Uplink

KIDS AND VIOLENCE

Behind school walls

By John Kelly
The Associated Press, Indianapolis

Newspaper headlines and nightly news teasers delivered the good word: fewer kids were caught bringing guns into America's schools.

The news stories cited a federal survey of the states, which reported a drop in gun-related expulsions during the 1997-1998 school year.

The timing was bizarre for me in Indianapolis.

For months, I had been eyeballing a similar school discipline database I knew the In-diana Department of Education was compiling. Just days before the national report came out, I finally got the disk with the Indiana schools' data.

One dataset listed every expulsion at every public school along with fields showing the student's grade, gender, date of birth and the reason for expulsion. Another tallied the number of suspensions by reason for each school.

We were going to be able to do exactly what I had hoped: provide parents with a school-by-school accounting of the discipline happening behind the local schoolhouse walls.

Within days, we gleaned from the database how many students were kicked out of each public school for bringing a handgun to campus, using drugs, alcohol or tobacco or just acting out. Then, we went to work reporting.

The data was easy to confirm and hard for school officials to dismiss. That's because it comes straight from the local principals' offices.

Annual reports

Each school year, every public school in Indiana must file paper reports with the Indiana Department of Education documenting their disciplinary actions – some of which is tallied for an annual report the states must submit to the U.S. Department of Education under the Gun-Free Schools Act.

The paper reports are certainly public records. Better yet, Indiana keys the information into a database. Other states probably do the same.

In our analysis and reporting, we found some weaknesses in the data. First, the tallies for each school only showed how many times students got caught and punished for bring-

---

CRIME AND COURTS
Conviction rate

By Bill Wallace
San Francisco Chronicle

More than a year ago, a source in the San Francisco District Attorney's office told me the city's new DA, Terence Hallinan, was plea-bargaining criminal cases at a disturbing pace, and many serious criminal charges were being dismissed in the process.

The tip was intriguing, but I needed to find some way of checking it out short of simply examining every case Hallinan's office had handled – a daunting task given the thousands of cases filed annually.

Late last year, I discovered a type of data called Offender Based Transaction Statistics (OBTS) that was used by the California Attorney General's office to generate a section of the state's annual report, "Crime and Delinquency in California." The information is known as "Adult Felony Arrest Dispositions," available from the Criminal Justice Statistics Center of the state Department of Justice.

Inside Uplink

SCHOOL VIOLENCE
Because of the media attention given to school violence this year, we decided to focus this issue on how computer-assisted reporting has played a role and offer advice for attempting similar stories in your communities.

Scott Smallwood of the *Albuquerque Journal* relates how he looked at violence on the campuses of Albuquerque, N.M., schools last spring. Both print and broadcast reporters share advice on how to deal with inconsistent reporting in school violence data.
See page three

FIRST VENTURE
Jodi Nirode, who covers the police beat for *The Columbus Dispatch*, shares how she learned that working with data isn't always the hardest part of a CAR story.
See page five

THE BUSH FILES
Aaron Diamant of the IRE and NICAR staff reports on how to convert George W. Bush's campaign contribution reports from the .pdf format on the Web to useable data.
See page seven

STAFF TRAVEL REPORTS
Derek Willis and Jackie Koszczuk of *Congressional Quarterly Weekly* tell how they uncovered a previously neglected area of reporting on Congress: expense-paid trips for key staff members.
See page nine

Campus crime

Continued from page one:

ing guns to school or breaking other rules.

There is no way to know how often offenses actually occurred on campus. Some schools may crack down on certain offenses while others turn a blind eye.

Second, like most crime data, you are going to find inconsistencies when you ask hundreds of school officials to lump incidents into broad categories based on vague definitions.

In Indiana, for example, one school confiscated what looked like a grenade from a student and promptly expelled him, state officials said. The grenade turned out to be a cigarette lighter, but the school reported the expulsion to the state under the category of "other firearms."

Beyond data

We overcame those weaknesses with reporting. We talked to state and local educators, students and others. We also carefully studied the paper reports that schools filed and the instructions they were given in filling out those forms.

> **As with most CAR stories, there was more traditional reporting involved here than number-crunching. The whole project took less than a week from the day I popped the floppy disk into my computer.**

All offered insight that helped us explain what conclusions people could and could not draw from the figures. Interviews with school principals gave us added opportunities to check our data against the paper reports they had filed with the state. We checked many schools' figures against individual reports provided by the state.

Along with our stories, we filed charts listing the expulsions and suspensions at every public school in Indiana for various offenses. That allowed local newspapers and broadcasters to do their own stories.

We filed the package within days of the federal report's release and just as Hoosier children were headed back to school.

Under-reporting

Along the way, we uncovered a problem with the national report: it was at least partly wrong. Indiana had reported 62 gun-related expulsions to the feds. But our database included 129 gun-related expulsions.

We rechecked our database work, and I called the state Department of Education official in charge of collecting and reporting this data and asked why the totals were different. He called back to say the state totals were right and the ones in the federal report were wrong.

We continued working on our stories, while the state education staff met to figure out what went wrong. The official called back to say the state had under-reported their figures and said, "We just have to own up to this one."

Education officials in other states reported discrepancies in gun expulsion figures too, and the AP moved a national story highlighting the national report's errors.

As with most CAR stories, there was more traditional reporting involved here than number-crunching. The whole project took less than a week from the day I popped the floppy disk into my computer.

We got the data for free, but Indiana's education department is accustomed to forking over large databases to us for nothing. Once we had the database, the computer work was no more sophisticated than importing a text file into Microsoft Access and running a half dozen sum and count queries.

Reporters in other states, which collect and compile the same data from local schools, could easily replicate the reporting.

John Kelly may be reached by e-mail at
jkelly@ap.org

Columbine aftermath

By Scott Smallwood
Albuquerque Journal

A few days after the Columbine High School (Colo.) shooting last April, we asked our local school district for statistics about violence at individual schools around Albuquerque. At first, we were planning to turn around a simple daily, but the district's foot-dragging eventually prompted us to turn the story into a small project.

