible -- or affordable.

By the time I left the *News Journal* in June, reporters could search campaign finance records or voting records using our internal intranet, or, those with more advanced CAR skills could connect to any database in our library using Access as a graphical user interface. The tools we used were all open source: Linux, Apache, PHP and MySQL. The cost to the organization was zero.

No matter the size of your newsroom or budget, open source software may be worth a look for you too.

What is it?

Open source software is often confused with freeware or shareware, but it is neither. It is a radical departure from traditional methods of developing and distributing software.

In practice, open source means just what it implies: The source code – or lines of instructions that tell the computer what to do – is available for anyone to read, modify and redistribute. But there is a catch: Everyone must play by the same rules. Anything based on an open source application must itself be open sourced.

The open source license ensures that enhancements, fixes or other changes are freely available to everyone and offers incentive for collaboration. Linux, probably the best-known open source project, involves thousands of programmers worldwide. (For more information, start with the Open Source Initiative web site, *www.opensource.org*).

The approach has been remarkably effective, much to the chagrin of corporations like Microsoft (which just launched a public relations blitz to discredit open source software). Linux has steadily increased in popularity, and was the operating system of choice on about 30 percent of all servers sold in the U.S. last year.

Apache is the Web server application running about 60 percent of all Web sites on the Internet – ironically, even a few owned by Microsoft. And MySQL is becoming one of the most popular database servers for mid- to large-scale Web sites, including Yahoo Financial, slashdot.org and NASA.

Open source: the good

• **Cost:** Open source software is almost always free for non-commercial uses and usually can be downloaded straight off the Internet ready to install or purchased on CD-ROM for about the price of the media

Continued on page twenty-two

Danger below

By MaryJo Sylwester
Center For Public Integrity

The world of underground fuel storage tanks has largely gone unnoticed by the media and environmental groups in the two decades since Congress first authorized environmental regulations for tanks.

At the same time, vast changes in the industry brought about by these new regulations have caused numerous problems that should be brought to the public's attention. Computer-assisted reporting can make that investigation a little easier.

With the help of CAR, *Kansas City Star* environment writer Michael Mansur and I uncovered the fact that thousands of leaking underground storage tanks continue to pollute water and soil, despite the new regulations and billions of dollars spent on prevention.

Through the data work and searching paper records we found leak detection was not working.

Our story, published April 22, 2001, in *The Star*, came a month before a General Accounting Office report that called for congressional review of these problems.

We found that the regulations and problems vary from state to state, and as a result, you'll find a slightly different story in each state. But here are some guidelines to get you started.

The laws

Congress authorized the first environmental regulations for underground storage tanks (UST) in 1984, requiring owners to register their tanks with the state. In 1988, Congress passed a series of laws that took effect gradually over the next decade. By December 1998, all tanks were required to have leak detection, corrosion protection and overfill protection. The law also banned the use of bare steel tanks and required owners to prove financial responsibility in case of a leak.

More information about the federal laws is available on the Environmental Protection Agency Web site, *www.epa.gov/swerust1/*.

State laws

States were required to either adopt the federal laws or enact their own, more stringent, regulations. Idaho is the only state that has not done this. Of the other states, most enacted a mirror of the federal regulations. A few, such as Florida and California, have more stringent laws. For example, Florida also requires tanks to be in an outer container, called secondary containment. Others ban fuel delivery to tanks that don't meet the state regulations.

We wanted to determine whether the new regulations were effective. We discovered that it was difficult to assess the effectiveness of the newest regulations (overspill and corrosion protection), but a much more important piece of the environmental laws – leak detection – had been in place since 1993.

Through the data work and searching paper records we found leak detection was not working. Even those leaks discovered after all tanks were required to be fully "upgraded" in late 1998 were primarily found when the owner took the tanks out of the ground, sometimes months or years after the leak first started.

Databases

To uncover this, we started with the Missouri and Kansas databases of tank registrations and leaks. Generally this information is kept in separate tables and possibly separate databases.

All states should have a database of tank/facility registrations. It may include both underground and aboveground tanks, depending on how your state program is set up. Most states don't have regulations for aboveground tanks, but they collect information from those who volunteer it. Leaks from aboveground tanks

Continued on page eighteen

The Kansas City Star story is available from the IRE Resource Center (Story #17723).

To order, call the Resource Center at (573) 882-3364.

Data

Tank registration, LUST, enforcement and inspection data for Missouri and Kansas underground storage tanks

Source

Missouri Department of Natural Resources and Kansas Department of Environmental Quality

Software

Access, FoxPro and Excel

Continued from page seventeen: **danger**

must be reported, however.

We obtained these databases from the storage tank units of the state environmental agencies. Both states required payment for copying the data, but it was less than $50, and they were prompt in responding. In both cases we received the information either in database format (dbf) tables or an Access database.

