Uplink

June 1996

A newsletter for the National Institute for Computer-Assisted Reporting

Oops!

Uplink update

To err is human; to forgive, divine; and to print a correction, humiliating.

While computer-assisted reporting allows us to write industrial-strength stories loaded with facts and details, it also lets us make eye-popping errors at light speed.

In this issue of Uplink, reporters share their own CAR mistakes in hopes of preventing the rest of us from repeating their errors. Plus, NICAR's Brant Houston and the AP's Drew Sullivan share tips on performing integrity checks that may head off data catastrophes.

Also in this issue, Penny Loeb describes how *U.S. News and World Report* used train data to examine railroad safety problems around the country, NICAR's Andy Lehren reveals how reporters exploited FAA data after the ValuJet DC-9 crash, and David Milliron of Gannett News Service shares some handy Web sites.

Inside

The perils of computer-assisted reporting

Errors can wreck stories

By Brant Houston
NICAR managing director

The recent correction of a page-one story by the *Atlanta Constitution Journal* was a nasty reminder of how easy it is to make a critical error when doing computer-assisted reporting.

The newspaper reported that Atlanta, on the eve of the Olympics, has the highest violent crime rate among United States cities. In fact, Atlanta is in the Top 5, but it is not No. 1. The newspaper accidentally left out several major cities when typing in data.

This certainly is not the first mistake made while doing a CAR story (or any news story). Mistakes have been made since journalists began using software to collect and analyze data. Some mistakes have made it to print or on air. Many did not because the journalists were aware where things could go awry and because they followed rudimentary guidelines for preventing errors.

Data snafus

To get an idea of some of the data perils out there, NICAR recently asked journalists for examples of errors and snafus. Here are a few of them:

- The *Miami Herald* was examining the sentencing patterns for drunken drivers, using court files. The data showed a category for jail time and a category for a fine. In looking at the data, the *Herald* deduced that a small percent of drunken drivers received no punishment.

In fact, judges always gave some kind of punishment, including community service. But there was no field in the data for community service. The clerks using the data knew "0" for jail time and "0" for fine meant community service, but no one put the information in the code book for the database, and no one told the *Herald* about the lack of codes.

- In another story, the *Herald* examined problems in local schools and in-

Continued on page eight

Tech Tip

Checking data integrity

By Drew Sullivan
Associated Press

A pair of mantras that NICAR's Richard Mullins teaches his students at the Missouri School of Journalism is: "All databases are bad," and "All data is dirty."

If you've worked with many databases, you soon realize the truth in these statements. Underpaid and overworked government data-entry clerks will forever confuse similar postal codes, such as Missouri (MO) and Montana (MT) or Connecticut (CT) and Colorado (CO). And even the most clever database designers cannot account for incomplete

Continued on page six

Uplink
June 1996
Volume 8, Number 6
A newsletter for the
National Institute for
Computer-Assisted
Reporting

Editor
Brant Houston
brant@nicar.org
Managing Editor
Jo Craven
jo@nicar.org
Assistant Managing Editor
Andrew Lehren
andy@nicar.org
Senior Contributing Editors
Richard Mullins
richard@nicar.org
Rosemary Armao
jourrca@muccmail.missouri.edu
Staff
Wendy Charron
Seth Hemmelgarn
Neil Reisner
John Sullivan
Wallace Winfrey
Copy Editor
Marilyn Lehren

Uplink is published every month by the National Institute for Computer-Assisted Reporting, 138 Neff Hall Annex Columbia, MO 65211. (573) 882-0684. Subscription price is $35 (U.S.) annually, $50 outside the United States. Postmaster: Please send address changes to NICAR. Send e-mail to nicar@muccmail.missouri.edu.

NICAR is a joint effort of Investigative Reporters and Editors and the University of Missouri School of Journalism. NICAR services include hands-on newsroom training on computer-assisted reporting, special academic and advanced training in data analysis. NICAR is supported by a grant from The Freedom Forum and other foundations intended to help the institute deliver its services nationwide to news organizations and associations.

Data help stories soar

Crash course on plane crashes

By Andy Lehren
NICAR staff

When the ValuJet DC-9 crashed near Miami, many reporters poured through airline data to find out what went wrong.

A common step was taking the airplane's tail number and running it through key Federal Aviation Administration databases, such as Service Difficulty Reports and Accidents and Incidents. Some went further, using the airplane's serial number and studying its history before ValuJet began using the jet.

Here are some lessons from reporters who mined data for covering the crash:

- Be prepared. The *Cleveland Plain Dealer's* Beth Marchak had been reporting about ValuJet's problems before the crash. She began by sorting SDR data by date, by tail number, by serial number and by the troubled part of the plane. She checked for structural work. And she used her search results as a springboard for additional reporting. SDRs helped prepare her for interviewing passengers and airline workers. They also helped her craft Freedom of Information requests for FAA documents. She said paper records were vital for showing the depth of ValuJet's problems.

- Watch out for dirty data. For instance, in the FAA Accidents and Incidents database, Delta had mistakenly reported the fatal DC-9's serial number as its tail number. Severity codes in the SDR database are inconsistent. The FAA violates its own rules for the key fields in the SDR table. Operator codes in the Accidents database are often missing.

