Uplink

May • June 2002
Volume 14, Number 3

Published bimonthly by the National Institute for Computer-Assisted Reporting

www.nicar.org

Bits & Bytes

Data additions:

The IRE and NICAR Database Library completed two additions of its government database collection: the Department of Transportation's database of truck inspections and data files from the Census Bureau's Federal Assistance Award Data System (FAADS).

The truck inspections database includes seven relational tables, covering millions of inspections and violations across the United States. The library offers the data nationwide or by state.

The FAADS files include tables for each fiscal year, covering grants and other financial assistance, including the recipient and location, plus a description of the program. FAADS files, too, are offered both nationwide or by state.

Data updates:

The IRE and NICAR Database Library recently updated the following databases:

Boating accidents

The U.S. Recreational Boat Accident Database includes 7,740 accidents in 2000, obtained in April from the U.S. Coast Guard. The database consists

continued on page 2

TRANSPORTATION

Databases highlight tired truckers

By Greg Reeves, *The Kansas City Star*

Investigating fatigue among over-the-road truckers last year was frustrating work: we had to analyze not one, but several databases, all of them missing crashes and key pieces of information. But when we were done, special projects reporter Judy Thomas and I were able to piece together enough solid information to show the fatal consequences of truckers driving tired.

In "Dead Tired," a three-day series that ran in *The Kansas City Star* starting Dec. 16, we reported that the long hours truckers log on the road contributed to hundreds of deaths every year.

continued on page 26

SPOTLIGHT: NICAR IN PHILADELPHIA

A conference report

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

More than 400 journalists passionate about computer-assisted reporting in March gathered in Philadelphia to trade tips, brush up on their skills and network. With newsrooms across the country slashing training budgets, many of the journalists attending IRE and NICAR's annual computer-assisted reporting conference paid their own way, and invested in their own careers.

Journalists packed 55 hands-on classes, learning everything from basic spreadsheet skills to statistical analysis and advanced programming tricks. And they had the chance to sample from more than five dozen presentations, panels and roundtable discussions aimed at broadcast, print and online journalists.

John Allen Paulos, the Temple University mathematician and author of "A Mathematician Reads the Newspaper," kicked off the four-day conference on March 14 with a spoken tour of his book, outlining the numeric pitfalls that journalists have fallen into.

continued on page 24

SPOTLIGHT:

FOR MORE ON NICAR IN PHILADELPHIA SEE:

- Census data for stories, p. 6
- Intranets for projects, p. 8
- Measuring the local economy, p. 10
- Tracking donors in city hall, p. 17

Bits & Bytes

continued from page 1

of four relational tables containing data on each accident. Details for each accident include when and where the accident occurred, its causes and details of deaths or injuries. Those purchasing the database from the Database Library will also receive data files covering 1969 through 1999.

AIDS

The AIDS Public Information Data Set contains information about AIDS cases reported to state and local health departments since 1981. These agencies share this information with the Centers for Disease Control. The data set includes each AIDS patient and summary information by state, metropolitan area, mode of exposure to HIV, sex, race/ethnicity, age, vital status and case definition category.

SDR

Service Difficulty Reports consist of maintenance incidents collected by the Federal Aviation Administration for the purpose of tracking repair problems with private, commercial and military aircraft and aircraft components. The data is reported by tail number and aircraft serial number, so it is possible to trace the maintenance history of a particular airplane with this database.

Contact information

To find out more about these or other databases, visit the Web page www.ire.org/datalibrary/ or call 573-884-7711. To suggest articles for *Uplink*, send your e-mail to uplink@nicar.org.

CAR projects garner top honors

Sari Horwitz, Scott Higham and Sarah Cohen of *The Washington Post* won a prestigious Investigative Reporters and Editors medal for exploring the deaths of children in the District of Columbia. For the same series, the *Post* team won the 2002 Pulitzer Prize for investigative reporting.

A student IRE Award went to University of Missouri School of Journalism graduate Mary Jo Sylwester, now at the Center for Public Integrity.

Both Sylwester and *Post* database editor Cohen have worked with IRE and NICAR. Cohen is a former training director for IRE and NICAR. Sylwester, as a graduate student, served as IRE and NICAR data library administrator. She's now the database editor for the Center for Public Integrity.

The *Post* team discovered that 229 children died during a seven-year period after their dangerous family situations came to the attention of the district's child protection system.

> **"The [*Post*] project, which has resulted in wide-ranging reforms, answers the highest call of investigative journalism."**

Despite strict confidentiality laws, the team pieced together records for 180 of those deaths, compiled them into a database and found that one in five – mostly infants and toddlers – lost their lives after government workers failed to take key preventive action or placed the children in unsafe homes or institutions.

"The project, which has resulted in wide-ranging reforms, answers the highest call of investigative journalism," the IRE Award judges said in granting the medal.

Sylwester won the student award for "A Flood of Problems." Her package for the *Columbia Missourian* exposed problems with a local storm water permitting system, the consequences for residents and why readers should care.

The May-June 2002 edition of *The IRE Journal* includes a complete listing of winners.

The annual awards of IRE – a 4,200 member professional organization – recognize outstanding investigative work in print, broadcast, online media and for work furthering freedom of information.

The awards will be presented during a June 1 luncheon at the IRE Annual Conference in San Francisco. The conference, scheduled for May 30-June 2 at the Hyatt Regency San Francisco, will feature many of the winners speaking about the techniques they used to develop their stories.

Contest entries are screened and judged by other working journalists. The final judges this year included two IRE board members and several IRE members selected by the membership and the board president.

Copies of all contest entries are available from the IRE Resource Center (*www.ire.org/resourcecenter*). The center can be reached via e-mail at: *rescenter@ire.org* or by calling 573-882-3364.

CONSUMER REPORTING

Databases add depth to pool stories

By Mike Sherry, *IRE and NICAR*

It's a good bet your news organization will soon be doing the quick-hit pool opening story. Why not add depth to an otherwise shallow story?

You can do that – and many other stories – with U.S. Consumer Product Safety Commission (CPSC) death and injury data. The data should be manageable for anybody with a basic knowledge of software such as Microsoft Access, and you can get a decade's worth for free.

Along with swimming pools, the data include incidents involving thousands of other consumer products – from air conditioners to vacuum cleaners.

Pool problems

Many of your readers, viewers or listeners might not realize how dangerous swimming can be.

Of the more than 44,000 deaths included in CPSC data running from January 1990 to October 2001, more than 5,300 records cited swimming pools as the main product involved with the death. Other oft-cited products include bathtubs and showers and stairs and steps.

Swimming deaths increase when slides and other related products are taken into account. One entry in the database – which is compiled from state health department records – details the September 1998 death of a 3-year-old girl in Tampa, Fla. She died after tumbling from a pool slide, hitting her head and falling into the water. The table's four narrative fields provide a detailed account of the incident.

CPSC's codes even get as specific as categorizing deaths related to pool chemicals (product code 938). The death database includes eight instances – in states as diverse as California, New York and Wisconsin – where pool chemicals contributed to a death. In one incident, a 66-year-old Cocoa Beach, Fla., woman died after inhaling chlorine fumes.

From January 1990 to October 2001, more than 5,300 records cited swimming pools as the main product involved with the death.

More than death data

The death database is just one of four CPSC provides with data compiled in the 1990s. (Data going further back is available, but it is likely to cost you and take longer for CPSC to provide). Also available are:

• National Electronic Injury Surveillance System (NEISS). Record Count: 3,228,474. Conducted for nearly 30 years, the survey currently collects reports from 100 hospitals nationwide. This survey is considered a representative sample of emergency rooms, allowing the CPSC to estimate the number of injuries caused each year nationwide by products. The commission makes these estimates available – in paper format – through its annual "Product Summary Reports." For example, the 2000 summary estimates there were 80,110 swimming pool-related injuries that year. The 2001 summary report should be available from the commission in June or July.

CPSC periodically alters its sample to account for hospital closures and other similar changes. The last change was in 1997. Yet, CPSC statistician Tom Schroeder said the commission's annual estimates could be added together for a historical account of injuries caused by a particular product. For instance, it would be OK to add up the CPSC estimates for swimming pool injuries for each of 10 years and report the number of injuries for the entire decade.

One caution, however; CPSC explicitly warns against saying these products "caused" the injuries. The commission says emergency room information only tells the products that were "related" to the injury.

• Injury/Potential Injury Incident File (IPII). Record Count: 222,396. This file contains summaries of reports to the agency's hotline, newspaper accounts, reports from medical examiners and letters to CPSC. Helpful fields in the database include whether the injury was work-related, a general description of the cause (such as an explosion or fall), and the outcome of the hospital visit (such as if the person was treated and released, or dead on arrival).

• In-depth Investigations (INDP). Record Count: 46,420. This file contains summaries of investigations into product-related injuries or incidents. According to the CPSC, these investigations are prompted by information gathered for the NEISS or incident tables or through a separate initiative by the commission. Helpful fields in this database include the body part injured and injury diagnosis (such as poisoning or drowning).

Like the death data, the narrative fields are one of the best features of these three datasets. Additionally, the IPII and INDP tables include the city and state where the incident occurred. Given that the NEISS is just a sample, it does not include state or city fields.

The incident and NEISS tables can be linked with the investigation table

continued on page 4

consumer

continued from page 3

through a unique identifier known as the task number.

Data documentation

Some of the key documents for these datasets are readily available on the Web. Those include an overview of the NEISS database (*www.cpsc.gov/cpscpub/pubs/3002.html*) and information about obtaining the databases (*www.cpsc.gov/about/clrnghse.html*).

Also available is the very useful "Product Code Comparability Table" (*www.cpsc.gov/Neiss/jove-asp/default.htm*), which documents the changes in product codes throughout the years. For instance, from 1972-1978 the CPSC used code "1220" for swimming and water sports equipment. But, from 1978 on, it split that code into three more-specific codes – such as "3276" for water polo.

However, the record layouts sent with the data were not great.

Problems included field names that were out of order or were not in the database at all. When Investigative Reporters and Editors and the National Institute for Computer-Assisted Reporting obtained the data, we had to call the CPSC for a code sheet called "Interpretive Sheet for Computer Printouts of Reported Incidents." You should make sure you request this with the data, since it contains codes that help decipher some of the fields.

Finally, the CPSC agreed to resend us the data with updated record layouts and code sheets.

The CPSC originally sent the databases in a text format (delimited by "|") that imported easily into Access. If you ask, you should be able to get the data as an Access database or as a .dbf. The NEISS dataset came as four separate text files; the other three databases came as three individual text files.

Data come with caveats

When dealing with the age fields, keep in mind that CPSC uses a three-digit convention for children under the age of 2. For example, a child that is 23 months old is listed in the data as "223." Obviously, you'll want to keep this mind if you're calculating average or median ages, since the age cited in the example would be read as a person who is 223 years old.

Each of the four databases includes multiple product fields. For instance, the NEISS has "Prod1," "Prod2" and "Prod3" fields. This *does not* mean that Prod1 was the main product involved with the injury and that the other two were involved to a lesser extent. The fields just tell which products were involved with the injury or death.

Mike Sherry can be reached by e-mail at mikes@nicar.org

readme.txt

For more information

The IRE Journal May-June 2002 edition offers detailed articles about consumer affairs reporting.