The first request was a little tricky in both states because we found neither agency was accustomed to giving out this information. At first, Kansas gave us their "public" version that doesn't include all of the information. Missouri, on the other hand, gave us everything the first couple times, then suddenly decided the comments field in the leaking tanks data wasn't public.

Many states offer a "public" version on their Web sites that would be worth checking out before making a request. Links to state Web sites are available at *www.epa.gov/swerust1/states/stateurl.htm*

The full databases could consist of multiple files, such as:

- One listing each facility/site that has underground and/or aboveground tanks.
- One listing each tank located at those facilities.
- Owner information.
- A leaking sites database.
- Enforcement actions data.
- Inspections records.

Facility/tanks

The facility portion will identify the particular gas station and its location. It might include geographical locators such as latitudes and longitudes that could be used to map the data. If so, be sure to also ask what datum they used to determine the "lats" and "longs."

The tank portion will provide details about each tank: whether it is steel or fiberglass, when it was installed, the type of leak detection used, whether it meets the regulations, and many other details. The data may also include those tanks that have been "closed" or removed from the ground, as well as a date when it came out of service. This can be useful for understanding whether a site might be contaminated because of old tanks used prior to the regulations.

The Kansas data kept a separate record on each tank for each year. As a result, you could easily track changes in the tank status or compliance with the laws over several years. Missouri changed the information in the same record whenever the tank status changed; as a result, we lost a lot of historical perspective.

We found that about 30 percent of the active tanks in Missouri had not been inspected during the last three years.

Owner information might be included with the facility or tank data, but it might also be kept separately. This should include the name of the company, a contact person, phone numbers and addresses.

State oversight

We also obtained enforcement and inspection data for Missouri. The state only provided data on "closed" enforcement cases and said the open cases were exempt from the open records law. Since most of the tank laws had just gone into effect a year or two earlier, most of the enforcement cases pertinent to our story were still open.

The inspections data was useful to study the aspect of state oversight.

By doing a subquery between the facility table and the inspection table, I created a list of all active sites (those still pumping gas) that had NOT been inspected between 1998-2000. Then I created a list of facility ID's and the date of its last inspection. To get the last inspec-

Continued on page nineteen

Useful Web sites:

EPA Storage Tanks Web site: *www.epa.gov/swerust1/*

Links to state tank agency Web sites: *www.epa.gov/swerust1/states/stateurl.htm*

Corrective Action Measures reports (semi-annual) to the EPA: See some of the basic data in PDF tables at the EPA Web site: *www.epa.gov/swerust1/cat/camarchv.htm*

GAO Report, May 4, 2001, GAO-01-464: *www.gao.gov* **(choose to search the GAO archives for the above report number)**

The Association of State and Territorial Solid Waste Management Officials: *www.astswmo.org/tanks.htm*

EPA Report to Congress and other reports: *www.epa.gov/swerust1/pubs/index.htm#rtc*

Continued from page eighteen: **danger**

tion date from the list of inspections (there might be multiple for each site), I did a subquery to find the "maximum," or most recent inspection date.

From that I created a table listing the name, address, city, owner, region, whether it had been upgraded and date of last inspection. We found that about 30 percent of the active tanks in Missouri had not been inspected during the last three years, the maximum time-span recommended for adequate state oversight.

The leak database should list one record for each facility that has reported contamination from a leaky tank. Sometimes it's called the "remediation" or LUST data.

Continued on page twenty

ENVIRONMENT

More tank stories

By MaryJo Sylwester
Center For Public Integrity

The *Kansas City Star* story covered the big picture – the effectiveness of the regulations – but there are many smaller pieces to this puzzle. You might find a story here that is easier and less time-consuming.

Insurance funds: Most states have a gas-tax-funded insurance fund that helps clean up contamination from leaking tanks. Tank owners pay premiums to be covered in case their tanks leak, but the bulk of the cleanup funding comes from a gas tax. Originally, the states planned these funds as temporary measures to ensure sites were cleaned up, with the hope that private insurance companies would take over after the new regulations stemmed the bulk of the leaks. Since the leaks are continuing, a few state funds are starting to run out of money.

Inspections: Focus on the state oversight aspect by studying the inspection system. How many sites do the inspectors visit annually? What's the workload like for the typical inspector? What do the inspectors generally find in terms of compliance with the laws?

Equipment: We found that some tank owners did not properly use leak-detection and corrosion-protection systems. Some incorrectly programmed the computer for the leak detection, ignored alarms indicating a leak, or turned off the electricity necessary for corrosion protection. There is also growing evidence nationally that the interior lining in steel tanks is sloughing off, leaving the tank susceptible to corrosion.