Officials first told us that the district had never tallied any of its crime and violence data by school. They said they didn't want to embarrass principals of the more crime-ridden schools. Instead, each year the district sends the state a general report listing the total number of assaults, gun violations, drug charges and incidents of vandalism. We were welcome to that, they said.

We imagined that in Albuquerque Public Schools, a district of 120 schools with 85,000 students, district-wide numbers probably told just a small part of the story. We asked the district for its database of criminal reports. (The district has its own police force that patrols the elementary schools and middle schools. City police monitor the high schools.) They wouldn't give it to us electronically and our state law seemed to support them on that.

Starting from paper

They did give us paper printouts of three years of data, which we were able to scan in. Using a fax server that's generally used to scan classified ads and death notices, we were able to quickly scan dozens of pages at one time. Luckily, the paper reports were very clean and a couple of Excel macros later, we had recreated their database.

We then went to the city police to get records of every criminal report filed at the city's 11 high schools for the same time period. We were the first to combine the two data sets – a total of 13,000 criminal reports over three years – to get a more complete picture of crime in city schools.

While district officials knew of most gun violations in the schools and correctly reported those to the state, they didn't have any record of most of the city police reports, for everything from aggravated assault to rape.

We found that the district's annual reports underestimated nearly every category, especially assaults and drug violations. We also found, to no surprise, that in a district with 11 high schools and about 30 middle schools, reports of crime and violence weren't spread evenly across the city. Some schools reported four or five times the number of incidents as others.

The superintendent acknowledged that principals and parents should have this information broken down by school site.

Afterthoughts

There were a number of things we probably should have done differently.

We didn't realize going in that the underreporting would be such a large part of the story. We should have gone back to the city and requested reports for the three years from the 79 elementary schools and 30 or so middle schools, but by that time we were well into the project and wanted to get it into the paper. Thus, we focused mostly on high schools.

Also, we probably should have fought longer on getting the data electronically, despite the poorly worded state law.

> ## We found that the district's annual reports underestimated nearly every category, especially assaults and drug violations.

The stories appeared in the paper a month after we first asked for the violence statistics. About two weeks were wasted negotiating with the district about the data. Once we had created the database, we spent a few days using Access and Excel to analyze the data and another week reporting and writing the stories. We combined the crime data with other school data in our files, including drop-

A few of the stories on school violence at the IRE Resource Center:

After the string of multiple school shootings in the early and mid '90s, Education Week reporter Jessica Portner examined easy access to firearms as one of the main causes in explaining why one student killed another on the steps of a Savannah, Ga., school (File 14853).

Jeff Meader, an Education Week journalist, reported in November 1998 that "during the months that Americans were transfixed by the tales of students killing other students, there were nearly as many stories – if not more – of children setting fire to their schools." (File 14883).

A 1998 story by Royal Ford, a *Boston Globe* staff writer, examines possible causes of the increasingly violent nature of teenage girls, as well as solutions such as neighborhood and school programs. (File 14905).

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

Inconsistent data

By MaryJo Sylwester
IRE and NICAR staff

Reporters across the country who have looked at school violence statistics say there is one big problem with the data: inconsistent reporting.

In Raleigh, N.C., WRAL-TV reporter Stuart Watson found cases of sexual harassment reported as sexual assault and bullets reported as guns.

In Virginia, the state's first-ever "report card" had numbers that seemed to defy reality for three school districts of roughly the same size, according to Jeff South, a journalism professor at Virginia Commonwealth University. Here's what the state said: City of Richmond, 188 reported incidents; Henrico County, a suburb, 859 incidents; and Chesterfield County, a suburb, 11 incidents.

The *Richmond Times Dispatch* published articles at the beginning of April telling the public about the discrepancies in the data.

"Educators have cautioned that some of the numbers — particularly those that record incidents of violence — are subject to the interpretation of the schools reporting them and may not be reliable as points of comparison from school to school and division to division," one of the *Times Dispatch* articles said.

So how do you overcome this problem?

Several reporters from around the country — both print and broadcast — chimed in with some suggestions, based on their own experiences. What follows is an overview of what they had to say.

The first step is to study the data, looking for "numbers that seem to defy reality," and ask for written guidelines on how school officials defined or categorized violent incidents, South recommended.

Watson suggested running an internal check by comparing school data against alternative sources such as police logs or reports and other school reports. Even if the school has its own police force, city police or sheriff's deputies may also take reports of criminal activity at schools.

Be sure to ask data collection or analysis people to explain any discrepancies and call the schools directly to double check the numbers.

"We found keypunch errors and principals who did not read definitions or follow simple instructions," Watson said.

Scott Smallwood of the *Albuquerque Journal*, said it's important to ask how crimes are reported — whether the principal tells someone or students go directly to officers on campus.

He also found those schools with more serious crime and drug problems tended to be more diligent in reporting.

"(Under-reporting) is probably more likely with the lesser incidents — from vandalism to simple assault to disorderly conduct. Our feeling was it would be harder to keep quiet about robberies, rapes and serious assaults with officers patrolling the campuses," he said.

But ultimately the only way to overcome discrepancies in the data is an old-fashioned technique called interviewing people. Talk to students, educators, police officers, state officials and anyone else who can shed some light on the "reality" behind the numbers.

And finally, be extremely careful in how you report the numbers. Be sure to clarify any discrepancies and throw out any numbers or comparisons that you can't verify as accurate.

MaryJo Sylwester can be reached by e-mail at maryjo@nicar.org

More resources on school violence:

National Criminal Justice Reference Service (*www.ncjrs.org/ojjhome.htm*) — Here you can access reports generated by the National Juvenile Court Data Archive, which collects, stores, and analyzes data about young people referred to U.S. courts for delinquency and status offenses. Their data include demographic information about the juveniles, the reasons for their referral to court and the court's response.

A second NCJRS site (*www.ncjrs.org/jjvict.htm*) offers a long list of available reports and links to other sites regarding juvenile justice, including school violence.