Resources

Tipsheets and stories from the IRE Resource Center can assist journalists working on similar stories.

Story #9182
Inside Edition (New York) shows how the swimming pool industry's own minimum standards do not prevent paralyzing injuries when an adult dives from a diving board into a backyard swimming pool.

Tipsheet #1122
This tipsheet, by Arnold Diaz of ABC News *20/20*, lists sources and Internet addresses for consumer groups, medical sites, journalists' tools, and consumer law and government agencies.

To order, call 573-882-3364 or go to ***www.ire.org/resourcecenter*** and search with the keywords "consumer affairs."

Ordering data
The U.S. Consumer Product Safety Commission (CPSC) databases are available from:

U.S. Consumer Product Safety Commission
Clearinghouse, Room 504
Washington, D.C. 20207
Telephone: 301-504-0424
Fax: 301-504-0025

Web address
You can search the CPSC's National Electronic Injury Surveillance System (NEISS) database at ***www.cpsc.gov/neiss/default.html***

Magazine Writers/Chicago presents:
July 26 to July 28, 2002
WRITERS & EDITORS ONE-ON-ONE

Learn what editors from: Self, Reader's Digest, Cooking Light, National Geographic, Meredith Integrated Marketing, Continental Airlines, Profit and Modern Maturity want from experienced freelance writers.

Attendees in 2001 made more than $93,300 as a result of the conference.

More details and applications are available at:
www.magazinewriters.com

This line-up, like everything in life, is subject to change.

CONSUMER REPORTING

Resources for consumer-related stories

By Ian Demsky, *IRE and NICAR*

IRE and NICAR offer several databases, stories and tipsheets for reporters investigating consumer-related issues.

The Manufacturer and User Facility Device Experience (MAUDE) database from the Food and Drug Administration contains information about problems that occurred with medical devices. The problems include injuries, deaths or product malfunctions.

The data include detailed narratives about the device failures. Specific devices can be searched to determine how often they're reported to have failed. Patient-specific and hospital-specific information has been removed for privacy reasons.

Several stories have been done using the data. For example, a 2000 *Chicago Tribune* investigation discovered that because of understaffing and a lack of training, nurses were killing and injuring patients every year. In 1992, *U.S. News and World Report* examined medical devices that concerned the FDA, including breast, penile and testicular implants as well as heart-bypass pumps and cranial stimulators.

More health-care data

The National Practitioner Data Bank from the Department of Health and Human Services has information about malpractice payments, licensure, clinical privileges, professional society membership and adverse action reports concerning physicians, dentists and other licensed health-care professionals.

Practitioner- and hospital-specific information is excluded from the file.

Despite that, the *Hartford Courant* used the data and a variety of public records to examine several doctors who continued to practice medicine despite being sued for malpractice. The *Courant* was able to focus on several of the most dangerous doctors.

The *New York Daily News* revealed 15 of New York's most sued doctors had between them 412 malpractice suits, and their insurance firms had spent more than $26 million in compensation payments.

Vehicle complaints, recalls

The vehicle recalls and complaints database from the National Highway Transportation Safety Administration contains information about vehicles that have been recalled by manufacturers and consumer complaints about them dating back to 1966.

Dateline used the recalls data to launch an investigation into "sudden acceleration."

About 75 percent of recalls are initiated by the manufacturers, because of problems – identified or perceived — with a vehicle. For the rest, NTSHA contacts the manufacturer after receiving complaints, although the agency itself never issues recalls on its own.

Examples of the use of this database include stories done by NBC News Dateline and the *Los Angeles Times*. Dateline used the data to launch an investigation into "sudden acceleration," finding evidence that electrical shorts, rather than people accidentally stepping on the wrong pedal, were to blame. The *Times* looked at data concerning safety hazards that contributed to the deaths of rear-seat passengers in collisions.

For more information about each database, go to *www.ire.org/datalibrary*, or call 573-884-7711. To obtain copies of stories or tipsheets, go to *www.ire.org/resourcecenter* or call 573-882-3364.

Ian Demsky can be reached by e-mail at iand@nicar.org

readme.txt

Stories

The stories mentioned in this edition of *Uplink* will soon be available from the IRE Resource Center. Search the story database at ***www.ire.org/resourcecenter*** for order numbers. Call 573-882-3364 to order these and other investigative stories.

Databases

To view the list of databases available at the IRE and NICAR Data Library, view the Web site ***www.ire.org/datalibrary***. From this page, you can also download an order form. The library's telephone number is 573-884-7711.

SPOTLIGHT: NICAR IN PHILADELPHIA

Using census data for non-census stories

By Paul Overberg, *USA Today*

If you're looking for local school stories, it's easy to be put off by this table:

P36. SEX BY SCHOOL ENROLLMENT BY LEVEL OF SCHOOL BY TYPE OF SCHOOL FOR THE POPULATION 3 YEARS AND OVER [47]

Universe: Population 3 years and over
Total:
Male:
Enrolled in nursery school, preschool:
Public school
Private school
Enrolled in kindergarten:
Public school
Private school
Enrolled in grade 1 to grade 4:
Public school
Private school
Enrolled in grade 5 to grade 8:
Public school
Private school
Enrolled in grade 9 to grade 12:
Public school
Private school
Enrolled in college, undergraduate years:
Public school
Private school
Enrolled in graduate or professional school:
Public school
Private school
Not enrolled in school
Female:
(Repeat)

This data will be published this summer for areas as big as states and as small as block groups (a few hundred homes) in Census 2000 Summary File 3 data files. While it's not a good idea to analyze data for block groups – the data's too thin – the groups let you build custom areas. Usually census tracts, areas that average 1,500 homes, are the smallest standard areas worth analyzing. The school enrollment data list is a good example of how census data can yield lots of non-census stories – stories that aren't just about basic demographic trends like changes in population and racial totals. The exercise below focuses on just this table, but you can adapt the approach to any of them.

Doing your homework

There are 812 tables in SF3 (www.census.gov/population/www/cen2000/SF3-pop.html and www.census.gov/population/www/cen2000/SF3-housing.html). Look at tables for subjects that you care about and think: "People asked to see this data this way. What are they looking for?"

Think geographically as well to use the rich local detail only the census offers.

Invert your usual process. Instead of looking at the data for stories, bring your knowledge of a subject – like your community's schools – to this data. See if this data can back up or knock down hunches and ideas. Look at its variations across your area. See if this data, linked to local data on tax rates or test scores or charter schools or attendance zones or home prices, produces new insight. You should also look at what's *not* there, the way you'd analyze a story draft for holes.

It's fair to ask, "Why do I need census enrollment data when I get it each year from the school district?"

• Reason 1: No school superintendent has all the data. Nationwide, one student in eight attends private schools. And then there's home schooling. Even states with strong rules don't always require registration with the local district.

• Reason 2: School systems collect data the way they need it – by school. That doesn't help if you want to see it by neighborhood. Or within 1.5 miles of a school building.

• Reason 3: Annual data may be too much. Census data from 1990 and 2000 lets you see broad patterns, and in ways that are comparable around the state and nation.

Preparing for possibilities

It's easy to see basic story possibilities in the table above, like public-private and male-female breakdowns in local schools. But think geographically as well to use the rich local detail only the census offers. If you map these ratios for each neighborhood, you will see variations that could produce stories.

One way to get ready for the new data is to get local data from the comparable table in the 1990 Census, table P54 in STF3:

P54. SCHOOL ENROLLMENT AND TYPE OF SCHOOL [7]
Universe: Population 3 years and over
Enrolled in preprimary school:
Public school
Private school
Enrolled in elementary or high school:
Public school
Private school
Enrolled in college:
Public school
Private school
Not enrolled in school

You can probably get local P54 data from planners, school officials or the state data center. You can also download it. Browse to *http://factfinder.census.gov*, click on "Data sets" and follow the instructions.

The 1990 table is less detailed, but you

can collapse 2000 categories to match.

When 2000 data arrives

• Who's not there: If you subtract an area's population age 18 and over and ages 3 to 5 from the sum of the two "not enrolled" categories, you get kids ages 5 to 18 who aren't in any school. How many? Are they home-schooled? Nationwide, about 1.7 percent of kids 5 to 17 were home-schooled in 1999 (*http://nces.ed.gov/pubs2001/HomeSchool/*) but that rate varies widely. Dropouts? (*http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2002114*) Where are they – clustered or scattered? If you want to focus on dropouts, check another table, P38. Set up mostly to track potential military recruits, it nonetheless tells you how many people ages 16 to 19 are dropouts and their employment status.

• Private school popularity: What's the ratio of public vs. private in grades 1-4 compared to grades 5-8 and grades 9-12? One index of a public school system in trouble is people leaving as their kids get older. In particular, calculate the public-private ratio for each neighborhood and map it, even if it's just using crayons on paper. See any surprises? How well do you know your private schools? At the National Center for Education Statistics web page *http://nces.ed.gov/globallocator*, check "private schools," punch in a local ZIP and a 20-mile radius. Use the links to find enrollment data.

• Starting earlier: Over the last 10 years, nursery schools have swelled. Compare 1990 and 2000 totals and public school-private school market share splits in each year. You might find overall increases even though birth rates in most areas dropped from 1990 to 1995, cutting the number of 4- and 5-year-olds from 1995 to 2000.

• Middle-age students: Look how many are enrolled in graduate or professional programs. You might be surprised — distance education, weekend and night programs are boosting middle-age enrollments. If you want to delve more into the age breakdown of local students, see tables PCT23 and PCT24 (*www.census.gov/population/www/cen2000/SF3-pop.html#seea3*).

• College town: You may know the enrollment of the local college, but where are the students? This table will tell you. You probably know the obvious undergrad neighborhoods near campus, but look at the change since 1990, both in density and extent. Grad students, who tend to be less raucous and more settled, will probably show a different pattern.

• Commuter colleges: Your town has colleges you probably don't think much about – community colleges, technical and trade schools, etc. Look to see where the most college students live *away* from known college neighborhoods. They're probably working-class and immigrant neighborhoods with lots of hopes and dreams and not much money. Where are those students enrolled?

• Magnet and charter schools: They're growing in popularity and steadily scrambling the neighborhood school picture. It's quite possible to find suburban neighborhoods where kids go to five different elementary schools. What does this do to neighborhood life?

• Race and enrollment: How different is the racial profile of your district from the public school enrollment? Or a school's attendance zone from its actual student body? NCES has summed Census 2000 race/ethnicity data for 8 million census blocks to get totals for 14,400 districts nationwide. The data is tallied for the total population and the population 18 and over, so you can subtract to get kids up to 17. Start at *http://nces.ed.gov/surveys/sdds/c2000d.asp* to get a quick profile or to download data. The 1990s saw school segregation grow in most places. (See *www.law.harvard.edu/civilrights/publications/resegregation01/pressseg.html*.)

• Striving: Take the private school enrollment percentage for every neighborhood you cover from this table. Take the poverty rate percentage from P87 or P90. Put them in Excel and make a scatter-plot graph, so that each pair is plotted as a dot. You'll see a pretty logical pattern – higher private school enrollment corresponds with lower poverty rates. But you'll see some dots standing out from the cluster – where a relatively poor neighborhood has a relatively large private school enrollment. What's going on there?