Advice

This area can be a quagmire. But it's worth wading into. Here are my recommendations, if you're going to dive in:

- Spend as much time as possible in advance getting to know the lingo and the laws of underground storage tanks before starting the reporting. I found people in this industry are particularly harsh to reporters who don't know what they're asking about. Once I could speak the lingo, industry officials were frank and forthcoming, as if I was a colleague. Look for brochures and newsletters geared toward tank owners.
- Request a handful of paper files for leaking sites and read those before you get too far in your investigation. By going through even just a few documents, I really got a better sense of how the system operated and what all the lingo meant. I could also clearly see how the database was constructed and identify the pitfalls.
- Be prepared to spend a good deal of time on this story, even if you pick out one of the smaller stories. Simply getting to know the lingo will take you a few weeks. This is a good story for picking at slowly while you're working on other stories.

MaryJo Sylwester can be reached by e-mail at msylwester@publicintegrity.com

Beat Books available from IRE:

Understanding Crime Statistics:
"A Reporter's Guide," by Kurt Silver.

Covering Aviation Safety:
"An Investigator's Guide," by Marie Tessier.

Home Mortgage Lending:
"How to detect disparities," by Jo Craven McGinty.

Numbers in the Newsroom:
"Using Math and Statistics in News," by Sarah Cohen

Each book is $15 for IRE members and $25 for non-members. More information about these and other books available from IRE can be found at *www.ire.org/store/books*.

To order call (573) 882-3364.

Uplink
is NICAR's bimonthly newsletter on computer-assisted reporting written by top reporters around the country. 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. The newsletter also includes information on **upcoming NICAR events.**

One-year subscription (IRE Member) ($40)

One-year subscription (non-member/ international) ($60)

Single volumes ($10/12 international)

Continued from page nineteen: **danger**

The term LUST is the one spot of humor in an otherwise serious industry. Instead of what you might think, LUST stands for Leaking Underground Storage Tank.

The Kansas data made it easy to determine how many leaks were discovered by leak detection because they had a field identifying that information in the LUST data. In Missouri, we pulled that information from paper records and created a new field for it in the database. In both states, leak detection found less than 5 percent of the leaks that occurred after December 1998.

Tank status

The most difficult part of the analysis was determining the status of the tanks at the time the leak occurred. We wanted to know if the tanks had been either replaced or "upgraded" to meet the new laws prior to the leak, or if the contamination occurred before the regulations took effect but didn't get reported until later. Missouri and Kansas didn't directly document this information in the data. Generally, we found it was available indirectly either in other parts of the database or in paper records.

Within the database, I used the tank installation and "out of service" dates for the tanks and compared those to the date the leak was discovered. I used paper records to fill gaps in the database. Both Missouri and Kansas were terrible about consistently filling in fields. The Kansas database could be a gold mine for a reporter because they have fields to track every little thing, but the field agents responsible for entering the information do so haphazardly.

EPA data

This is one of those occasions where the federal level collects very little data. States submit two reports to the EPA each year. These include the number of leaks, active tanks, closed tanks, sites that started clean-up, sites that finished clean-up and number of emergency responses made by the agency. These numbers are then compiled into a list that is posted on the EPA's Web site in PDFs.

The paper reports include other details that provide a good overview of the state program and what things they are working on. You should be able to obtain the reports from either the state agency or the regional EPA office in your area. (In Missouri, their records department kept all of the EPA reports in one file that you could request).

I put the numbers for end-of-the-year reports 1995-2000 from the PDFs into Excel and did some basic calculations.

MaryJo Sylwester can be reached by e-mail at msylwester@publicintegrity.com

---

Continued from page thirteen: **drivers**

shortcomings in the court system, putting a face on the issue. We constructed pivot tables and examined trends in prosecution of drunken driving cases, using this analysis and individual cases we had uncovered as the basis for interviews with those responsible. In the end, it was clear that drunken driving cases were one of the lowest priorities for prosecutors and judges.

In looking back, Katie and I think one of the reasons the project was successful was the pairing of her beat reporting with my computer skills. Having tackled a similar investigative project on drunken driving at the Tulsa World, the experience provided a point of reference on the issue.

The biggest hurdle to the project? Compiling a complete, accurate, picture of the chronic offenders' drunken driving histories — a fact not lost on these drivers who count on missteps by police, prosecutors and judges.

David Fallis can be reached by e-mail at fallisd@washpost.com

A pet project

By Sean Mussenden
The Orlando Sentinel

Never completely trust government data. Sounds like a no-brainer for veteran CAR disciples.

But it took a bit of simple spreadsheet work and a chance discovery to hammer that point home for two young reporters.

My colleague Mark K. Matthews and I, both working for a student wire service run by the University of Maryland's College of Journalism, broke a national story last fall that embarrassed federal safety regulators investigating Firestone tire failures.

... the error exposed a serious flaw in a dataset that every major newspaper, wire service and network news program in the country was using ...