National Center for Education Statistics (*www.nces.ed.gov*) includes reports and statistics, such as:

• Violence and Discipline Problems in U.S. Public Schools, 1996-1997 (released March 1998)

• Digest of Education Statistics, 1998, also includes some stats on school violence.

School violence resources:

Bureau of Justice Statistics Clearinghouse (*www.ojp.usdoj.gov/bjs/abstract/iscs98.htm*)
• Includes the Annual Report on School Safety 1998, which has national school crime data, successful strategies to reduce school violence and a local safety checklist for parents.

• Also includes Indicators of School Crime and Safety 1998, which presents data on crime at school from the perspectives of students, teachers, principals and the general population using a comprehensive array of surveys as sources.

Learning the hard part

By Jodi Nirode
The Columbus Dispatch

I was so idealistic. After finishing IRE and NICAR boot camp this summer I thought the hardest part about doing my first solo computer-assisted reporting project would be writing Access queries. How quickly I learned how many pitfalls there can be along the way.

The project was looking at the response times of Columbus police.

On Mother's Day, a 48-year-old man was shot and killed in front of his children and grandchildren – 30 minutes after a neighbor warned a police dispatcher that a drunk man was firing shots from his front stoop.

Police called the delay a fluke – that rare case when more important calls bogged them down from getting to a run that should have been the division's second highest priority.

The Dispatch wanted to know for sure.

I requested a year's worth of the 911 dispatch records. In my request, I stated I wanted nothing omitted and I wanted them in a comma-delimited text format. Simple, right?

> **I should have asked for them to give it to me in Access originally. I just never dreamed they had it. It taught me the first of many valuable lessons that would resurface later – never assume.**

However, the woman placed in charge of my request gave me separate files for each day – 366 in all. So I'd have to input and link each file for every Access query.

The dispatchers comments weren't included, I later learned, because they are stored separately.

So, I had to make the request again.

And again, the files came back in separate databases.

The head of the department gave me a technical argument about the information being too large to store in one text file; if I wanted it in one file, he could do it, but in Access only.

God, did I feel dumb.

I should have asked for them to give it to me in Access originally. I just never dreamed they had it. It taught me the first of many valuable lessons that would resurface later – never assume.

The data work

Alas, finally after about a month of waiting and re-requesting, I had my data. I was so excited. I signed out a new IBM Winbook and interviewed the 897,870 records or calls that Columbus police took that year. (By the way, I was defining response time as the time it took between the time the call was dispatched to when the first police cruiser arrived on the scene.)

I queried the data every which way. Average response times for gun runs, domestic calls, dead bodies, burglaries, stabbings – you name it.

I looked at medians. Then I looked at response times by shift and precinct using Excel's pivot table. (What a marvelous invention.)

I knew going into the project that nearly 50 percent of police in Columbus weren't logging their arrival times. It was disappointing, but we planned to use the data anyway and note the problem.

As I saw it, the information I had to use was the only information the division had to make important decisions such as staff allocations, requests for more manpower and citizen complaints. The response times might not be dead on, but I thought they would still show disparities in response times between the larger and smaller precincts. (Of course, with the help of data processing I did some more sophisticated queries first to see if the lack of reporting was similar across the city's 19 precincts and it remained fairly constant for all types of calls.)

What I didn't anticipate was a programming quirk that plugs in the same time for

First venture

when the call was received, dispatched and when the officer arrived.

The self-initiated run, as Columbus police call it, is supposed to be used when an officer witnesses a crime or calls the run into a dispatcher shortly after it happens.

Those runs, though, comprised 28.9 percent and couldn't be used. Factoring those out, we had about 21 percent of the total calls to use.

Three independent statisticians assured us the data would still be fine to use if we stipulated that the averages were averages of the data available – and not averages of the division as a whole.

A new direction

As controversial as it might have been, the thought of using it didn't last long. Doing queries on the data without the self-initiated runs showed how skewed the data can get when you start slicing it too much.

Response times for officer in trouble – a call officers use when their life or well-being is in danger – jumped to 8 minutes. (Without self-initiated runs, the average was about 3 minutes.) I knew the 8 minute-average was not only absurd, it was false.

Though a young reporter, I've covered police on and off for six years and knew that was the call that police likely responded to the quickest.·

Going back to police communications I got the answer I dreaded. The 911 system plugs in arrival times for the first car on scene who logs his arrival time.

So for example, if an officer shows up an hour later to handle traffic and is the first on scene to hit a button on his cruiser console signifying his arrival, that's the time the 911 data shows.

I knew then, the project as planned was dead.

Of course, there was still a very important story: that the division has no idea how long they take to get to crime scenes. They can't.

Of course, in retrospect I felt dumb for assuming the arrival time listed in the database was for the same officer listed on the department's run. Next time, I'll be very thorough in finding out how information gets plugged into each column.

In the end, the story prompted outrage by the City Safety Director and others, but no change.

The director ordered the officers be forced to record their arrival times, but the police contract has a provision that the city can't alter any past practices, unless changed in the contract.

And the Fraternal Order of Police didn't see the lack of response times as a problem, so the director's order was moot.

The division is caught up in a battle too with the U.S. Department of Justice over claims that they abuse citizen's rights, so even with follow-ups, the issue is still likely going to be lost in the shuffle.

It was a lot of work to uncover a record-keeping problem, but I sure learned a lot along the way.

Jodi Nirode can be reached by e-mail at jnirode@dispatch.com

Aftermath

out rates and test scores, to examine any possible correlation. Those went into a sidebar.

Our immediate editors were helpful in freeing up some time for us to finish the project. But because the newsroom doesn't have a long history with computer-assisted reporting, others thought the stories would be ready the day after the school district gave us 400 pages of printed reports.

A few weeks ago, the district released its newest violence and vandalism report. A new state law requires them to break it down by school site and for the first time officials gave individualized reports to each principal about crime and violence in his or her school. Also, after our fight in the spring, we were able to get the data electronically that same day.

Scott Smallwood can be reached by e-mail at ssmallwood@abqjournal.com

The Bush files

By Aaron Diamant
IRE and NICAR staff

Amidst a great deal of self-congratulatory hype, George W. Bush has made available on his Web site (*www.georgew bush.com*) a complete list of every single campaign contribution he has ever received. Sounds great, except it's nearly impossible to work with.