Paul Overberg can be reached by e-mail at poverberg@usatoday.com

readme.txt

Cautionary notes about SF3 data

1. This data comes from the census long form, which about one household in six fills out. If the school enrollment question isn't answered despite the best efforts of census-takers, the Census Bureau will use rules it has tested to fill in the database on known characteristics of the household and its neighborhood. To see what percentage of any area you're studying has this imputed data for this question, see table P110.

2. Because this data comes from a survey, each number has a little cloud of uncertainty around it, just like an opinion poll. We usually call these "error margins." The smaller the geographical area you're studying, the bigger the cloud is relative to the totals. Make sure you're not overstating what the data says by understanding that error margin. Use the SF3 documentation to compute the applicable margin, or get help from a local statistics professor.

SPOTLIGHT: NICAR IN PHILADELPHIA

Intranet tools and preparing for disaster

By Tom Torok, *The New York Times*

The New York Times taxed traditional reporting tools to the limit shortly after the first plane slammed into the World Trade Center.

Within 24 hours, more than 600 files of story notes accumulated in the newsroom's Atex computer system metro queue. There were more related files in the foreign, national and business queues. Searching for any snippet of information became increasingly difficult and slow as torrents of information poured in.

Some reporters tried other means of organization. One started a spreadsheet to track suspects that soon grew to monstrous proportions. The spreadsheet was difficult to read and almost impossible to print. Attempts to share the spreadsheet over the network compounded the unwieldiness.

A terrorism team was assembled and editors had trouble tracking more than 80 reporters and their assignments, phone numbers and e-mail addresses.

I didn't know it at the time, but I had an intranet ready for a story of the magnitude of Sept. 11.

Managing the information

A web application to search the accumulating notes was ready the next morning.

A virtual online newsroom began to take shape by the second afternoon, with applications that allowed lightning-fast searches of accumulated notes, suspects, reporters, phone numbers — anything that involved Sept. 11. Reporters and editors could add or update information whenever they wanted, from wherever they wanted.

The intranet within an intranet wasn't always the perfect tool – no such tool exists for what had happened — but it did help significantly in managing information.

In all, about 20 pages were created to deal with the coverage. But with the exception of one of those 20 pages, I either already had similar pages that could be modified to suit the task or had a program ready to write the code needed.

Every "borrowed" Web page needed tinkering – some needed extensive rewrites – but anything I had on hand was better than starting from scratch. And what I had on hand was a virtual newsroom waiting for a project.

Many of the pages were written for coverage of the 2000 national party conventions, where they were never used. The Sept. 11 pages were better received than the convention pages. In the first two months, the newsroom was calling up those related pages between 2,200 and 2,600 times a day.

Pages for preparation

Rather than enumerate which pages the *Times* used for Sept. 11 coverage, let's concentrate on pages that you should have ready for the next breaking story or project, be it big or small. To start, you want to prepare four pages to display information:

1. A news-brief or announcements page on which information important to the project can be posted.

2. A whereabouts page that will allow anyone involved in the project to find reporters, editors and assignments quickly.

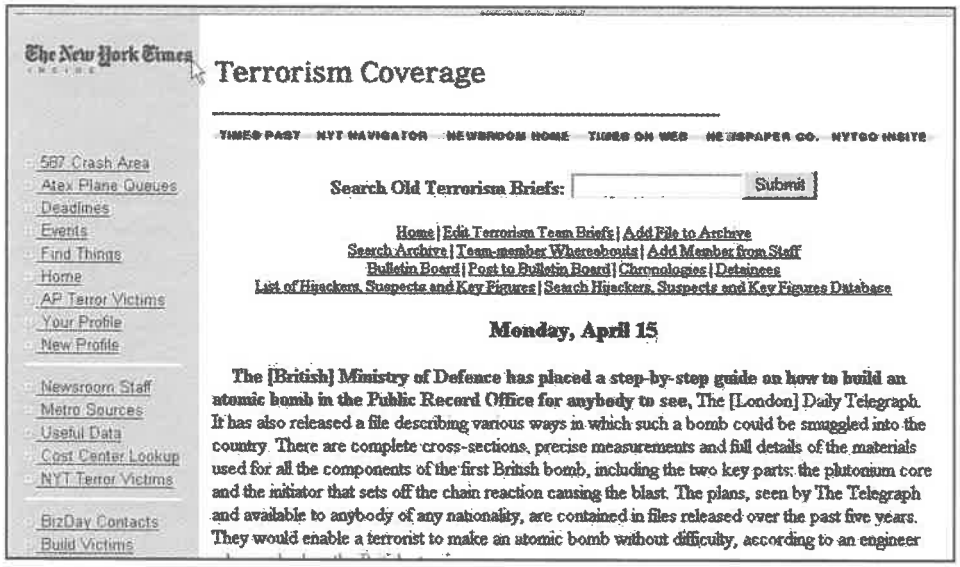
3. A text-search page to search collected text or HTML documents related to the project.

4. An events page to track events related the project, such as press conferences or meetings.

Before you create a monster, ask yourself some questions about these pages.

Who will keep them current? Who can add breaking information? Who can delete information? Who can edit embarrassing typos? If all answers point to you, you're heading for trouble before you start.

Take a look around the newsroom. A good look. Is there *anyone* who can edit a Web page while you're sleeping, on



vacation or on deadline? Probably not.

As important as you are to your newsroom's computer-assisted reporting efforts, you are not indispensable. In fact, you will become a bottleneck if your intranet cannot be brought up to date without you. Talented writers and news gatherers surround you. Why not marshal their talents to help? Probably all but two people in your newsroom can type into a Web input box. (The other two can't do anything with computers. Accept that and get on with life.)

Code-writing programs also are invaluable for making interactive forms for databases

To bring more hands into the process, try following a few guidelines that will help keep your intranet current:

• Guideline 1. Do not post anything on your intranet unless there is some mechanism (or persons in addition to yourself) in place to keep the information current.

• Guideline 2. Whenever possible, construct information pages so they are fully and easily editable by any persons you choose at anytime they choose.

We're taking it as a given that no one else in your newsroom knows HTML, so we've got to add a few more guidelines to make 1 and 2 practical:

• Guideline 3. Use databases to drive virtually all of your content that is not archived text.

• Guideline 4. Use interactive templates to display the contents of the databases.

With these guidelines in mind, you'll need one more page in your basic package to allow selected users to upload new text files for your text search page. You'll also need at least three more pages to make the announcements, whereabouts and events pages fully editable on the Web. The interactive templates for the latter tasks are complicated, tedious to build and difficult to debug. In that light, consider another guideline:

• Guideline 5. Use one or more programs to create the code to display database content and/or to make database content fully editable on the Web.

Such programs would be useful in creating your four basic pages and the pages that are used to edit them. (A quick way to allow others to upload text via the Web is a program called SA-FileUp which you can sample at *www.softartisans.com*). Code-writing programs also are invaluable for making interactive forms for databases that you are creating on the fly, as we did at the *Times* for victims, suspects, detainees and chronologies.

Choosing the software

There are a variety of excellent ways to serve up the pages and the data. Software and hardware have evolved to the point where communications between Web pages and databases are so robust that even pages with multiple complicated queries will work *every* time.

If at all possible, I would avoid using a desktop application, such as Microsoft Access or FoxPro, as database servers. They'll work in low-volume situations, but they will not match the performance and ease of administration of a dedicated database server, such as Microsoft SQL Server, Oracle or MySQL (which is free).

Many professionals at high-volume Web sites prefer Web servers that operate in Unix/Linux-like environments, but most newsrooms are using either Cold Fusion or Microsoft Internet Information Server (or both) for their intranets because of relative ease of use in a PC environment. Vastly improved versions of FileMaker Pro also are gaining popularity in both PC and Macintosh environments.

Cold Fusion and FileMaker Pro have the shortest learning curves. Cold Fusion has been used to create excellent interactive sites at *The Seattle Times*, the *St. Paul Pioneer Press* and elsewhere. Server versions start at about $1,300. FileMaker starts at about $1,000 for server versions.

If you have Windows 2000 Professional, you have enough free software to start a small intranet (10 concurrent users) with text-searching capabilities.

The learning curve can be quite steep to use Active Server Pages, the programmable pages used with Microsoft's Web server. Some journalists with little or no programming skills have had success creating interactive pages with Microsoft's FrontPage, which is packaged with premium versions of Microsoft Office.

I've written Web-based applications that will create Active Server Pages for database searching and database editing (updates, deletes, additions as well as searches). I've also modified Microsoft's sample text-searching pages so they can be quickly deployed in your newsroom. (These are available at: *www.nicar.org/techtips.html*.)

Whichever platform you choose to host your intranet, you should immediately create a package of the pages recommended here and then, make two copies. While making the copies, take extensive notes in a text document that details which changes are needed to create new packages. That way, you'll be ready for two new projects immediately and you will be able to deal with any new projects that follow.

Tom Torok can be reached by email at tomtorok@nytimes.com

SPOTLIGHT: NICAR IN PHILADELPHIA

Measuring your local economy

By John Byczkowski, *The Cincinnati Enquirer*

Your editor asks, "How's the local economy?" What your editor probably means is "Is the local economy expanding, standing still or contracting? Is business in a mood to expand? Are consumers in a mood to spend?"

You'll find that there's actually a fair amount of data out there about local economic activity. The challenge is whittling it down to a manageable number of key indicators that you can obtain reliably, and then track regularly.

These nine indicators should be obtainable on a monthly basis for just about anywhere in the country:

Employment — total non-agricultural wage and salary, manufacturing, new unemployment claims

Construction — new home starts (number of starts, rather than dollar value), nonresidential construction (dollar value of building permits)

Spending — retail sales

Production — average weekly hours, manufacturing, industrial electricity sales

Getting the data

• Employment data: The pace of job growth is probably the best economic-health indicator, not to mention consumer confidence. Consumers don't spend if they don't feel secure in their jobs. Start at *www.bls.gov/data/home.htm*. This is the gateway to the huge stores of data at the Bureau of Labor Statistics, U.S. Department of Labor. You'll be able to get data – in some cases down to the county level – stretching back as far as the 1940s. Focus on this section: "State and Area Employment, Hours, and Earnings." You'll find numbers on total employment, manufacturing employment and average work week. That will help you get the historical databases set up, but new data – as it's issued – will come from your state's agency that collects labor market information. Ask for a schedule of release dates and ask what other data are collected. Talk to an expert before you use data older than 1990.

The pace of job growth is probably the best economic-health indicator, not to mention consumer confidence.

• New unemployment claims: One of the few "leading" indicators you'll get on a local level. Check with your state's employment agency and ask about unemployment compensation data. Many states can provide this data weekly. It's unlikely that the data will be totaled for your Metropolitan Statistical Area, so you may have to add it up yourself.

The data typically report how many people file claims at particular offices, but keep in mind also that many states are switching to telephone reporting systems, so try to find data on the people who file claims by phone in your area.

• New home starts: Strong activity means jobs for builders and reflects consumers' willingness to make long-term financial commitments. Check with your local homebuilders association. A state agency might also provide this.

• Commercial (nonresidential) construction: Activity here means jobs for construction workers and reflects the willingness of business to invest and expand. The F.W. Dodge division of the McGraw-Hill Construction Information Group is the main source of data on construction activity. The division issues press releases monthly with construction totals for nonresidential construction. That includes manufacturing, commercial, education and medical buildings.