- Beware of how information is entered into the database. The SDR database allows the airline to report one character for the precautionary procedure taken if the plane was in the air. The *St. Paul Pioneer Press'* Dan Browning found Northwest often did not use the code for dumping fuel. But the airline routinely dumped fuel when facing a problem in the air. The *Seattle Times's* Bryan Acohido reported about rudder problems in Boeing 737s and found those difficulties were not always coded the same way (see Uplink, February 1995).

- Look how often an airline submits reports. They are supposed to report any mechanical difficulty. They often don't. Airlines report at different rates. Browning said he tries to adjust for reporting rates. He also looks at what the airline owns by using the FAA's Annual Inventory of Airframes and Aircraft Engines by Carrier.

But the problem raises questions about how the agency works. "Why does the FAA tolerate the disparity?" Browning asked. The University of North Carolina's Phil Meyer said that, with this problem, a lack of reports may well tip off a reporter to probe into an airline.

- Look at other databases. The ValuJet crash was linked with a hazardous chemical. The U.S. Department of Transportation's Hazardous Materials database includes airline accidents that involve hazardous chemicals. The *Miami Herald* used this information to put the crash into context, and show how rarely such accidents happen. It also gave readers a chronology of earlier accidents involving such substances.

Dateline NBC's David Hinchman used the FAA's Enforcement Information System to examine lax enforcement in commuter airline safety (see Uplink, April 1996).

CBS's Roberta Baskin recommended using NASA's Aviation Safety Reporting System, where pilots make anonymous reports on problems. By using dates, she has been able in other cases to strip away that anonymity and learn more about an incident.

The FAA's collection of advisories showed the agency was already concerned about bad wiring in the cockpits of DC-9s. National Transportation Safety Board records are often valuable for understanding airline safety, said the *New York Times'* Steve Engleberg.

- Keep in mind that human errors play a big role in crashes. The *Minneapolis Star Tribune's* Greg Gordon recommended that, along with other databases, include a look at FAA records for runway incursions (when mistakes happen on landings and takeoffs), pilot diversions (when pilots make mistakes), and near mid-air collisions. These can provide great anecdotal information.

- Look overseas. *Newsday's* Ford Fessenden used Aviation Information Services Ltd. and Flight International, both in the United Kingdom, to probe airline accidents worldwide. For tracking who owns planes outside the United States, he uses a book called the Jet Airliner Production

Continued on page five

Indices illuminate data

Playing by the numbers

By Andy Lehren
NICAR staff

The holy grail in aviation data is figuring out what predicts accidents.

What causes plane crashes: The airline? The type of aircraft? Pilot mistakes? Poor maintenance? Violent weather? How can reporters sift through patterns when so much changes over time — airlines change management, the aircraft they fly, and even owners.

After the ValuJet crash, a challenge remains in trying to develop an airline safety index. And even if this is not successful, University of North Carolina Professor Phil Meyer, author of "The New Precision Journalism," said this problem sheds light on how to think about developing an index.

Indices are no stranger to news pages. Look at the Dow Jones Industrial Average or the Standard & Poor's 500. Both are used daily to reflect the stock market's health. The *Philadelphia Inquirer's* Neill Borowski wrote in April's Uplink how he indexed various local economic segments against the nation. That helped him determine the Philadelphia area's strengths and weaknesses. Baseball fans track how many games back a team is from the leaders.

More than ranking

The idea is that these numbers do more than rank things. They are barometers. They show relative strength or weakness. A baseball fan knows more than whether a team is in second place; games back tell whether the team is close or far from the division leader. The S&P 500 indicates how much better or worse the stock market is doing. The index used by Borowski showed how much the local economy's makeup is different.

At *Newsday*, Ford Fessenden had previously looked at accidents worldwide to show a history of airline crashes, including predecessors in the numbers for the airlines that took them over, and included information such as the odds of a plane crashing. That, he said, showed readers that foreign carriers in certain countries have troubling accident rates.

The stories also underscored how human mistakes — and companies with a culture that did not emphasize safety and training — play a great role in crashes.

Meyer suggested reporters can try to go one step further. The goal would be to create a safety index for airlines, gauging the likelihood that someone would get killed riding a particular airline. It would include an analysis looking at whether factors in older crashes predict more recent accidents.

Drawing from the data

An index would use one or more factors for measuring safety. Journalists look to Federal Aviation Administration data such as Service Difficulty Reports, which include equipment problems, engine flame outs and emergency landings. They also look at the FAA's Accidents and Incidents reports, which include when people are hurt or killed, or the airplane suffers major damage.

The FAA also keeps information on accident rates, and one recent report, separating start-up carriers from major airlines, claimed there is a difference between the two groups. The report ranks airlines by accident rates, but does not include a statistical analysis.

Reporters have already documented airlines failing to report SDRs. Instead of representing the whole universe of U.S. airline problems, or even a random sample, the database is skewed because of contrasting reporting rates. Airlines are required to file SDRs when they repair aircraft, but reporters have found many don't comply with FAA rules, and the agency lacks the teeth for enforcing regulations.