The problem is that the data are only available as one giant .pdf file (and I mean a giant 1,750 pages and counting as of this writing and that's only for the 3rd quarter of 1999). So although the data don't cost you a dime, you really can't do anything with them unless you have lots of time and specialized software to convert them into a workable format.

A few of you have written to the NICAR-L listserv in the last few months asking for help converting the monster .pdf file into a format that can be imported into Excel or Access.

And as the 2000 election gets closer, more of you will be asked to take a look at, and produce stories on, the Bush campaign finance data from your area.

The following are some methods others have used:

MacGyver method

This is the method used by reporters at *The New York Times.*

Step 1. Write two PostScript programs that go automatically to the Bush Web site. One version should read the page and download the .pdf files you need. However, you don't want to download all the files after the first time. A second version of the PostScript goes to the directory and checks for the most recent additions.

Step 2. Load the data into Adobe Exchange with a "Redwing" plug-in designed by Monarch. This extracts text from a .pdf file. Use the "Extract all to Monarch" command that parses the text. You may run into a problem with the multi-line format of the original data (sometimes there is a second line added for comments which can be lost). To solve it, go to step 3.

Step 3. Run a PostScript that checks for spaces at the beginning of lines. If there is a space, write the script such that it will save the comment as a separate field.

Step 4. Export as Tab Delimited.

Step 5. Load into Access.

Degree of difficulty: 9.5. Not recommended for anyone except the most serious programmers.

Mother Theresa method

For those of you who are fairly patient, this method may be for you. It is much simpler than the previous method, but there are more limitations.

There is an Acrobat Reader plug-in called Aerial, which a number of you may be familiar with. You can use Aerial to convert a .pdf to an .rtf document. This is a good tool for copying a table at once and pasting it into Excel.

It can convert the whole .pdf document at once, but is not infallible. This is where the Mother Theresa part comes in. Once you convert the .pdf to an .rtf document you can use a word processing program such as Microsoft Word to clean the data and convert it to the format of your choice for import into a spreadsheet or database.

You can download a trial version of Aerial from *www.ambia.com/aerial.htm*. For help using Aerial you can refer to an article written by IRE's Training Director Tom McGinty in the July/August edition of *Uplink.*

Degree of difficulty: 5.

Fed Ex method

If you need to get data from a .pdf file that's posted on the Web, you can send a message to pdf2txt@adobe.com with the URL of the .pdf in the body of the message. The contents of the .pdf should be returned to you as a text message. You should get a response in a few minutes (sometimes a lot longer for larger files like the Bush .pdf). But be warned, you are going to have to clean the data yourself.

Degree of difficulty: 4

Indiana Jones method

Indy was known for hunting for things that are already out there. Believe it or not there are actually a number of sites that

---

Bush data sources:

www.georgewbush.com – George W. Bush's Web site, where the public can access his campaign contribution reports in .pdf format.

www.ambia.com/aerial.htm – At this site, you can download a trial version of Aerial, a software program that can help translate .pdf documents.

pdf2txt@adobe.com – An e-mail address to Adobe; send .pdf files here to be returned as a text message.

www.tray.com – Web site of Public Disclosure, Inc., where electronically filed presidential data is easily downloaded.

www.foodnews.org/bush.html – Food News' searchable database of Bush's campaign finance reports; however, data can't be downloaded.

http://metalab.unc.edu/javafaq/bush – Elliotte Harold's Web site includes Bush's original data as tab-delimited documents, available for downloading.

Confidence rates

These are takes from a Sept. 30 discussion on the NICAR-L listerv regarding confidence levels used in Census Bureau data. See sidebar for information on how to subscribe to NICAR-L.

• No doubt some of you are working today with the Census Bureau's just-released income and poverty rates. Naturally, I'm looking at Colorado's numbers in comparison to the other states.

The state-by-state estimates of median income and of poverty rates are accompanied by measurements of standard error. Does anyone know the confidence level associated with these standard errors? The appendices dealing with how to work with standard error seem to indicate that the confidence interval, at least for U.S.-level estimates, is 90 percent. Does that same confidence interval hold for the state-level estimates?

A specific example: Colorado's three-year average median income is $44,349. The standard error is $1,075.

Would I, then, be correct in saying something like this: "The Census Bureau is 90 percent certain that Colorado's median annual income ranged between $43,274 and $45,424." ?

A corollary question: Which of the median income (or poverty rate) measures is best to use? The three-year averages? The most recent 2-year rolling average? – Jeff Thomas, *The Gazette*, Colorado Springs, Colo.

• The state-level data also is reported at the 90% c.l. [confidence level] This is noted in Table D in the footnote for the percent change column (those percent changes that are statistically significant). The Census release uses the three-year average for income (see the bulleted item on Alaska's income). This gives them a dartboard that's a bit larger. – Neill Borowski, *The Philadelphia Inquirer*, Philadelphia, Penn.

• The Census Bureau usually reports error margins at one standard error, which means the confidence level is 68 percent. I don't have the report to which you refer, so can't evaluate the indication that it is 90 percent. But for 90 percent confidence, according to sampling theory, you need an error band that is 1.65 standard errors wide. – Philip Meyer, University of North Carolina, Chapel Hill, N.C.

• This is a common mistake in interpreting confidence intervals. We cannot be certain that the true mean (or whatever parameter) falls within the confidence interval. The correct way to interpret a confidence interval is to say: "In repeated sampling, our range (confidence interval) would encompass the true mean 90% of the time." Thus, if we were to sample from a population 100 more times and determine the mean and respective intervals for each of those samples, then 90 of the intervals would contain the true population mean. – Mary A. Davis, Center for Environmental Studies, Virginia Commonwealth University, Richmond, Va.

Y2K on the Web

Below is an excerpt from a handout compiled by Russell Clemings of The Fresno Bee, for a regional Investigative Reporters and Editors conference in Los Angeles in September. A copy of the full tipsheet is available by calling the IRE Resource Center at (573) 882-3364 and asking for Tipsheet #1095.