One caveat: The totals in the press releases are preliminary and you don't get the final number until the next year. The revisions can be as much as 50 percent, but this is the best available data. Contact F.W. Dodge PR at 212-904-2439, or Bob Murray, VP economic affairs, 781-860-6623.

• Retail sales: The U.S. Commerce Department used to provide these figures but stopped doing that around 1996. If you have a local sales tax, you might be able to get those collections from your state's taxation department or local government.

• Industrial electricity sales. This reflects the level of manufacturing activity in your area. Ask your local electrical utility if they'd provide this for you. My local utility folks used to provide me with the figure in kilowatt hours, but today they provide an index so they don't have to disclose Kwh. They can also break this down by type of industry.

Using the data

• Chart it. You'll understand it better if you visualize it.

• Run comparisons. Keep in mind almost all the data you get will not be seasonally adjusted. That means you

can't really compare February's numbers to January's; you have to compare them to February a year ago.

• Calculate change both as percentage change and nominal change. Data will be clearer to the reader. When writing, you say the economy created 10,000 new jobs rather than say the number of jobs grew 2 percent.

• Calculate year-to-year change. Run a 12-month moving average and chart it against the main data line. This will smooth out seasonal changes and allow you to see trends.

• Calculate a year-to-date total and change from a year earlier. This is helpful with indicators that are cumulative, such as construction or unemployment claims. Be mindful of revisions in the data. Labor data is revised every March for the previous two years. As mentioned above, the construction data can be substantially revised.

Analysis of the data

Analysis is tough for a lot of reasons. If you're collecting data for March, you'll have some in April but you won't have all of it until the end of May. Also, it's sometimes tough to know exactly what these indicators are indicating.

All this makes it difficult, for instance, to build a local economic index. Think your readers will be interested in June to see your March index?

Instead, the data is a conversation starter. The best you can do is eyeball it, get a sense of where things are going and who's most affected and begin your reporting.

John Byczkowski can be reached by e-mail at johnb@enquirer.com

readme.txt

Other data measuring local economy

1. The unemployment rate While it's less important than the rate at which the number of jobs is growing (or not growing), the unemployment rate is still worth knowing. Keep in mind your local rate may not be seasonally adjusted, which means it's not comparable to the national rate A national unadjusted rate is available at ***http://stats.bls.gov.***

2. Browse Census Quickfacts ***http://quickfacts.census.gov/qfd/*** for state and county business information.

3. The Economic Census: Every five years the Census Bureau does an economic census, and it gives a great picture of business in your area. Trouble is the last one was 1997. The next one will be done this December, and the data will be released starting in 2004.

4. Read the Beige Book. It's at the Federal Reserve Web site ***www.federalreserve.gov***, under "Monetary Policy." You'll get story ideas about current economic conditions in your region.

5. Bureau of Economic Analysis: ***www.bea.doc.gov***. There's a "State and Local Area Data" section, and check for more data under "Publications."

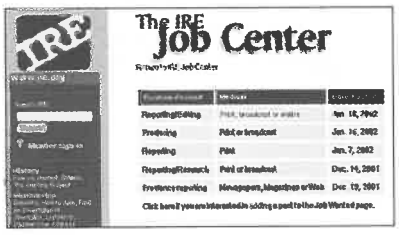
6. Commercial real estate: Somebody in your area is tracking commercial and industrial properties, watching vacancy rates and rents. Most surveys of this type are done quarterly.

7. Notices of big layoffs: Your state should be able to provide these to you. Are big layoffs increasing in your area?

8. Manpower survey: A quarterly survey of company hiring plans. Contact Manpower in Milwaukee at 414-961-1000.

9. Economic Cycle Research Institute: 212-557-7788 and talk to Anirvan Banerji, director of research (ext. 11), or Julienne Kim, publications manager (ext. 10). Alternatively, go to ***www.businesscycle.com.*** This used to be part of Columbia University and pioneered leading indicators.

JOB-WANTED ADS
for IRE Members



IRE makes it easier to find a job!

Your anonymous job-wanted listings can be posted for free on IRE's active Web site to be seen by Web-browsing news managers.

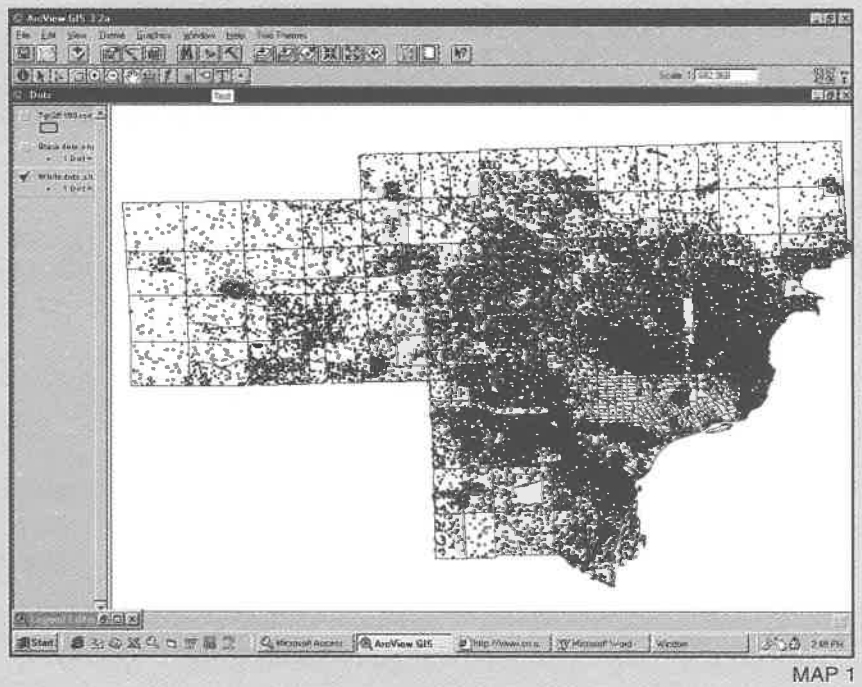
Describe your qualifications and the job you want and invite potential employers to contact you through a blind e-mail address we provide. It will automatically forward responses to you – without revealing your identity.

For more information go to:
www.ire.org/jobs

MAPPING IT OUT

This is a new Uplink feature that shows off analytic maps created by journalists for their news reporting.

Plotting race by population

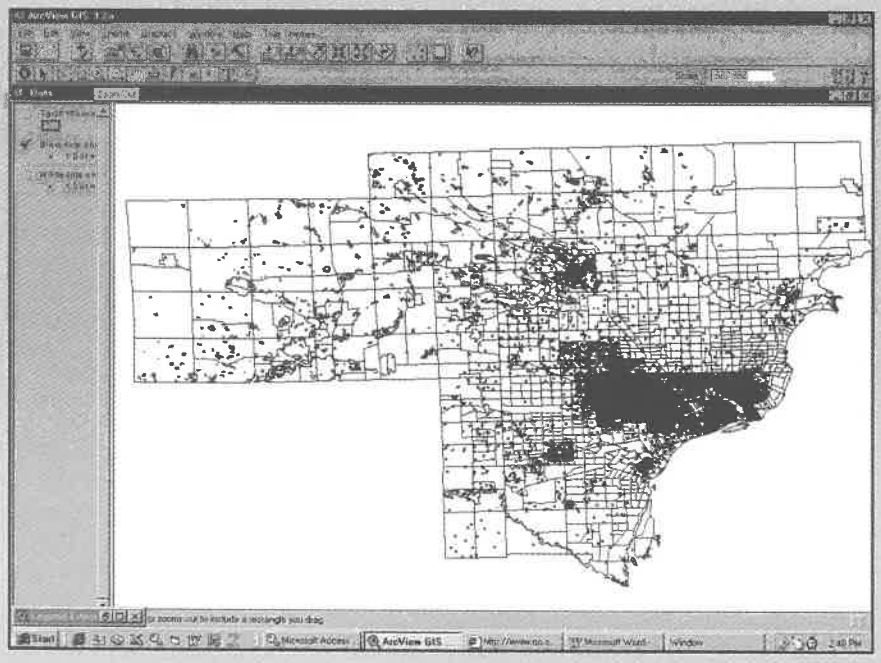


MAP 1

Census 2000 provided journalists a once-in-a-decade opportunity to explore racial population patterns in their communities. Using geographic information systems (GIS) and census data, journalists across the country mapped the population patterns of blacks and whites, looking for segregation and migration. As you can see from the examples on these pages, journalists took different approaches in their mapping, but all found interesting stories.

'Cost of Segregation'

Gordon Trowbridge of *The Detroit News* created two dot-density maps of the Detroit metropolitan area to explore segregation between blacks and whites in 2000 for a special report, "The Cost of Segregation." (See the report at *http://pc99.detnews.com/specialreports/index.hbs*). Trowbridge used ArcView to plot each race on the separate maps, with one dot representing 100 people inside a census tract. (See Map 1 for white and Map 2 for black). The juxtaposition of these two maps – later dressed up for print – clearly showed readers the high segregation in the city of Detroit. Trowbridge says that the maps were useful internally in helping editors understand the severity of segregation.



MAP 2

Trends in 'tipping'

Jeff Hansen of the *Birmingham* (Ala.) *News*, used ArcView data to examine "tipping" – shifts of 20 percent or more in the percentage of black or white population. Hansen's map (Map 3) here shows the city of Birmingham as a black outline and the tipped city block groups as white. The block groups outside

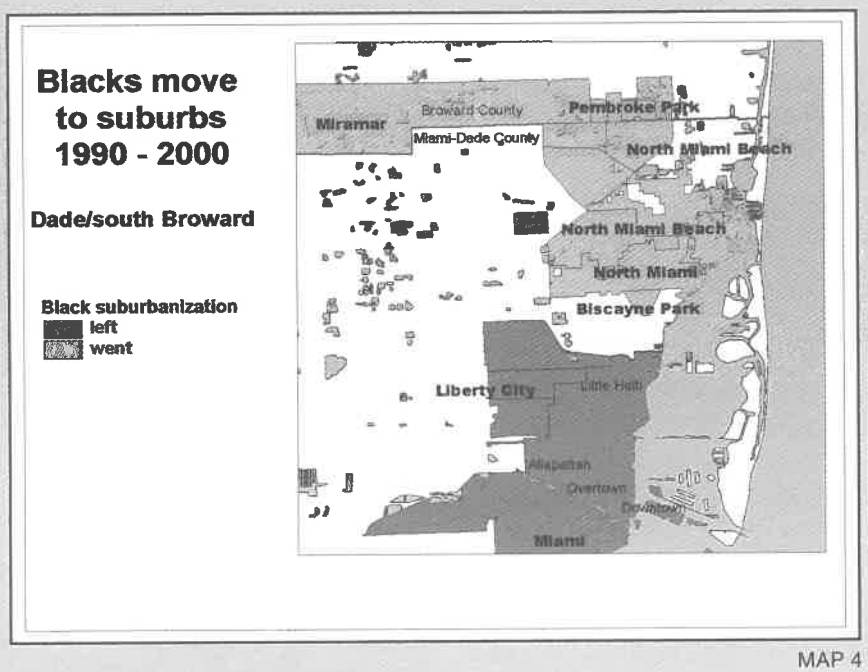
MAP 3

Do you have a map that you'd like considered for Mapping it Out? Just send an electronic copy of the map – published or not – along with a note to David Herzog, the editor of this feature, at dherzog@nicar.org.

the city that tipped are shown as gray. Digging deeper into the data, Hansen found that about one in 11 people in Alabama live in an area that saw a steep racial change during the 1990s. The mapping also led Hansen to two interesting trends: First, middle-class black families moving out of older, smaller homes, to larger, new homes on the city's east side. Second, surges in white population in former rural black areas where subdivisions had since sprouted.