We disclosed that the National Highway Traffic Safety Administration had miscalculated the number of human deaths linked to the tires with dead pets.

Granted it was only two animals – a German shepherd and a Persian cat – out of 103 total deaths reported to the agency as of Sep. 15, 2000.

Major problem

But the error exposed a serious flaw in a dataset that every major newspaper, wire service and network news program in the country was using to supplement their blanket Firestone coverage.

My first venture to the NHTSA Web site began as a quest to localize the national Firestone story. Stories by the University of Maryland's wire service, Capital News Service, are largely run by newspapers in Maryland.

In September, my editor in the Maryland state capital bureau of Capital News Service, Adrianne Flynn, wanted to find the number of Maryland accidents linked to Firestone tires.

First step

Like many journalists across the country, I downloaded NHTSA data from the Web into an Excel spreadsheet. I filtered for accidents reported from Maryland and did some quick analysis, mostly looking at the type of accidents (lots of rollovers), vehicles (usually Fords), and injuries (no fatals). After double and triple checking my work, I wrote a short sidebar for our Maryland clients to run along side the national stories.

Two weeks later, when NHTSA updated their data and added more records, it seemed natural to again trot out the Maryland data for a story. Again, I downloaded the spreadsheet, and did some quick analysis.

But this time there was a fatal listed for Maryland, a surprise given that most of the fatal accidents linked to Firestone tread separation occurred in warmer Southern climes.

We needed to verify that it was correct with the hard copy of the accident record given to NHTSA. We also needed more information, specifically names and contact information, to begin fleshing out the story (We would learn later that NHTSA redacted all names.)

D.C. help

From the spreadsheet, I pulled the individual record number and called Matthews, stationed in the Capital News Service Washington, D.C. bureau. He went to the records office, and, after a predictable hassle, obtained a copy.

Like the spreadsheet, the hard copy showed two fatals. But the short narrative said nothing of people dying, just two animals. We knew we had a story, if we

Continued on page twenty-three

The story by *Capital News Service* on the NHTSA/Firestone data is available from the IRE Resource Center (Story #17269).

Stories

Other stories on the NHTSA/Firestone problems are available from the IRE Resource Center (*www.ire.org/resourcecenter*):

Story #16832. *Newsweek* reporters reveal that Firestone knew of tire flaws back in 1998 and obtained documents that show Firestone was chronicling a pattern of tire failures.

Story #17613. CBS News investigates the Firestone and Ford companies. And even after they initially insisted there was no problem and no need for a recall, CBS discovers compelling evidence of coverups.

Stories

Other stories on campaign finance are available from the IRE Resource Center (*www.ire.org/resourcecenter*):

Story #17624. Mother Jones investigates how "campaign contributions [during the 1999-2000 election cycle] are divided into 10 broad industries ranging from agribusiness to transportation." A major finding is that "candidates who raised more money than their opponents captured all but 29 of the 469 seats up for grabs in Congress – and the White House as well."

Story #17242. "In The Buying of the President 2000," Charles Lewis and the Center for Public Integrity shine a spotlight on the special (and often secret) interests that have heavily invested in the politicians who are seeking the nation's **highest office."**

Continued from page sixteen: **affordable**

and shipping.

• **Quality:** Because so many developers are working on a project at one time, serious programming bugs are rare and are usually fixed almost as quickly as they are found. Enhancements are added remarkably quickly as well. At present, for example, Linux is the first and only operating system that supports Intel's newest generation microprocessors.

No matter the size of your newsroom or budget, open source software may be worth a look for you too.

• **Cross-platform compatibility:** Open source developers have no incentive to "lock" users into a particular operating system or platform – in fact, just the opposite. MySQL, for example, runs on 11 different hardware/operating system configurations, including Windows and Macintosh. That makes porting data or code from one platform to another virtually painless.

• **Speed:** Open source software does not suffer from performance-killing "feature bloat" that plagues so much proprietary software. Enhancements are added only when needed, and with a premium placed on performance.

Open source: the bad

• **Documentation:** Great programmers rarely make great technical writers, and it shows. The larger projects (including those mentioned here) all are extremely well documented, but that is not true across the board. Patience is a must.

• **Learning curve:** Much of the software useful for CAR is designed for the enterprise-level applications, and not targeted at desktop users or novices. Do not expect to always find a user-friendly interface, a manual free of highly technical compuspeak or point-and-click simplicity. Did I mention patience?

• **Support:** Official support is nonexistent. The software is usually distributed as-is, without warranty or bundled support. However, there is almost always an active community of users, who are happy to help solve problems that crop up. There are also third-party vendors that offer support packages at a price.

• **The "L" word:** Linux is the project that gave real momentum to the open source movement, and remains the operating system of choice for most developers. Porting applications to Windows or Macintosh is a recent phenomenon. Most ports I have tried run great in Windows, although they are sometimes more stable and faster under Linux. Bottom line: You may have to learn to love the penguin.