Some good Y2K Web sites:

• President's Council on Year 2000 Conversion: *www.y2k.gov*

The most official of the official sites. Contains links to Y2K information by sector, although much of the material is out of date.

• U.S. Senate Special Committee on the Year 2000 Technology Problem: *www.senate.gov/~y2K*

The best and most authoritative of the congressional Y2K sites.

• Sanger's Review of Y2K News Reports: *www.sangersreview.com*

This used to be an excellent neutral source of Y2K information; lately, under a new author, it has tilted toward hysteria. But original editor Larry Sanger still checks in from time to time, most recently with a perceptive analysis of the Navalgate scare.

• Year 2000.com: *www.year2000.com*

Another archive of news clippings and other Y2K information. The section called "Bug Bytes" tracks reports of Y2K-related and Y2K-like failures.

Staff travel records

By Derek Willis and Jackie Koszczuk

Congressional Quarterly Weekly

In the past decade, Congress has tried to distance itself from some of the perks of lawmaking, especially gifts from lobbyists. The law on Capitol Hill is that members and staff members can't accept gifts such as meals and baseball tickets that exceed $100 a year per donor – a relatively paltry sum by most standards.

But that doesn't stop officials from taking trips to Hawaii, Switzerland and Las Vegas on the tabs of corporations and special interests who often have an interest in influencing legislation. An exemption in the law allows private travel as long as it is related to "official duties" and is not paid for by lobbyists.

> ## Travel records are stored in two places on Capitol Hill... Lawmakers and staffers fill out one of two basic forms, but early on we discovered that each body differed in its organization of the records.

In an effort to examine the link between these privately-financed trips and legislation, *Congressional Quarterly* looked at the travel records of leading members of Congress and their staffs. We found that private interests spent nearly $3 million on 2,042 trips for committee chairmen, ranking members and key staffers between January 1998 and May 1999.

How It Started

I had been looking for alternate sources of data on Congress, and while other organizations have examined travel records

before, most excluded staff travel from their analyses. We decided to include staffers because they often specialize in certain areas and because some top aides are nearly as powerful in shaping legislation as the members themselves.

Travel records are stored in two places on Capitol Hill: the Senate's Public Records office, in the Hart Office Building, and the House Legislative Resource Center, in the Cannon Office Building. Lawmakers and staffers fill out one of two basic forms, but early on we discovered that each body differed in its organization of the records. In the Senate, we obtained travel records for 1998 on microfilm for $20, but they were arranged alphabetically by the last name of the person who traveled. That made the process of ensuring that we had every record for each of the lawmakers that we wanted difficult.

The House did not offer microfilm but grouped records by member and included staffers with the member they worked for. Neither body had the records available electronically, and no plans are in place for that to happen, despite the fact that travelers often download the form off a House Web site and fill it out by hand. Instead of paying photocopying fees for most of the remaining records, I took a portable scanner to the Capitol and scanned in the forms.

We did the data entry ourselves, with help from two members of CQ's research staff. The research staff also helped fact-check a percentage of the records before publication to avoid any mistakes. Creating the Microsoft Access database of trips took about 5 weeks, although none of us worked on the project full-time during that period.

The data was fairly clean, although we encountered several forms filled out improperly and some that neglected to mention basic information about the trip, including the destination and time span. We answered those questions by contacting the traveler. After the input process, we added several fields to calculate the length of each trip and the total cost, since the forms break it down to several categories.

Reporting the Story

Analyzing the data was much faster: we were

For more information on lobbying, check out the following tipsheets, available from the IRE Resource Center at www.ire.org/resourcecenter

Tipsheet #923 "Lobbyists and local elections" from the 1999 National CAR Conference.

Tipsheet #827 "Uncovering the secrets of campaign finance and lobbyists" from the 1998 National CAR Conference.

Tipsheet #116 A one-page source list of government offices to contact for lobbying activities information.

Travel records

interested in seeing which lawmakers and their staffs traveled the most, or what foreign destination was most popular. For the story, we tried to focus on aspects of the trips that would link travel and legislation, and to describe details of junkets to Palm Springs, Calif., Las Vegas and Cape Cod, Mass.

We also produced several charts for the story that listed the leading destinations and top sponsors of travel, among other things. Internet search engines such as Northern Light *(www.nlsearch.com)* were helpful in gleaning more information about some of the leading sponsors, including Taiwan's main business lobby, the Chinese National Association of Industry & Commerce.

What we found was that some staffers spent more than 30 days on all-expenses-paid trips during the 17-month period, and that several groups had penchants for taking key members and their staffs to exotic destinations in the thick of winter: recycling seminars in Key West, aviation conferences in Hawaii and telecommunications roundtables in Palm Springs.

But the story also found that despite the ban on lobbyist-paid travel, most of the groups that sponsored trips concentrated on members who headed a committee or subcommittee with jurisdiction over their industries. The Nuclear Energy Institute, for example, paid for seven trips by Sen. Frank Murkowski, R-Alaska, and his staff. Murkowski heads the Senate Energy panel.

As with campaign finance stories, examples of direct quid pro quo, where a trip was traded in exchange for beneficial legislation, were scarce. Instead, we chose to shine a light on the loophole that allows organizations lobbying Congress to pay for members and staffers to travel for free, and show just how friendly the skies can be when you work on Capitol Hill.

Derek Willis can be reached by e-mail at dwillis@cq.com. Jackie Koszczuk can be reached by e-mail at jkoszczuk@cq.com

Bush files

have the Bush data available without doing all the grunt-work. It has all been done for you. Here are three possible options:

First, Tony Raymond of Public Disclosure, Inc. and the Web site *www.tray.com* has worked to put electronically filed presidential data in a Microsoft Access database or .dbf format that you can download right off the site. He has the data online shortly after they're filed with the Federal Elections Commission.

Another choice: check out *www.food news. org/bush.html.* This organization prints the data to a PostScript file from Adobe Exchange, then parses them based on the position in the file. The original dataset is not available here but you can search their database by name, address, city, date, occupation, employer, etc.