Tracking racial shifts

Tim Henderson, a database editor at the *Miami Herald*, looked at black population shifts a little differently. He used ArcView to map the places in Dade and Broward counties where blacks had left during the 1990s and where they went. (See Map 4) After some trial and error, Henderson settled on these definitions for shading the map: Places that had been at least 20 percent black in 1990 and lost 10 percent or more of its population during the decade was shaded as "left." Places where the population increased by 10 percentage points during the decade were shaded as "went." The resulting ArcView layout, Henderson says, shows that blacks tended to move out of the older, inner-city neighborhoods of northwest Miami and neighboring Liberty City and into the more suburban areas in northern Miami-Dade County and southern Broward County.



MAP 4

POLICE SHOOTINGS

Gunning for safety

By Brian D. Crecente and Sarah Huntley, *Rocky Mountain News*

When the Denver Police Department landed on *The Washington Post's* national 10 worst list for police shooting fatality rates, department officials were outraged.

They said the numbers weren't a fair comparison and painted a picture of the department that wasn't true.

A three-month investigation by the *Rocky Mountain News* found, that while the fatality rates could indeed be construed as misleading, the department's system for tracking police-involved shootings is a secretive and outdated process that is protected by police unions, the department's top brass and city hall.

The search for data started at the Denver's District Attorney's Office.

The legwork

The district attorney reviews all police shootings that result in significant injury or death to determine whether the officer should be charged with a crime. After making a determination about whether to prosecute, the district attorney's office issues a lengthy and detailed decision letter.

When the letter is released, so are the investigative files. The files, which include police reports, written statements and videotaped accounts given by officers and witnesses, were a data gold mine.

We requested all of the decision letters going back to 1990 and then spent the next month poring over the letters, making notes and highlighting relevant information.

We decided early on to let the data dictate what stories we would be writing, so we created a basic Access database and frequently modifying it to meet our emerging needs.

Tailoring a database

The first database we created had fields for the basics about the incident, the person who was shot and details on up to six officers.

After reading a half-dozen reports, we decided to add a few yes/no checkboxes to the database. We wanted to know how many of those shot were on drugs or drunk when the shooting happened and how many showed signs of possible mental illness. We created other checkboxes to denote when an officer had a gun pointed or fired at him and if the officer or a civilian felt the person shot threatened them.

All of this data, while open to interpretation, came directly from the decision letters. The only exception to that rule was a yes/no box created to show if we thought the shooting was questionable. We used that to help highlight the cases that we needed to go back and re-read and eventually pull the entire case files for.

As we went through the data, we also decided to create fields for a description of the civilian's weapon and later a generic description for the weapon. So, in time, an unloaded pearl-handled .22-caliber Rohm revolver became a gun.

We did the same thing for the events leading up to the shooting. Each incident had a brief description of an incident and then a matching field that used a generic version to help us categorize the cases.

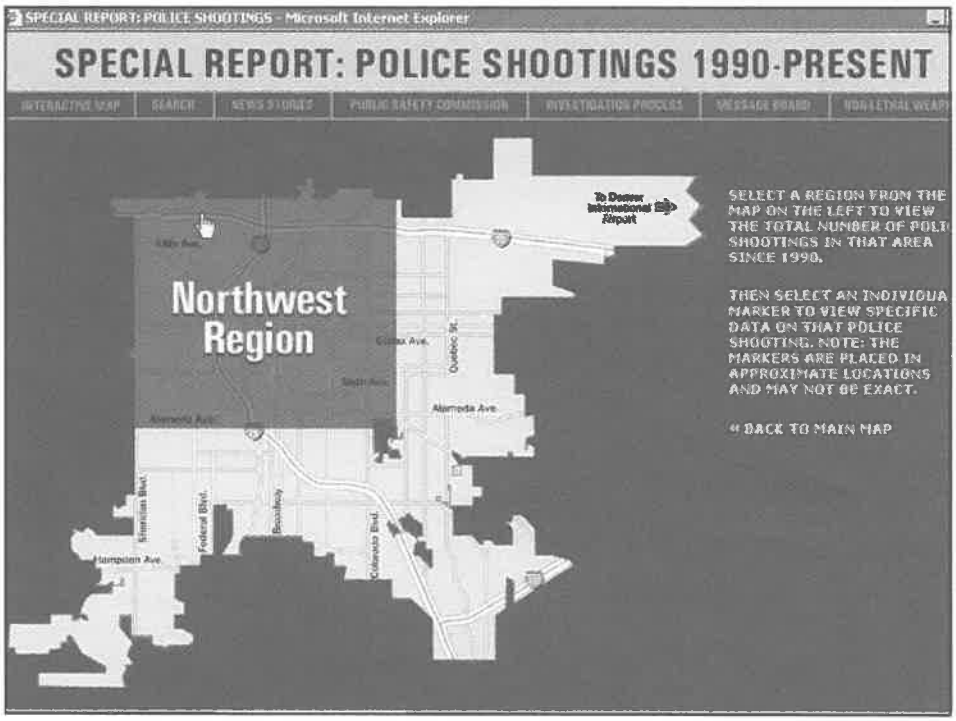
There was some data we entered but never used.

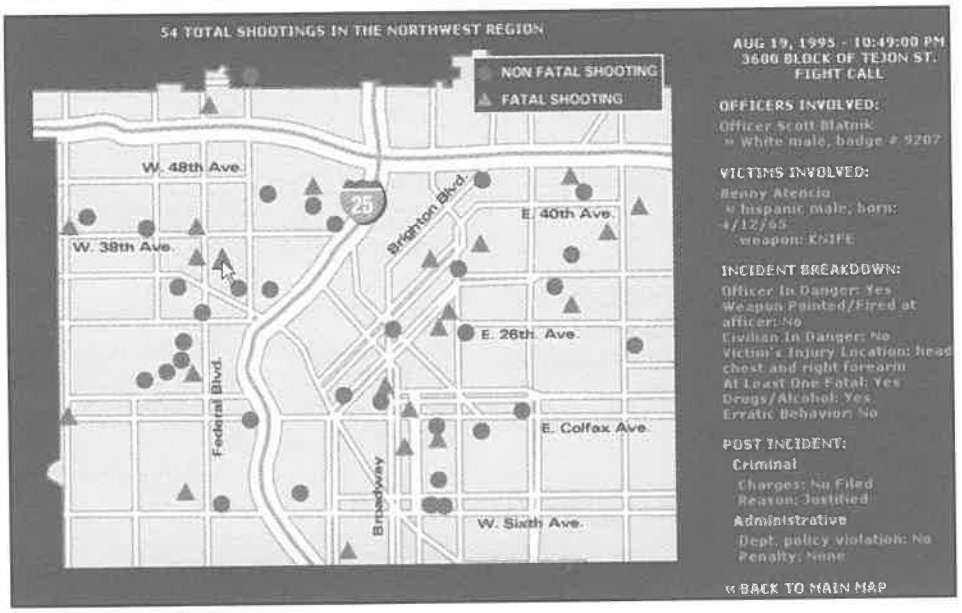
We included the number of times a person was shot, but after reading through 86 letters decided the information wasn't accurate enough to include. The same held true for the location of the injuries.

Quality control

By the time the database was complete we had a file that was as wide as it was long, with 86 records each with 76 different fields.

Once the database was created, we





decided to run it by the police department to make sure we had all of the incidents that happened over the past decade.

The problem was that the department's policy, which leaned heavily on state law, restricts meaningful access to officer discipline.

The department finally agreed to check our data by comparing our list of every officer we felt was involved in a shooting with their internal affairs records.

Over the next few weeks, we were able to find a handful of missing cases and finally completed an accurate list of officers involved in shootings that the police and we agreed upon.

At first, department officials provided us only with a numerical breakdown of officers disciplined in connection with shootings, refusing to give us the actual names of the officers or specifics about the cases.

As we continued to probe, however, we learned that the department had created a spreadsheet of cases reviewed by the Firearms Discharge Review Board, an internal group made up of the department's top officials.

We had already done extensive research on the discharge review board and determined it would play a major part in our stories. The board is supposed to review every police shooting, including those that do not result in any injuries, and come to a conclusion about whether an officer followed departmental policy. Our study had found that the board rarely conducted any additional independent investigation.

The department faxed over a complete list of every officer involved in a shooting, whether the officer was disciplined and the discipline received. The list didn't include names of officers, but did have the dates the incident happened and a brief description of each incident.

After a little cajoling, officials agreed to e-mail the list over as an Excel file. We converted it to an Access database and then cross-referenced the data using the date.

Finally we went through each case individually and double-checked that the descriptions matched.

With the help of staff writer Peggy Lowe, we published seven articles in a two-day series.

The results are in

Our analysis found:

• From 1990 to 2000, Denver police shot 91 people, killing 35. The city has paid nearly $3.7 million to settle lawsuits and claims filed by the injured and the families of the dead.

• Of the 126 officers who shot people, five were disciplined. The most serious punishment was a three-day suspension plus a one-day fine. In four of the cases, the officers were penalized not for shooting a person, but for shooting at a car.

• More than a third of the people shot were mentally ill, high on drugs or alcohol or acting irrationally in the moments before officers fired. Only seven survived. Denver police are planning to implement a new crisis intervention program in hopes of lowering these types of shootings in the future.

• The Denver Police Department does not have a comprehensive, computerized method for tracking officers or identifying possible problems. A deputy chief said the current system is a "hodgepodge" that relies upon institutional memory and occasional reporting.

• The city's only citizen watchdog group, the Public Safety Review Commission, has failed to bring meaningful oversight to police shootings. The group's advocates say the accountability process is stymied by politics, a lack of funding for full-time investigators and a growing caseload.

Community response

We worked with our online team to put a searchable form of the database up on our Web site and to publish a map that showed the location of every shooting and details of each. (see screenshots on this and previous page.)

Putting the database online required stripping out some information, renaming fields to better explain what they were supposed to represent and deciding how best to allow readers to search the voluminous information.

The stories received mixed feedback, with some saying it was time someone took a closer look at law enforcement conduct and others accusing us of

continued on page 22

E-PANELS

Building databases of sources

By Jim Hopkins, *USA Today*

It's 7:05 p.m. The deadline for your story about the tumbling stock market is an hour away. At the last minute, your editor wants quotes from "real people" – investors whose panic-stricken reaction will add color. Now what?

You could grab the telephone book and make random calls, hoping to find someone with an articulate, punchy quote.

A better-prepared reporter would turn to a list of investors' names and phone numbers who've already agreed to be called in just such a situation. The list – let's call it a panel – might be scribbled out on paper or kept in a computer file.