What's out there?

Plenty, and I am finding more every day. The best repository of open source applications is *sourceforge.net*, here are a few I have found useful over the past year:

• **MySQL:** One of the fastest database managers and servers available, period. It is optimized to serve data for Web applications, but MySQL can function as a heavy-duty database manager for desktop use. No built-in GUI, but Access (or any ODBC-compliant database manager) can be used as one. Easy to install on Windows or Linux and great documentation. Web site: *www.mysql.com*

• **PostgreSQL:** Another high-end database server. It isn't as fast as MySQL but it does support many more features, including views and subqueries. It is not nearly as easy to install under Windows, however, and there is no Macintosh version I am aware of. Web site: *www.postgresql.org*

• **Mapserver:** An open source application to build Web applications using GIS, similar to ArcIMS, but without the hefty price tag. Runs under Windows as well as Linux. Check the Gallery section of the Website for examples of Mapserver in action. Web site: *mapserver.gis.umn.edu*

• **PHP:** A Web scripting language similar

Continued on page twenty-three

Continued from page twenty-two: **affordable**

to Active Server Pages, but faster and much more flexible. PHP scripts will run unchanged on just about any platform, and has built-in support for nearly every database manager on the market, including Oracle and SQL Server. Web site: *www.php.net*

Four more

• **Python:** A simple, yet powerful, programming language that is ideal for novice programmers or those who need programs that run on many different platforms. Python applications will run under Linux, Windows, the Macintosh OS, OS/2 and a number of other operating systems. Web site: *www.python.org*

• **Apache:** The most popular Web server application in the world, and with good reason. It is fast, secure and runs on 28 different operating system/hardware combinations. Configuration can be tricky, however. Web site: *httpd.apache.org*

• **StarOffice:** Sun Microsystems recently open sourced this integrated office suite, and it is available for download. The current version 5.2 is so-so, but Sun is promising big things in the upcoming 6.0 release. Web site: *www.sun.com/products/staroffice/*

• **Gnumeric:** A spreadsheet application that rivals Excel feature-for-feature. Gnumeric is Linux-only right now. Web site: *www.gnumeric.org*

Aron Pilhofer can be reached by e-mail at aron@ire.org

Continued from page twenty-one: **project**

could confirm in the few hours before deadline that NHTSA was including these two pet deaths in their oft-repeated number of deaths linked to the tires.

After a bit of pressure ... we finally got a spokesman to confirm the error.

After a bit of pressure – Matthews in person at NHTSA and I over the phone – we finally got a spokesman to confirm the error. He said the agency would remove the two pets from their totals and comb through their records in search of any more errors.

Some of our regular clients picked up the story, and several national newspapers, nightly news programs, and wire services ran stories about our discovery.

Reuters, in their version, had the good sense to call People for the Ethical Treatment of Animals for comment on the pets removal, something I did two days later in a follow-up I wrote about NHTSA naming the animals on their Web site. Predictably, PETA was outraged that the pets were removed from the list.

Lessons

I suppose this story taught us two things

• Always take great pains to verify the accuracy of a dataset before reporting anything. For us, that meant looking at the hard copy of the accident record. If we had not, and simply printed that two fatals had been reported from Maryland, our story would have been incorrect.

• It does not take a months-long database project to write a good story using CAR skills. Just using basic techniques in day-to-day reporting, like filtering in Excel for example, can lead to a decent story.

Sean Mussenden can be reached at smussenden@orlandosentinel.com.

Mark K. Matthews can be reached at mmthews@wam.umd.edu.

Attend a computer-assisted reporting bootcamp in Columbia, Mo.

CAR Boot Camps are weeklong, on-campus seminars at the Missouri School of Journalism from a Sunday afternoon to a Friday afternoon. They give journalists a jump start in computer-assisted reporting techniques. Our regular boot camps are scheduled about five times per year.

These unique seminars train journalists to acquire electronic information, use spreadsheets and databases to analyze the information and to translate that information into high-impact stories. In addition, the institute then provides follow-up help when participants return to their news organizations.

Upcoming bootcamps:

- Jan. 6-11, 2002
- March 24-29, 2002
- May 19-24, 2002
- Aug. 4-9, 2002

Continued from page six: **performance**

in these regions if they were within 50 miles of a major aerodrome.

4. **Organize!** Because database builders have to deal with lots of individual bits of handpicked data, you've got to computerize those bits as soon as you get them. We plugged our data daily into an Excel 6.0 master spreadsheet, which allowed us to add, subtract, move, and modify data fields and records according to what the concurrent reporting for the article was telling us. The spreadsheet served three important organizing functions: 1) a behind-the-scenes storehouse of all information that might go into our final ratings; 2) an easy calculator for converting raw numbers into more meaningful comparative indexes; and 3) a project tracking tool that told which data were in and which were still missing.