But the best place that I've found to get the original data is off of Elliotte Harold's Web site (*http://metalab.unc.edu/javafaq/bush*). Harold teaches courses in Java and object oriented programming in the Computer Science Department of Polytechnic University in Brooklyn, New York. He has extracted the ASCII text from those .pdf documents and posted them as tab-delimited documents that can be easily imported into any good spreadsheet or database.

Degree of difficulty: 2.

The fact that the data are so hard for the average journalist to deal with raises the question of the sincerity with which the Bush camp released them under the guise of full public disclosure. But as journalists, most of us enjoy a good challenge. And for that we thank you Mr. Bush, because most of us are paid by the hour.

Aaron Diamant may be reached by e-mail at adiamant@nicar.org

Economic census

By Neill A. Borowski
The Philadelphia Inquirer

"Imputed" and "estimated" should be two words that wave red flags and set off alarms for users of census and other government data. Those words mean the information in question didn't come directly from the source. In some cases, the data represent the agency's "best guess" in the absence of solid statistics.

The cautions of using estimated data are particularly true for the Census Bureau's Economic Census program (more on the program below), which this fall has been releasing data from the 1997 survey of establishments.

The Census Bureau's economic censuses, taken every five years in years that end in "2" or "7," are not full censuses. Questionnaires are sent to multi-unit and large single-unit businesses. But smaller firms aren't included in the survey.

When data are missing, the bureau will try to use administrative records from other federal agencies, such as the Social Security Administration or the Internal Revenue Service. And when even that information isn't available, it turns to imputation – estimations based on industry averages or historic company information.

An example of imputation gone wrong was apparent in September when the bureau released the 1997 Economic Census for Educational Services. This report focuses on proprietary schools of business, technology and cosmetology, among other subjects.

I was interested in building a table from this report for the counties in the Philadelphia metropolitan area. When I reached Montgomery County, which is Philadelphia's largest suburban county, the data leaped out of the page at me. Now, here was a great news story.

The census reported that total revenues from educational services firms were just under $300 million for Montgomery County – almost three times the total revenues for such firms in Philadelphia. The annual payroll of such firms was $108 million, again about three times that of Philadelphia. And the number of employees was 4,427, nearly four times that of Philadelphia.

Learning the truth

Such a good story. You almost hate to ruin it with a phone call to make sure the data are correct. But this report seemed far out of whack.

Indeed, the report was wrong – forcing the Census Bureau to not only recalculate the Montgomery County totals, but also the metro and state numbers. "We simply missed it (the error) during our analysis," said Barbara S. Tinari, a survey statistician on the economic census team. The formula used to estimate the Montgomery County numbers didn't work in this case.

Instead of $298 million in revenues, the figure should be about $75 million. Instead of $108 million in payroll, the figure should be about $25 million. And instead of 4,427 employees, the number should be about 1,500.

Tinari wasn't pleased about the error, but she was pleased to have it pointed out. The problem, she explained, was in the bureau's estimation procedure for Montgomery County. Nationally, only about 13.5 percent of the data for educational services came from estimates. However, 78 percent of the Montgomery County total was from estimates. (Only 11.7 percent of Philadelphia's data came from estimates.)

The lesson? Get to know the data and read the appendices to the report. They will let you know just how frail the data can be. The economic censuses are forthright about data from administrative records and estimation – a separate column lets the user know how much came from each source. For other types of government data, you may have to dig a bit deeper.

The program

Having cited the potential problems with imputed data, I want to also urge you to check out the economic censuses for your area. If the data don't spark solid business and economic stories, they should be great sources of graphics or supporting paragraphs for other stories.

The 1997 Economic Census page is at *www.census.gov/epcd/www/econ97.html.* You can track the roll-out schedule - each census release tends to begin with New England states and move across the nation to California – at *www.census.gov/epcd/www/ec97stat.htm.*

Here are Web addresses for the U.S. Census Bureau's Economic Census program referred to in Borowski's story:

1997 Economic Census page: **www.census.gov/epcd/www/econ97.html**

Roll-out schedule for each state: **www.census.gov/epcd/www/ec97stat.html**

1997 County Business Patterns reports: **www.census.gov/prod/www/abs/cbptotal.html**

Group Importing

This advice from David Milliron, CAR editor at The Atlanta Journal-Constitution, was posted on NICAR-L on Oct. 19.

How can you import a large group of files of the same type into an Access 97 database?

The Import Specification tool and the TransferText function (in the Macro tab section) is one way to accomplish this.

Let's say you receive weekly crime data in tab-delimited text files that you want to maintain in your "crime" database. You save the weekly crime.txt files to the "crime" folder on your "F" drive. (You can substitute any file name)

It may be easier to manually import the first week's data – where you can also use the "advanced settings" option to specify which fields are to be imported and any specialty data format, etc. Then save the "import specification" for future use.

To create a specification table for the "crime.txt" file:

1. Open your "crime" database.
2. On the File menu, click Get External Data, then select Import.
3. In the Import box, select Text Files under Files of type.
4. Select the crime.txt file from your F:\crime folder, then click Import.
5. Select delimited, then click Next.
6. In the Import Text Wizard box, select tab delimited (the first row of our data does not include field names).
7. In the Import Text Wizard box, click Advanced, make any necessary changes in the Field Information, then click Save As and accept the "Crime Import Specification" default name by selecting OK in the Save Import/Export Specification box.
8. Click Cancel in the "Crime Import Specification" box.
9. Click Cancel in the "Import Text Wizard" box.

You didn't import any data doing this, but you'll want to test your newly created "Crime Import Specification" file. Then:

1. Open your "crime" database.
2. On the File menu, click Get External Data, then select Import.
3. In the Import box, select Text Files under Files of type.
4. Select the crime.txt file from your F:\crime folder, then click Import.
5. Select Advanced.
6. Select Specs, then choose the "Crime Import Specification" import specification you created above. Click Open, click OK, click Finished.

Now you're ready to create a macro and put the "TransferText" command to use.