But the best-prepared reporters will turn to what I call an "e-panel" – a searchable database of investors' names, phone numbers and other identifying information: Gender. Race. Age. Annual income. Occupation. Investment goals. And perhaps most important, e-mail addresses.

Now technology makes it easier than ever for newspapers and other media to create these sorts of databases, then mine them for trends and other story ideas. Best of all, an e-panel can be built with minimal effort using software such as Excel, already available in many newsrooms.

Of course, e-panels don't have to be limited to stock market investors. Imagine you had a database of teen-agers to ask questions about music, fashion or peer pressure. Other topics: consumers, sports fans, TV viewers, frequent travelers. The sky's the limit.

E-panel on deadline

As *USA Today*'s entrepreneurs reporter, I couldn't do my job without our Small Business Panel. It comprises 2,100 U.S. companies, with e-mail addresses, phone numbers, annual revenue, location, number of employees and other identifying information – all stored in an Excel spreadsheet.

The panel was indispensable on Sept. 11, when we were writing stories about the terrorist attacks. I work in our San Francisco bureau. At 8:14 a.m. Pacific time that day, I sent 150 e-mails to companies asking how the tragedy affected them. Within minutes, the first of nearly 50 replies poured into my laptop. They were rich with quotes that I immediately fired off to our lead reporter on the East Coast.

Technology made it easy to build the Small Business Panel

Business consultant Anna Bernstein e-mailed from Manhattan: "If you go out on the street, it's strangely quiet. Business people all over the sidewalks at noon and no one speaking." Without our Small Business Panel, I would have spent hours of cold calling to find someone like Bernstein at ground zero. And anyway, phone circuits were tied up; e-mail was the only way I could communicate.

Another example: For a story about access to venture capital, I needed a half-dozen specific kinds of businesses – very young companies started by women. In the past, I would have called trade groups and other sources for a couple of days. Instead, I searched our Small Business Panel database for all women-led companies open three years or less. That produced a list of 128 firms. I copied their e-mail addresses into a single Microsoft Outlook e-mail form. Then I wrote one e-mail asking the firms if they fit my criteria. I included a deadline for replies, and wham! fired off 128 copies.

The entire process took less than 15 minutes. Within 24 hours, I had seven perfect replies. And as you can see, by storing the panel in a spreadsheet, I could quickly slice and dice the contents to find companies that were likely matches.

Gathering the data

Technology made it easy to build the Small Business Panel because we used *USA Today*'s print edition and Web site to gather names. A note published in the Money section directed readers to a *USA Today* Web page with an electronic survey form. Readers filled in e-mail addresses, phone numbers and other information. Responses were saved in a text file that was then easily imported into Excel.

Many newspapers may need to use a more traditional approach – asking readers to complete an in-paper survey form, then mailing it to the newsroom. Information from the forms can then be entered into an Excel or similar database by hand.

Tips on how to create your own e-panels:

• Get tech advice. Talk to the person who maintains your newspaper's Web site to see if they can build an online survey form.

• Plan ahead. Imagine the kind of stories you hope to draw from your e-panel then reflect that in your in-paper or online survey form. Example: I knew I would be searching the Small Business Panel for manufacturers or retailers, so our form asked people to specify their industry.

Jim Hopkins can be reached by e-mail at jhopkins@usatoday.com

SPOTLIGHT: NICAR IN PHILADELPHIA

Tracking campaign donors through City Hall

By Bob Warner, *Philadelphia Daily News*

If you're a reporter covering city hall or state government, maintaining databases on political contributors is essential. In most places, you can't really understand what's happening in local government or politics without knowing who is financing the campaigns.

Reforms at the federal level might push even more political money into state and local coffers, so you've got to develop a strategy to track it. Here are a few suggestions.

In city hall or your county seat, spend some time in whatever office is responsible for maintaining campaign finance data. Try to get to know the people who file away all the reports – it's a thankless job, and the people who do it well are probably just as shocked as you at the amount of money now influencing the political process. No matter how good your databases are, you'll occasionally need help from real humans tracking down paper records. Prepare a list of all the public officials you're interested in and take a fast look at their most recent reports. Make sure they're up to date, and figure out which ones are already using computers to prepare their paper filings.

If your employer will pay for it, you can hire someone to input data into your own databases. If not, you've got two choices: start typing yourself, or call the pols, asking them to share the donor data that they've already fed into their own computers. At first blush, this may go against your grain. Nobody wants to ask favors from the people they cover. But that's the wrong way to look at it. This is no personal favor. Your public officials are asking city contractors and other people to give them hundreds or thousands of dollars. All you're asking them to do is facilitate public disclosure of their fund-raising. Surely they have no objection to that, do they? If they do, ask them to explain it. Very few will be willing to say for the record that they want to obscure the names of their contributors.

In Pennsylvania, the major candidates for governor, mayor and other major offices are now usually cooperative and provide the news media, voluntarily, with diskettes or e-mail attachments identifying their contributors and campaign expenses. When they're disclosing $10 million worth of donations at 4:55 on a weekday afternoon, you'll have a hard time writing the next day's story unless the candidate provides a digital report.

Find a way to get lists of contributors into your newspaper. Every time we do it, we get story tips from people who suddenly understand why they got beat on that last contract awarded by City Hall. Don't be shy about calling people to find out why they donate. Sometimes they'll be remarkably candid.

Be creative about campaign finance stories and charts. Write about the biggest donations from people outside the city or state, who won't be able to vote for the candidate they're sending money to. Write about the people who are giving to three or four candidates in the same campaign. Pay attention to contributions from out-of-state PACs – frequently they'll be raising money under rules that are different from those that prevail in your own state; for instance, some can take money from corporations. When you write about various campaign issues, include a box with relevant campaign finance data – lists of union contributions to go with a story on organized labor, for instance, or lists of real estate and developer contributions to go with a story on suburban sprawl. Use a mapping program to show which areas of your city or state are making the heaviest political donations or how financial support for different candidates breaks down geographically.

Pay attention to campaign expenditures as well as contributions. Some of your best stories may involve transfers of money between different political organizations, or the use of campaign funds for personal expenses or other questionable purposes. And don't forget about in-kind contributions – frequently, they'll help identify a candidate's fundraisers, a particularly influential group.

Bob Warner can be reached by e-mail at warnerb@phillynews.com

readme.txt

Ideas welcome

Have you or one of your colleagues recently published stories using CAR, either one of particular note or someone's first CAR story after attending an IRE and NICAR Boot Camp? If you have a story idea, we'd like to hear from you. Just e-mail your ideas to ***uplink@nicar.org***.

Tech tips

Have you found a technical solution that saves you time? Offer it as an *Uplink* Tech Tip, a story of 500-1,000 words showing a technique that you'd be willing to share. Please e-mail suggestions to ***uplink@nicar.org***.

Tech tips...

Looking for the perfect — fuzzy — match

By Ron Nixon, *IRE and NICAR*

Matching information from two datasets is seldom an easy task, especially if they lack a matching pair of unique identifying fields. So often we're forced to join tables on several fields (name, address, date of birth, etc.) to find matches. This, however, does not find misspellings and other problems created by poor data entry.

Over the years various government agencies, medical professionals and genealogists have developed several name-matching algorithms that journalists can use. These algorithms allow you to make matches where the spelling is slightly off, something that is hard to do with traditional database joins. They can also be used to check for duplicates, even when the names are off slightly, and searching a database when you don't quite know the spelling.

Soundex can help

One of the earliest matching algorithms developed was the Soundex function. It was developed by the Census Bureau in the late 1800s to find names that sounded alike phonetically. Soundex is a built-in function in FoxPro, SQL Server and many other software packages, but not Access.

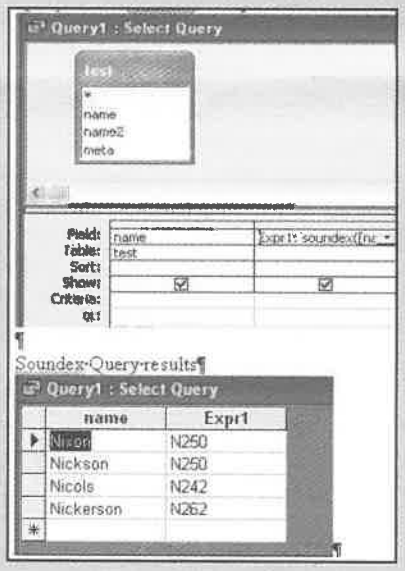
To add the function to Access, you can download an Access 2000 database file that includes – in the "Modules" section – the Visual Basic for Applications scripts for Soundex and more at *www.nicar.org/techtips.html.*

Soundex gets rid of all the vowels in a surname, assigns numbers to the remaining letters and changes it into an alphanumeric string. Using these codes, reporters can match names that are similar phonetically. Take, for example, the Soundex code for Nixon: N250.

This query in Access using SQL would match all names that had the same Soundex code:

```
Select a.name, b.name,
otherfield
From table a, table b
Where
Soundex(a.name)=Soundex(b.name)
```

Of course, you'll get some matches that aren't really the same. Using a database of surnames, I found that there were 63 Soundex matches for Nixon, many of which bore no resemblance to Nixon. And misspellings of the name might not match since the code could be different.



Soundex has other weakness. Since it is based on English pronunciation, many European names may not translate correctly to a Soundex string. For example, some French surnames will not translate according to pronunciation. Take the names Beau and Beaux. Both are pronounced the same, yet they would have different Soundex codes (B000 and B200, respectively).

These limitations in Soundex led the New York State Identification and Intelligence System in 1970 to create a new algorithm named after the agency — NYSIIS. Many intelligence agencies today continue to use NYSIIS in one form or another.

Downloading NYSIIS

NYSIIS is not native to FoxPro, Access, SQL Server or any of the other software products commonly used by CAR mavens. There is FoxPro code for the algorithm, however, downloadable from *www.civilsolutions.com.au/publications/dedup.htm.* Place the downloaded program in your FoxPro default folder. An example of testing NYSIIS:

NYSIIS works well with Spanish and southern European names. It supposedly has difficulty with Japanese and French derivative names.

Unlike Soundex, NYSIIS maintains vowels, but changes them all to As.

It also returns a purely alphabetic code. So under NYSIIS, Nixon becomes NAXAN. Whereas Nixon and Nickson would have matched under the Soundex algorithm, they would be different under NYSIIS. Nickson would be NACSAN. Using a FoxPro SQL query to match the names in two different databases with NYSIIS:

```
Select a.name, b.name;
From table a, table b;
Where
NYSIIS(name)=NYSIIS(name)
```

This would give you matches, based on stricter criteria than Soundex. NYSIIS resulted in only four phonetic matches for Nixon, with only one bad match. No doubt you'll miss some matches because of the more strict rules in NYSIIS; however, it will save you the trouble of trying to weed through hundreds of wrong matches.

Watching Metaphone

In 1990, Lawrence Phillips, a computer programmer, sought to improve on Soundex and created Metaphone.