Because database builders have to deal with lots of individual bits of handpicked data, you've got to computerize those bits as soon as you get them

5. **Use outside vendors.** Price is a major motivator for switching airports, and we used it as a key ratings measure. We hired GRA Inc., a Philadelphia airline consulting firm, to extract real airfare data from U.S. Department of Transportation nine-track tapes and give it to us in a medium and format we could use easily: Excel files on CD-ROM.

The little-known DOT source database we tapped-Origin and Destination Databank 1b-offers an amazing wealth of information: A 10 percent sample of the 40 million airline tickets actually sold for travel within the U.S. each quarter. We asked GRA to compile an abundance of very specific airfare data, which we ultimately distilled to a "low fare" score indicating which airport in each metro area had the lowest average restricted coach fares to their most popular destinations. This rating was based on more than 17 million tickets sold during the 12 months ended September 30, 1999.

6. **Get more data than you need.** As with all reporting, you don't always know which information will wind up being meaningful. So it's always better to have data and not need them than to need them and not have them.

Not important

For example, we gathered what we thought would be important numbers from the Federal Aviation Administration showing each airport's rate of runway incursions-planes in unauthorized proximity to other aircraft, people, or ground equipment. Incursions were a hot topic at the time, having risen 71 percent since 1993.

But interviews with safety experts and examination of the data showed the risk of actual collision was vanishingly small. There were just 328 incursions out of 68 million takeoffs and landings in 1999. And no airport we examined stood out as having a significantly worse incursion rate than others. So our final ratings did not include this measure. No sense alarming readers with meaningless comparisons.

7. **Use part of an existing database.** We put together on-time flight arrival and departure performance data from a year's worth of monthly Air Travel Consumer Reports published by DOT and available online at www.dot.gov/airconsumer/index1.htm. The ATCR is ostensibly designed to rate airline performance, but some of the data series are also organized by airport, which we used to calculate the percentage of flights that arrived or departed 15 minutes or more late at each airport the previous year.

Site visits

8. **Collect data in-person.** To find out how far travelers must lug their carry-ons to the farthest gates, I traveled to 45 airports and walked more than 35 miles of concourses with a wheeled surveyor's measuring stick. Of course, this tour of U.S. airports was also used to get information on food prices and payphone charges, insight into terminal layout and design, digital video

Continued on page twenty-five

Tipsheets

More information on creating a database is available from the following tipsheets in the IRE Resource Center (*www.ire.org/resourcecenter*):

Tipsheet #911. Shannon Harrington offers help for creating the most useful database from the information you have.

Tipsheet #447. Mike Wendland helps reporters decide whether a spreadsheet or a database best suits a task. Audiotape available through Sound Images, Inc., (303)649-1811, tape #CAR95-12.

To order, call the Resource Center at (573) 882-3364.

Continued from page twenty-four: **performance**

footage, anecdotes, and a firsthand understanding and feel for each airport.

9. Use a questionnaire. To count all the payphones, ATMs, restaurants, and other services, we developed a detailed questionnaire that we faxed or e-mailed to some 75 airport authorities.

10. Normalize and score. The raw numbers by themselves are not always meaningful. Sure, there are 40 restaurants to choose from, but that may not be a sufficient number if they're spread over seven terminals and you're prisoner at the end of one long concourse waiting to hear when your late flight will finally begin boarding. So instead of comparing the absolute number of locations where you can buy prepared food in each airport, we calculated the ratio of food outlets per terminal/concourse area. Similarly, we calculated the number of pay phones per 1,000 daily passengers, ATMs per teminal/concourse area, and available parking spaces per daily passenger whose travel itinerary originated at that airport.

Finally, Finance Editor Louis Richman, Statistics Director Michael Saccucci, and I worked together to translate all of the comparisons into the familiar *Consumer Reports* ratings symbols (known around our Yonkers, N.Y. headquarters as "blobs") to help readers make fast and easy comparisons.

Jeff Blyskal can be reached by email at blysje@consumer.org

Continued from page seven: **contamination**

address, the type of work it did, the dates of its operation, the current use of the property, etc. We also included notes fields to list any document we found that might be relevant to the operations of each company, as well as the location of that document in our paper files. For example, the field for the Simonds Saw and Steel Co., a steel mill in Lockport, N.Y., that cut uranium metal rods used to produce nuclear weapons fuel, would include references to documents in our "Health Studies" file, our "Uranium Processes" file as well as our file on the company itself.

Glorified index

We never actually crunched any data from the spreadsheet, other than a few basic queries to identify the number of firms in a particular state. We just used it as a glorified index. But that proved critical, because we were piecing together company histories from dozens of different types of documents. We might find a contract that named a particular contractor but didn't say what sort of work the company was doing. Another record might tell us what the work was, but it wouldn't include any information on whether the government had studied radiation doses to workers at the site. Another might offer information on typical radiation levels for workers in that type of operation, but it wouldn't list the company specifically as a place doing that work.