1. In the Crime database, click the MACRO tab and select New.
2. Under Action choose TransferText.
3. At the bottom of the screen in the Action Arguments section:
A. In the Transfer Type section, select Import Delimited.
B. In the Specification Name section (the specification name for the set of op tions that determines how a text file is imported, exported, or linked) select the "Crime Import Specification" import selection criteria you created earlier.
C. In the Table Name section, enter Crime (see assumptions above)
D. In the File Name section, enter *f:\crime\crime.txt* (see assumptions above)
E. In the Has Field Names section, select No.
F. Save the Macro as Update Crime Data and you're ready to append future file updates.

Note: All files must be the same name each week.

Conviction rate

The OBTS data included the total number of felony arrests for a given county, the number of cases dropped by law enforcement agencies, the number of cases declined by county prosecutors, total complaints actually filed and their outcomes – conviction, acquittal, dismissal, non-judicial diversion and prison or jail sentence.

Every California county has contributed to the dataset for more than two decades, and the federal Bureau of Criminal Statistics in Washington, D.C. considers California's data to be highly reliable.

Because I wanted to calculate and compare conviction, dismissal and diversion rates, the OBTS dataset was ideal for my study of the effectiveness of the San Francisco DA's office.

I contacted the state Attorney General and requested OBTS data for San Francisco County from 1988 through 1997 (the last year for which data was available at the time of the request). The reason for this historical spread was to compare DA Hallinan's new administration (1996-1999) with his predecessor. Eventually, I requested three years of identical OBTS data for all the Bay Area Counties, and for the entire state of California. When 1998 statistics became available this summer, I also requested those for all California counties.

Studying the data

I began receiving the data in the form of unprocessed hard copy spreadsheets within a couple of weeks, and immediately sat down to examine it using lead pencils, a ruler, a calculator and a pad of analyst's paper. After I was satisfied that I understood the relationships between the various subsets of numbers for each county, I used Microsoft Excel to construct computer spreadsheets, calculate various rates and make tables that ranked the results.

Because I was dealing with about 30 data fields for 58 counties for each year in my spreads, I hand entered the numbers directly from the DOJ spreadsheets, then cross checked them for accuracy before setting up my formula cells for calculation. The process took only a few hours at each phase of the project.

I used the same techniques for determining conviction rates that are used by the state Attorney General: dividing convictions by total arrests to obtain the rate as a percentage of the total universe of possible cases in each jurisdiction, and dividing convictions by complaints filed to obtain the conviction rate as the percentage of cases each DA considered worthy of pursuing.

I determined "nol pros" (decline to file) rates by dividing the complaints denied by the total number of arrests. I also calculated dismissal and diversion rates by the two methods I had used for convictions.

Finally, I corroborated the trends my statistical information seemed to indicate by looking at two independent and unrelated datasets. One was the number of cases that had gone to trial in San Francisco over the last seven years from the California Judicial Council. The other was the California Department of Correction's data on the total number of convicted criminals sent to state prison over the last seven years by authorities in San Francisco.

Behind the trends

All three datasets verified my original source's tip: conviction rates in San Francisco had dropped under Hallinan's administration while dismissals and refused cases had climbed steeply. The number of cases going to trial under the new DA was steadily decreasing, as was the number of cases resulting in prison sentences.

All of these previously unreported trends were newsworthy, but I still wasn't satisfied that I understood why they were happening. In order to find out, I examined 100 felony criminal files from San Francisco Superior Court just to see what was actually happening in a substantial – but still manageable – number of cases. I made no attempt to pull a scientific sample but simply pulled cases at random from the alphabetic court index by the criminal code section in the leading charge (187 for homicide, 245 for assault, 11,350 for drug possession, etc.)

This review, which I also summarized in Excel spreadsheet format, showed that most cases, including violent crimes, were indeed being plea-bargained, despite Hallinan's

Conviction rate

pledge to reduce plea bargaining "to zero or nearly zero" in violent crime cases. In the process, a substantial number of criminal charges were being dropped.

The case survey also showed that many defendants were being sentenced to brief jail terms and probation – and that a large proportion of those who received probation were rearrested again within months for new offenses.

I finished off the study with a 90-minute interview with Hallinan in which I queried him at length about this statistical picture of his office's performance and about specific cases that I had uncovered. The result was a package of stories, charts and photographs that ran on page one of the *Chronicle* and broke to two full inside pages.

Story response

In response to the package, Hallinan and his supporters argued three major points: First, that his conviction rate only appears low because his staff has to weed out many bad arrests by the San Francisco police, while other California police agencies themselves kick out cases without sending them to the district attorney's office. Second, that San Francisco sends many defendants to diversion programs instead of jail or prison; using non-judicial diversion is a progressive but non-traditional approach to justice that has the effect of reducing a prosecutor's conviction rate, since diversions are officially considered dismissals.

> **The number of cases going to trial under the new DA was steadily decreasing, as was the number of cases resulting in prison sentences.**

Finally, they said Hallinan's approach toward criminal cases had proven itself effective because reported crimes had dropped dramatically in San Francisco during his administration.

Because neither Hallinan nor his supporters understood the state data, all three arguments were completely off point. Cases thrown out by police agencies before review by county prosecutors accounted for only 4 percent of the total arrests in California in 1998. Police in the three California counties that Hallinan pointed to as typical, Los Angeles, Alameda and San Diego, dropped the vast majority of those cases. However, despite the police drops, prosecutors in those three counties still filed complaints in a larger percentage of arrests than Hallinan.

More importantly, all three had a significantly higher conviction rate than Hallinan after weak cases were weeded out – regardless of whether police or prosecutors had done the weeding.

Second, the number of cases sent to diversion by Hallinan – although higher than in any other county in California – were only 636 out of 16,614 arrests, less than four percent of the total. This minuscule percentage did not begin to account for Hallinan's low conviction rate or huge 32 percent dismissal rate.

Finally, while it is true that crime has dropped substantially in San Francisco in recent years, crime reports have dropped in most U.S. jurisdictions during the same period. In addition, the crime categories in which San Francisco's largest decreases have occurred are offenses like murder and rape that occur in relatively small numbers in the first place.