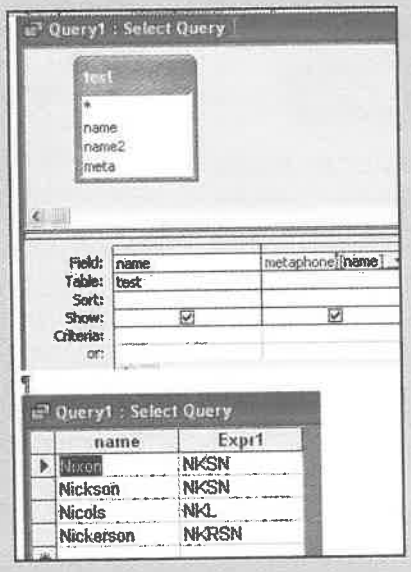
While similar to Soundex in many ways — it strips out vowels— Metaphone attempts to follow the commonplace rules of English. As an example, Soundex equates "g" and "j" no matter how they are used and also assigns "c," "q," "z" and "k" to the same number.

Metaphone reduces the alphabet to 16 constant sounds and gives a phoneticized spelling of a name. Under Metaphone, Nixon becomes NKSN and would match Nickson, which would also have the same code NKSN. Metaphone, because of the way it is constructed, limits matches more than Soundex. Metaphone, used on the database of surnames mentioned above, found seven phonetic matches for Nixon, with five being good matches.

Matt Waite at the *St. Petersburg Times* used the algorithm and received good results, unlike when he used Soundex and got hundreds of wrong matches.

"I took 650,000 records from one agency's check register, and 500 records from another licensing agency and POOF! I get 5,127 matches, and they were good matches. Misspellings, strange spellings, abbreviations, they're all there," he said.

Metaphone is included in the Access database file mentioned earlier, but is not part of other widely used database manager programs.



Phillips has since improved on the Metaphone algorithm with a new one called Double Metaphone, which reduces the alphabet to 12 constant sounds, and gives an even stricter match. Soundex, too, has been improved. Soundex2 debuted not long ago and corrects some of the problems with the original version.

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

readme.txt

Web addresses

For the use and rules of these algorithms see:

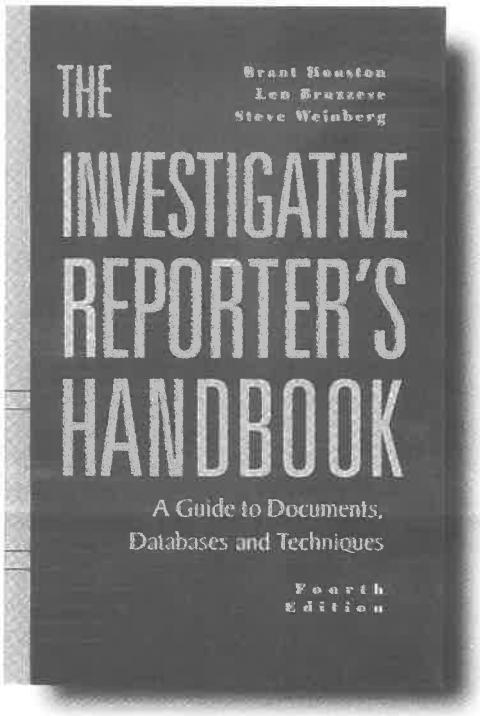
Soundex:
www.nara.gov/genealogy/soundex/soundex.html

NYSIIS:
www.civilsolutions.com.au/publications/dedup.htm

Metaphone:
www.webbtide.com.au/metaphone.html

The Investigative Reporter's Handbook, 4th Edition

by Brant Houston, Len Bruzzese and Steve Weinberg
of Investigative Reporters and Editors, Inc.



The new edition of this indispensable reference has been updated to include examples of local investigative reporting and features easy to find Internet address lists to help students and professionals in computer-assisted investigations. The most comprehensive text available about investigative reporting, this useful source book created by IRE shows how to investigate anyone or anything.

Table of Contents

- **Introduction:** Paper trails and people trails: an overview
- **Secondary sources:** Working from the outside in
- **Primary documents:** Obtaining the best evidence
- **Computer-assisted reporting**
- **Crossing borders:** International investigations
- **People trails:** Finding and interviewing sources
- **Investigating government:** The legislative branch and those who try to influence it
- **Investigating government:** The executive branch
- **Investigating government:** Education
- **Investigating government:** Law enforcement
- **Investigating government:** The judicial system
- **Where government and private sector meet:** Investigating licensed professionals
- **Investigating the private sector:** For-profit businesses and their workers
- **Investigating charities and other nonprofits**
- **Investigating health care**
- **Investigating insurance**
- **Investigating financial institutions:** Banks savings and loans, credit unions, investment companies and their mutual funds
- **Investigating energy and communications utilities**
- **Investigating transportation**
- **Investigating real estate:** Housing, commercial uses and zoning
- **Investigating environmental issues**
- **Investigating the world of the disadvantaged**
- **Writing compelling projects**
- **The ethics and accuracy of investigative journalism**

ORDERING INFORMATION:

ONLINE: www.ire.org/store/books
PHONE: 573-882-3364

Members $48
Nonmembers $60

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

CAMPAIGN FINANCE

Looking at campaign expenditures

By Ted Sherman, *The Star-Ledger*

"Follow the Money" is the mantra of campaign finance reporting.

Sometimes, though, the most interesting trail is not who gave — but where the money went.

The focus of campaign finance analysis usually looks at how much special interest groups are able to funnel into any given election. Each year, we build databases for statewide campaigns such as the gubernatorial race, the political action committees of the state's legislative leadership, as well as the U.S. Senate campaigns. The aim is to document the bundling of contributions from powerful law firms, developers, lobbyists and others looking to parlay contribution limits into a much louder voice.

How the campaigns actually spend the money, though, is not always so carefully analyzed. Campaign finance reporters will look at total expenditures and break down major spending — such as media buys and consulting expenses. The trouble is that expenditure reports are not available as electronic records for Senate or statewide campaigns, and it costs money to get them inputted for database analysis. It can be money well spent.

In this year's race for mayor in Newark, N.J. — a classic battle between an entrenched political organization and a well-financed, self-styled reformer — an initial look at campaign finance reports suggested the disbursements would be fertile ground for further examination.

The incumbent, Mayor Sharpe James, had raised $1.3 million going into the May campaign. Yet his expenditure reports showed he had spent just about all of it before even declaring he was a candidate for office. Eyeballing the paper records available from the state Election Law Enforcement Commission raised other questions. With more than a dozen reports and four years of data to examine in the current election cycle (more than 3,000 records overall), we opted to pay someone to type it into a database.

Grouping the expenses

We use a simple technique to examine bundling and to group election data. We create a separate column that shows ties among contributors or expenses. For example, we will carry over the employer records to that column, and then edit it to include spouses, children, or related companies, so that grouping shows the full extent of bundling. For expenditures, we will classify payments into subcategories — such as entertainment, fundraising, salaries and consulting.

After getting all of James' contributions and expenditures into an Access database, we documented where the mayor got his money — and it came from those who work for him. Nearly 30 percent of the money to the mayor came from 678 contributors on the Newark city payroll — including clerks, administrators and members of the police and fire departments.

"Employees come because they want to come," he said of his fund-raisers. "They come because they do the Electric Slide. They come because they know they're going to have fun, because I'm a fun person."

But those same reports revealed that the mayor spared little expense on those and other events. After cleaning up and grouping that data, we found he had spent $290,558 on catering over the past four years, $73,015 on entertainment, $8,241 on the design and printing of invitations, and $7,280 on centerpieces. He spent $10,624 in airfare from Los Angeles one year for the entertainment for one birthday fund-raiser.

The story basically showed how the mayor had spent little on actual campaign expenses, but instead utilized the trappings of old-style, ward politics — like Thanksgiving turkeys, summer picnics, and city jobs – to solidify his political base.

The first time we employed this kind of intensive expenditure analysis was during the 2000 campaign of Jon Corzine, who spent more than $62 million in his successful run for U.S. Senate in New Jersey. Again, the story was prompted by the level of spending.

As in the case of Sharpe James, we had every contribution — as well as every expenditure — input into a database, so that we could aggregate even the seemingly smallest expenses. There, we found that Corzine was not only spending money (mostly his own) on consultants and a media blitz, but was spreading thousands of dollars in contributions to local politicians, to county committees, churches and ethnic groups. He donated $10,000 to the Hunterdon County Democratic Committee—a county with just 8,916 Democrats, where Republicans outnumber

continued on page 22

readme.txt

Searching for stories

Are you exploring the political landscape with computer-assisted reporting? For the July-August Uplink, focusing on campaign finance, we'd like to hear about your stories. Please send details such as data and software used, to ***uplink@nicar.org***. If you want to share paper copies of your stories instead, send them to Jeff Porter or David Herzog, 138 Neff Annex, Missouri School of Journalism, Columbia, MO 65211.

money

continued from page 21

Democrats more than 2 to 1.

Corzine's campaign finance reports revealed he contributed to a local politician who was not up for re-election, but has a core constituency of Trenton voters he needed to win. In the two weeks before the election, our expenditure database was able to spell out how Corzine spent more than $2.8 million on canvassing, field organizing and other get-out-the-vote activities.

Uncovering stealth giving

Expenditure reports can also be a roadmap to stealth giving. Some county committee reports we examined, for example, showed huge contributions to other county committees. Why would any county raise money and give it to someone else's candidates? We learned later it was a case of conduit financing. Some vendors don't want the powers-that-be in one county to know they are giving to both sides. So they give it to another county committee, which later writes out the check.

Such reporting can lead to other questions as well. In one New Jersey city, the mayor raised eyebrows after his campaign disbursements showed he had taken $25,000 from the coffers of the city's Democratic organization to contribute to the campaign for president of the Dominican Republic.

In addition to helping pinpoint unusual spending in a well-funded campaign full of disbursements, analyzing expenditure reports by computer can also map spending over the timeline of a campaign, and allows the grouping of hundreds of seemingly small payments, like salaries, postage and telephone expenses.

Like so many CAR projects, though, the biggest hurdle is getting the data. To follow the money, you've got to be able to count it.

Ted Sherman can be reached by e-mail at tsherman@starledger.com

gunning

continued from page 15

pounding on police officers. Interestingly, many of the critics mentioned Sept. 11, as if to suggest that scrutiny of law enforcement was no longer appropriate after the World Trade Center attacks.

The police department had already started looking into other models for civilian involvement, including the creation of a disciplinary review board that would include two non-police officers. That work continues. The Public Safety Review Commission is redrafting its budget to request additional funding.

Brian D. Crecente can be reached by e-mail at crecenteb@rockymountainnews.com

Sarah Huntley can be reached by e-mail at huntleys@rockymountainnews.com

readme.txt

For more information

For more details about the *Rocky Mountain News* series, see the May-June edition of *The IRE Journal*. Also, the *Journal* March-April 2002 edition includes an article about *The Washington Post* series on police shootings in Prince George's County in Maryland.