In the end, it all had to be pieced together like a maddening puzzle. The spreadsheet simply helped us keep track of all the pieces – and where they fit together – in a much more efficient way.

When it came time to run our list of weapons contractors in the paper, we already had all the information we needed in the database. We simply stripped away the notes fields, wiped out whatever coding we'd entered, and we had all the information we needed in a handy format for our graphics people. And even that wasn't the end of it. After the series ran, when readers began calling to ask where they could get more information on contracting sites in their area, I could call up the database and tell them exactly what records they needed to request from the government. It was only a few weeks before I got a call from an official at the Department of Energy: He wanted a copy of the database.

Pete Eisler can be reached by e-mail at peisler@usatoday.com

Tipsheets

More information on nuclear-weapon plant dangers is available from the following tipsheets:

Tipsheet #1341. "Exposing Hidden Hazards." Get pointers on how to start investigating nuclear industry cleanup from an expert panel at the 2000 IRE National Conference.

Tipsheet #1245. "Newly Discovered Perils in the Nuclear Industry." Sam Roe of the *Toledo (Ohio) Blade* digs into activities of the government and the beryllium industry.

To order, call the Resource Center at (573) 882-3364.

Uplink story ideas

Have you or one of your colleagues recently published a story using CAR that has not been done before or involved particularly difficult data work?

Do you know of a technical problem (or its solution) that others may like to hear about?

Or is there some issue or beat that we haven't covered?

If you have a story idea, we'd like to hear from you. Please contact associate editors Amy Sherrill or Mike Sherry either by e-mail at amys@ire.org or mikes@nicar.org or by phone at (573) 884-7711.

Continued from page fifteen: **justice**

and arraignments. Some of my best information came from chance interviews in the courtroom hallway. Once word got around about my project, many court regulars made a point of pulling me aside and giving me their two cents. This resulted in some good tips about cases that illustrated my findings. It also confirmed for me that the results of my data analysis were on target.

Internet sources included the Bureau of Justice Statistics at *http://www.ojp.usdoj.gov/bjs/*. This site had good reports on drug arrests and convictions. The White House Office of National Drug Control Policy at *http://www.whitehousedrugpolicy.gov/*. This site had good links to other drug-related sites.

Lessons learned

Since court computer records in Massachusetts are limited, when I began my investigation I was forced to make my own database. Being a complete novice with computer-assisted reporting, I know now I went about this the wrong way. Instead of thinking up front about how I might want to analyze my data, and hence how to organize it, I just jumped in with collecting information and inputting it into Excel. This meant that I had to go back into my database after the fact and spend days cleaning it up. This included simplifying columns by doing things like parsing out addresses and charges.

Laptop problem

Also, I started using the newspaper's old Dell laptop. This broke down part way through the data collection and since the newspaper did not have another one, I was forced to use my own Macintosh Powerbook. This meant running Access through Virtual Windows, which added a layer of complications to my data analysis. Were I to do something like this again, I would think through equipment questions up front.

Polly Saltonstall, formerly of the Standard Times, can be reached by e-mail at PSalty@aol.com

Continued from page eleven: **offense**

down to a reduced charge? Which judge handed out the heaviest sentences? The lightest? How many of the defendants were children themselves? How many charges each defendant faced? Since there were only 207 defendants in 435 cases, it's obvious most were repeat offenders.

Searching for reasons

On to the reporting: I talked to social researchers about why most of the offenses occurred in the area of the county with the lowest income. To prison officials and counselors about how they (don't) provide therapy to offenders while they're in prison, and their prognosis for change. To prosecutors and defense attorneys about why the overwhelming number of cases were pleaded down and most offenders serve only probation.

I spoke to social experts about why children rape other children. To police about how hard it is to help children tell strangers their deepest, darkest secret while the offender is watching from just 15 feet away across the courtroom. To prosecutors and police about why many cases are reported, but never prosecuted.

I interviewed the mother of a pair of young boys repeatedly molested by a trusted neighbor.

And I spoke to police and counselors about what parents should really be looking for. It's not strangers you have to worry about. In 435 cases in 6-1/2 years, only three strangers were prosecuted for sex assaults on children.

The result was a three-day series with nine or 10 nine stories, and additional graphics, including a ZIP code map indicating where the assaults occurred.

Susan L. Oppat can be reached by e-mail at soppat@aa-news.com

Continued from page two: # insecurity

The violation had to have occurred since 1989, and we're only interested in cases in which the violators are airports or airlines. The main table can help with both of those. One of its fields shows the dates of violations. Another, called CERT_CODE, describes the certificate type held by the violator.