In those categories, even a modest numerical decline yields a whopping percentage decrease. In addition, those large decreases began before Hallinan took office, so it is more likely that they are related to shifting U.S. demographic patterns than Hallinan's law enforcement strategies.

In the month since the story first appeared, neither Hallinan nor his supporters have found one error of fact in the *Chronicle's* study. As the district attorney himself said in responding to the story, "the proof of the pudding is in the eating."

Bill Wallace can be reached by e-mail at wallaceb@sfgate.com

Economic census

As of mid-October, for example, the Educational Services and the Health Care and Social Assistance reports had been released for each state.

One advantage to the Economic Censuses is that they go beyond the county level and offer some data on larger cities and townships. The other major census business product, County Business Patterns, doesn't go below the county level. However, the advantage to County Business Patterns is that it comes out every year. The 1997 County Business Patterns reports were released in early October (available at *www.census.gov/prod/www/abs/cbptotal.html*).

The Economic Census also reports on total receipts (revenues) for each sector. Both the Economic Census and County Business Patterns report on total establishments, employees and payroll in each sector.

Other 1997 Economic Censuses that will come out over the next two years include Wholesale Trade; Retail Trade; Real Estate and Rental and Leasing; Professional, Scientific and Technical Services; Administrative and Support and Waste Management and Remediation Services; Arts, Entertainment and Recreation and Accommodation and Foodservices.

Yet another wrinkle in the Economic Census is the move to the new North American Industry Classification System (NAICS), which is a revision of the Standard Industrial Classification (SIC) system of separating industries. The change makes comparisons of industries over time a problem, if not impossible for some sectors.

As the bureau notes: "The implementation of NAICS will cause major disruptions in the availability of comparable information across time periods. In the last 30 years, the SIC system was updated three times...and each time a significant number of new industries was introduced into the existing framework. What is different for 1997 is that the whole framework has changed."

The new NAICS structure, however, does make some of the data richer. For example, under the SIC system, hotels and motels were classified as 7011. Under NAICS, there are separate listings for casino hotels and bed-and-breakfast inns.

Each geographic report is an Acrobat .pdf file. The Aerial plug-in and Acrobat Exchange easily turned the tables into spreadsheets. You also can purchase the data on a series of CD-ROMs.

Neill Borowski can be reached by e-mail at nborowski@phillynews.com.

Conviction data advice

Here are some suggestions from Bill Wallace for others attempting to look at conviction rates:

First, be sure you understand the relationship between the various subsets of numbers in the statistical data you obtain, and use only stats that are directly comparable to each other. In other words, use numbers that measure the same things and don't mix them up.

The key is understanding the subsets of numbers you are working with. If you are completely clear on what it is that each section of your data actually represents, you can't go wrong. A good tip is to take a single column of data and use simple arithmetic to clearly understand how each subset relates to the total.

Second, look for other sorts of statistical information from unrelated sources – like data on the number of prison remands per year, jury trials versus pleas, the number of felony cases versus misdemeanors, etc. – that can be used to cross check your primary data or illuminate aspects of it.

Third, be ready to test any conclusions you draw about how cases are being handled in your jurisdiction by going down to the courthouse and looking at lots of cases. Case reviews will provide much of the detail and color you need to flesh out a largely statistical story, and will also give you information about the way cases are actually being processed in the criminal justice system.

Finally, never hesitate to discuss your approach and findings with people who are knowledgeable about how the system works. I got some of my most valuable guidance by discussing my project with law professors, trial lawyers and statisticians and sharing my findings with them as I went along. No doubt this prevented me from making serious errors of fact and helped me to more fully understand what my data were showing me.

Bits, Bytes and Barks

Upcoming Boot camps

Dates for Computer-Assisted Reporting Boot camps next year have been set.

More information, including registration forms, costs, a typical schedule and travel and lodging options, is available at the NICAR Web site, *www.nicar.org*. The boot camps are held at the Missouri School of Journalism in Columbia, Mo., with hands-on training from NICAR staff members.

The dates for the three sessions are:

January 5-9, 2000
March 26-31, 2000
May 7-12, 2000

There are several fellowships and scholarships available for boot camps. Check out NICAR's Web site for more information. Applications can be obtained from John Green, membership coordinator, jgreen@ire.org, (573) 882-2772.

Uplink Survey

If you haven't already filled out the *Uplink* survey posted on the NICAR Web site, please do so in the next couple weeks. We'll be compiling results after Nov. 15 and we'd like your opinions included. We'll be using the results of the survey to better tailor *Uplink*'s content to your needs.

The survey can be accessed at *www.nicar.org*

SAS and SPSS

CAR specialists who use SAS and SPSS software for data management and statistical analysis — and those who want to learn more about these high-end programs — are invited to join the new SASCAR-L e-mail list recently created by Steve Doig of Arizona State University.

SASCAR-L is for reporters who need specific technical help with those programs. More general questions about computer-assisted reporting, including help with statistical techniques, still should be posted to NICAR-L.

To join SASCAR-L, send this message — SUBSCRIBE SASCAR-L firstname lastname — to this address: listserv@asu.edu. You can also visit the SASCAR website at *http://cronkite.pp.asu.edu/sascar/* for more information and tips.

IRE Bookstore

Just getting started with CAR? Check out Computer-Assisted Reporting: A Practical Guide, by Brant Houston. It can be ordered from IRE and NICAR for $25 for IRE members or $30 for non-members plus shipping. Call (573) 882-2042 to order.

Other books available from the IRE Bookstore:

• 1998 Uplink Collection, a bound edition of the 1998 issues of Uplink; $35 for IRE members, $50 for non-members.

• 100 Computer-Assisted Investigations; $20 for IRE members, $25 for non-members.

• A Journalist's Guide to the Internet, by Christopher Callahan; $17 for IRE members, $21 for non-members.

• Find it Online: The Complete Guide to Online Research, by Alan M. Schlein; $16 for IRE members, $20 for non-members.