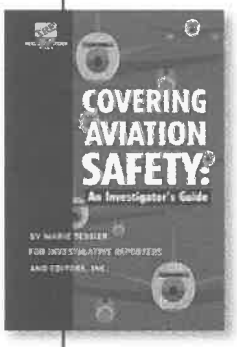
Web addresses

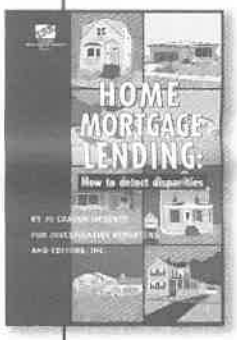
The complete police shootings project, including the *Rocky Mountain News'* searchable database, is available at ***http://cfapp.rockymountinsnews.com/deadlyforce***

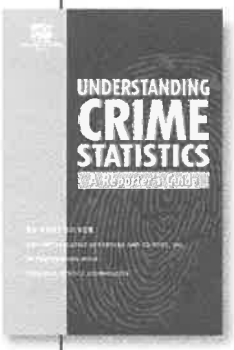
Copies of recent decision letters can be found at the Denver District Attorney's Office, are available at ***www.denverda.org/html_website/denver_da/police_shooting.html***

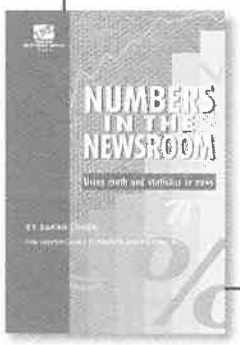
REQUIRED READING
For Your Newsroom

ORDER NOW!









The IRE Beat Book Series

BY PHONE
Call 573-882-3364 with your VISA or MasterCard

BY MAIL
Send your check to IRE:
138 Neff Annex
Missouri School of Journalism
Columbia, Mo. 65211

BY WEB
Visit our Web site at www.ire.org for online ordering or order form downloads

IRE MEMBERS: $15 each
NON-MEMBERS: $25 each

Plus Postage:
First Class-
$4 for the first book,
$2 for each additional book

Q&A From the Lists

Listserv messages you may have missed between members of NICAR-L, Census-L and IREplus-L. To learn about joining these mailing lists and searching their archives visit www.ire.org/membership/listserv.html.

Per 100,000 calculation

Q. I have a total population of 8.9 million — of that population, 325 died from a certain cause. How can I get that to reflect a per 1,000 deaths number so I can compare it to a different population and a different death count?

A. You probably want deaths per 100,000, not per 1,000. Divide 8.9 million by 100,000 which will give you 89. Then divide 325 by 89, which means 3.7 deaths per 100,000. If you did it by 1,000, the answer would be 0.037 deaths per 1,000, which seems pretty meaningless.

Results with Excel

Q. I'm trying to put together an Excel worksheet that shows which candidate carried each precinct. The layout would be Precinct, Candidate, Candidate, Candidate, Candidate, Winner with the winner being determined by who gets the most votes.

A. Try putting this formula in: =INDEX(B$1:F$1,MATCH(MAX(B2:F2), B2:F2,0)) presuming your headers are on row 1, and you have a candidates in columns b through f, with precinct results starting on row 2. You'll probably want to do a count on the occurrences of Max(b2:f2) just to be sure there isn't a tie.

Paradox problem

Q. I'm trying to import some data into Paradox 8.0. It's delimited text. I get a "table is full" error that stops the import. The table is only about 131 megabytes. Paradox is supposed to have a two gigabyte upper limit. Any idea what the problem is?

A. This error message generally appears when you have probably hit the limit on the number of blocks that a Paradox table can contain (65,535 blocks). In order to store additional records in the file, you'll need to increase the table's block size to allow each block to hold more records. See *http://community.borland.com/article/0,1410,10318,00.html*

ArcView attributes

Q. If I want to show only some of the attributes for my table, how can I eliminate the ones I don't want from the HTML Mapper output?

A. In your source view, go into table properties and uncheck the fields you want to disappear. They won't appear in the output.

Changing names via FTP

Q. We got raw election returns via FTP. The returns were sent incrementally over the evening as "results.asc" The problem is we had to grab the file quickly and rename it lest the next file overwrite the old file. Is there some kind of script we could run to re-name each file a unique name? We're running FTP software by Ipswitch on Windows 2000.

A. We used Magic File Renamer, a cheap little piece of shareware, to rename all the Census FTP files. It is slick and easy and it's $15. Here's the site: *http://doodu.tripod.com/*. ... If you are using WS_FTP you can go into Options—>Session (cont'd) and then click the check box "Receive Unique (PC assigns name)".

Select in SQL Server

Q. I use SQL Server for most of my data analysis, and love it because it's quite a bit faster than other programs I've used. However, one little thing really gets on my nerves, and I'm hoping one of you has a solution. When I'm working with data in diagram pane, which is similar to Access' view of tables and joins, I often use SELECT INTO to create new tables based on the joins in two or more current tables. You can easily select the fields you want to use, but every time you do, the SQL that is generated includes an AS, like such:

```
SELECT info9899$.ID AS
Expr1, info9899$.CYACTIVE
AS Expr2, etc, etc, etc.
```

This happens whether you choose two fields with the same name or not. It's a pain, because I've got to edit out all of those "AS Expr" before I run the query. Anyone know how to turn that off, or have a better way to approach this type of problem?

A. When a query is generated, by default, it starts with "select *". After a select all any field you select will duplicate the name. Hence, SQL inserts Expr1 etc. Solution: Remove the asterisk. If you don't have another field selected, SQL will give you an error message. Tell it to buzz off. Also, in the grid, if you already have a series of Expr's generated, simply click on the alias column. All aliases will be highlighted, then hit the delete key.

continued on page 24

conference

continued from page 1

Another special guest speaker, Mark Monmonier, geography professor at Syracuse University, spoke to dozens of journalists about the best and worst uses of mapping in the news.

And, on that Saturday night, IRE and NICAR honored University of North Carolina Professor Philip Meyer, one of the guiding lights of computer-assisted reporting with a special tribute, recognizing his achievements in promoting the use of social science methods to uncover the news.

Many of the panels and discussions recapped issues from the big stories of the past year: terrorism and war, the Florida election ballot count and the 2000 Census. Presenters on other panels explained how they used data to cover beats (local government, the statehouse, police) or drive the big projects.

During the beginners' track sessions on Thursday, panelists offered entry-level CAR tips, everything from getting started to grabbing data off the Internet.

With so many panels and presentations, it's impossible to mention all the great happenings at the annual conference. Instead, in this issue of Uplink, we're passing along a sampling of tips from presenters.

John Byczkowski, of the *Cincinnati Enquirer,* who was on the "Describing a Community's Economic Health" panel, lists ways you can see if your region is vibrant or sickly.

Bob Warner, of the *Philadelphia Daily News* and moderator of the "Finding Waste and Corruption in Local Government" panel, tells how you can track campaign finance in city hall.

Tom Torok of *The New York Times* explains how he created a special intranet (internal network using Web tools) after Sept. 11 so editors and reporters at his newspaper could manage the crush of information coming in from around the globe.

And Paul Overberg of *USA Today,* from his presentation on the panel "Mining Census for Non-Census Stories," tells how to make plans to use the next wave of census data.

If you missed the Philadelphia CAR conference, take heart: You can always attend the 2003 conference in Charlotte, N.C.

David Herzog, formerly an investigative reporter with *The Providence Journal,* can be reached by e-mail at dherzog@nicar.org

Q&A From the Lists

continued from page 23

Highlighting polygons

Q. I'm using ArcView and have a parcel map shapefile that's polygons, but I can't highlight a polygon I've selected — with either the pointer tool or the select features tool. It does turn the record in the attribute table yellow but not the polygon on the view.

A. If you go into legend editor and fill the polygons with some color, they will display as yellow when selected.

Splitting columns

Q. I have a worksheet with 24,000 rows and one column. I want to split the one column into three columns and separate the data by semi-colons — where is that function?

Here's what I have One column that reads: John Doe; 10 Short Street; Lafayette, La. 70506

I want to make that three columns that read:
John Doe 10 Short Street
Lafayette, La. 70506

A. Highlight the column containing the information and go to Data, Text to Columns. Choose "delimited" and put in the semi-colon as the delimiter.

Cleaning up spreadsheets

Q. I've got an Excel spreadsheet that thinks it's bigger than it really is. There are about 400 rows of data, but the sheet thinks it has 65,000. This makes saving, copying and everything else take a long time. I know Access has a "repair and compact" function for databases. Is there something similar for Excel?

A. Go to cell A401 (assuming that's the first blank below your data), then hold down the Shift key, hit the End key, and then the Down arrow. That should select all the rows. Then delete the selected, empty rows.

2002 IRE Regional Conference

October 19-20

The Coastal Georgia Center
305 Martin Luther King Jr. Blvd.
Savannah, Ga.

Panels that touch on every beat:
Hot issues in the South, election coverage, freedom of information and open records, backgrounding individuals, investigating City Hall, courts, cops and education

CAR classes:
Internet, spreadsheets, databases, mapping, statistical software

For hotel reservations call:
Radisson Hotel Historic Savannah
912-790-7000
(ask for the IRE room block)

Sponsors:
Morris Communications Corp.
Savannah Morning News

REGISTRATION FORM.

To register, please complete this form. Visit our Web site regularly at **www.ire.org** or **call 573-882-2042** for the latest details. *Please write carefully! This information will be used to make your name tag.*

Name: _____

Employer/Affiliation/School: _____

Address: _____

_____

City, State: _____ Zip: _____–_____

Office Phone (required): _____ E-mail (required): _____

Home Phone: _____ Fax: _____

To register, mail this form and a check to IRE, 138 Neff Annex, Missouri School of Journalism, Columbia, MO, 65211. To register by credit card, you must have a Visa or MasterCard. We cannot accept American Express. You may fax your credit card registration to 573-882-5431 or register online at **www.ire.org/training/savannah02**.

Cancellations need to be sent via e-mail to jgreen@ire.org. There is a $25 processing fee for all cancellations until Oct. 17. Refunds will not be given for cancellations after Oct. 17.

Card Number: _____ Expiration Date: _____

Card Holder Name: _____

Card Holder Signature: _____

To attend this conference, you must be an IRE member through Nov. 1. Memberships are non-refundable.

MEMBERSHIP

____ I am a member of IRE through Nov. 1.

____**$40** I need to join/renew my U.S. or international ❑Professional ❑Academic ❑Associate or ❑Retiree membership.

____**$25** I need to join/renew my student membership.

REGISTRATION

____**$50** I would like to register for the main conference as a ❑Professional ❑Academic ❑Associate or ❑Retiree member.

____**$25** I would like to register for the main conference as a student member.

____**$15** Late fee for registrations postmarked or faxed after Sept. 20.

CAR DAY (optional)
Sunday, Oct. 20, is the optional Computer-Assisted Reporting Day and requires an additional fee. If you would like to attend, please check one of the following options:

____**$25** I would like to attend the CAR Day on Oct. 20 as a professional IRE member.

____**$20** I would like to attend the CAR Day on Oct. 20 as a student IRE member.

TOTAL $ _____

Uplink Info

A newsletter of the National Institute for Computer-Assisted Reporting

Editor
Brant Houston
brant@ire.org

Managing Editor
Jeff Porter
jeff@nicar.org

Asst. Managing Editor
David Herzog
dherzog@nicar.org

Contributing Editors
Mike Sherry
Amy Sherrill

Art Director
Lisa Triefenbach
lisa@ire.org

Editorial Phone
573-884-7711

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

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

Director of Publications
Len Bruzzese
len@ire.org

Advertising Coordinator
Pia Christensen
pia@ire.org

Subscription Administrator
John Green
jgreen@ire.org

Subscriptions
IRE members $40, nonmembers $60

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

Postmaster: Please send address changes to IRE-NICAR.

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

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

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