So we made a smaller main table with just the cases we're interested in:

```
SELECT main.* into smalltable
FROM main, cases
WHERE main.iid = cases.iid and date_viol
> #12/31/89# and (cert_code = '00' or
cert_code = '34')
```

Now, we have "smalltable," which includes security cases filtered by security code and further filtered by date and by type of violator.

For many passenger violations, those cases were reported by state or local police agencies.

But we're still just halfway there. We have a large table – of course, smaller than the one we started out with, thanks to filtering – but what can we do with it?

Comparing airports was a common theme, so let's have our newsroom do the same. For a basic query to count security problems by airport:

```
SELECT apt_id, count(*)
FROM smalltable
GROUP by apt_id
ORDER by 2 desc
```

More to think about

That gave a simple count at airports and violations, with the violations blamed on airports or airlines (airlines, of course, are in charge of screening of passengers). But even though the query filtered out violations by passengers, reporters must consider passengers. Why? Because big airports almost always have more violations than smaller ones, and the size depends on the number of passengers.

Reporters turned to enplanement data, showing the number of passengers in a year, by airport. The latest set of numbers, in Excel format, can be downloaded at *http://www.ire.org/related/airpassenger.html.* A search on the FAA's Web site should turn up additional historic data.

Our fictional newsroom took a semi-scientific approach, choosing airports within 50,000 passengers a year to compare.

Violation details

Our newsroom looked at specific information on our local airport's problems, trying to pick apart, with basic shoe-leather reporting, specific cases. The database can help with that, too. It shows when and where a violation occurred, plus at least a short description of what happened. There's a field called SRC_DESCR in the main table that shows how the violation was reported. For many passenger violations, those cases were reported by state or local police agencies. For the others, though, often the field shows designations of inspection or surveillance when airport screeners fail to detect weapons.

Newspapers and television stations, from one coast to the other, reported about security at their local airports. Often, they used other data sources or paper records to flesh out their stories – for example, John Perry of *The Daily Oklahoman* (Oklahoma City, Okla.) used local police reports to confirm much of what he found in the data and find narrative detail.

Other databases

The FAA enforcement database isn't the only potential data resource available through IRE and NICAR:

- Using the FAA's airmen directory – a listing of most pilots certified by the U.S. government – reporters found basic information about some of the hijackers.
- The *Dallas Business Journal* used a federal contracts database to report on the "568 North Texas companies that did business with the military last year … .Now, some of those companies are preparing for their Pentagon contracts to grow as war looms…."
- While airlines are seeking a federal bailout, reporters are considering stories about linking that industry to campaign contributions, using Federal Election Commission data.

Jeff Porter can be reached by e-mail at jeff@ire.org

Resources

For even more detailed advice – or to re-read some of the debates on how to use the data – IRE members can search the archive of NICAR-L messages at *http://www.ire.org/membership/listservs/nicar-l.html.*

FAA enplanement data is available at *http://www.ire.org/related/airpassenger.html.*

Bits, Bytes and Barks

NICAR national conference

Because of the Sept. 11 terrorist attacks on the East Coast, the IRE Board postponed the Philadelphia conference until March 14-17, 2002.

Much of the program will be unchanged and the conference will remain in Philadelphia (See details below). However, we are adding new sessions to discuss coverage of the attacks.

The conference site is now the Double Tree Hotel (Broad Street at Locust), in the heart of Philly's theater district.

For hotel reservations call: (215)-893-1600. Please ask for the Investigative Reporters and Editors room block.

Reservation cut-off is Friday, Feb. 8. Room Cost is $129 single/double, $139 triple/quad.

For more information call IRE at (573) 882-2042 or go to www.ire.org/training/philly/

DC regional conference

Plan to attend one of the most practical training events in Washington, D.C.

The 2002 regional conference, to be held on January 26, will address hot topics like the U.S. Census, terrorism, aviation, health, and education.

You can also learn about important Web sites and crucial Internet search techniques. There will also be an optional CAR day on January 27 at the nearby Medill School of Journalism Washington bureau.

The conference hosts are The National Press Club, Investigative Reporters and Editors, and the National Institute for Computer-Assisted Reporting.

For more information on the conference, including registration information, go to *www.ire.org/training/dc02*

Data update

The database library recently updated the NASA Air Safety Reporting System database. It is current to February 2001.

The ASRS database consists of anonymous reports about aviation safety. Anyone is eligible to file an ASRS report, including air traffic controllers, pilots, flight attendants and passengers. The database has information on more than 98,000 events back to 1988.

The cost of the database is: $65 for small news organizations, $130 for medium news organizations and $155 for large news organizations.

For more information about the ASRS database or other NICAR databases, go to www.ire.org/datalibrary/databases. To order, call (573) 884-7711 or (573) 884-7332.

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

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