Richard Newman of *U.S. News & World Report* is among those reporters who believe SDRs could never be part of an index on airline safety. In an unpublished analysis, *St. Paul Pioneer Press'* Dan Browning attempted to adjust SDR data by looking at reporting rates. The *Cleveland Plain Dealer* reported nine of 10 repairs were unreported in SDRs (See Uplink, February 1995). Even when reports are made, key information may be missing. Details may not be coded consistently. Accidents and Incidents data also suffer from missing information, though many reporters agree that accident data appears more complete.

What if . . .

If a reporter did not face these and other

Continued on page four

For more on ranking airline safety, read "Why No Airline Brags, 'We're the Safest'," by Adam Bryant in the June 9 edition of the *New York Times.*

For more on human errors in airline crashes, visit NASA's web page at http://olias.arc.nasa.gov/

For the FAA's library of advisories, visit http://www.fedworld.gov/ftp.htm#faa

IRS prosecution checked

Erratic enforcement found

By John Sullivan
NICAR staff

Sleepy Roanoke Virginia has never been associated with rampant tax fraud. But if you happen to live there, you are 57 times as likely to be recommended for prosecution by the IRS than if you lived in New Mexico.

The likelihood of prosecution in Pittsburgh is 29 times greater than that of all of Idaho, according to "Where the IRS Aims the Prosecution Ax," a recent *New York Times* article by David Cay Johnston.

Johnston documented capricious IRS prosecution after examining Justice Department data on 4,542 cases in which the IRS sought prosecution in 1994.

The data, which was stripped of taxpayers' names, was made available to reporters on the Internet by Susan Long and David Burnham, co-directors of the Transactional Records Access Clearinghouse (TRAC) at Syracuse University.

Johnston says the data raises some serious questions about the disparity in the way the IRS pursues tax cases, such as the case of Albert Taggi of Scarsdale, N.Y., who took a "voluntary" buyout from AT&T Corp. in 1985 along with 11 other engineers and middle managers.

According to Johnston, the 12 involved, all residents of New York, filed for refunds of income taxes on the buyout and got varying results: Some paid no taxes. Some paid partial taxes. And two, including Taggi, were taxed on the full amount. The federal appeals court that heard the case said it was irrelevant that the 12 were taxed differently.

Analysis of the data indicates Taggi's case may be closer to the rule than the exception, although IRS and Justice Department officials Johnston interviewed disagreed.

"Officials of the IRS and the Justice Department tried to disown the data after TRAC had spent five years in a protracted lawsuit trying to gain access to it. If the data truly was flawed, the Justice Department had five years to deal with that issue," Johnston said. Long and Burnham argue that the data is useful in showing how the IRS carries out enforcement.

Continued on page nine

For SQL tips, check out this web site: http:// www.inquiry.com/ techtips/thesqlpro/

Continued from page three:

Playing by the numbers

problems, Meyer said they could begin the quest by looking at whether fatal air crashes and major repairs during the 1980s predict accidents during this decade.

"It may be that the past doesn't predict the future," Meyer said. "If it doesn't, there is no reason to look at an index."

But if it did, reporters could also look at how well several factors — such as aircraft age and the number of most severe SDRs — fit together.

One test for reliability is called Chronbach's alpha, which looks at whether factors move together consistently. (In SPSS 6.1, with professional statistics, go to statistics ... scale ... reliability analysis). Those who attended the 1996 Advanced Computer-Assisted Reporting Bootcamp in North Carolina used it for studying polling data.

The most robust indices, Meyer said, are those that still predict even when some elements are stripped away.

Because the FAA's inspector general raised so many concerns about ValuJet before the crash, Meyer said the quest for developing a safety index is worth pursuing.

But one problem — for data hounds, not passengers — is the few number of crashes in U.S. aviation. That means an analysis trying to predict accidents is fragile. To get an idea of what that means, if you changed an airline's accident total just by one crash, it could drastically alter results. The FAA noted that in its recent report on start-up carriers.

One way to shore up that problem is to also look at severe mishaps and other incidents. But those records — along with those on pilot mistakes and other human errors — are less complete.

A recent *New York Times* story noted these problems for developing a safety predictor. Trying to use airline data to do this kind of analysis, said the *Minneapolis Star Tribune's* Greg Gordon, "is a lot like squeezing molasses and watching it go in all directions."

Andy Lehren can be reached at (573) 882-0684, or send e-mail to andy@nicar.org

Don't forget to check out the audio tape recording from the 1996 IRE Conference's panel on covering the FAA. The session included *The New York Times*' Steve Engleberg and CBS' Roberta Baskin. Order the cassette by calling (303) 649-1811. Audio tapes are also available from previous conferences.

Answers found in the data

Tracking rail safety

By Penny Loeb
U.S. News and World Report

A train crashes in your city. What's the accident record of that railroad? Several people in a car are killed at a crossing. How many people have been killed at crossings in your state? Some of the answers are in the databases distributed by the Federal Railroad Administration at its Web site, and also by NICAR.

In May, *U.S. News and World Report* and ABC's PrimeTime Live released a joint investigation of safety problems with railroads. It revealed that accidents have remained pretty steady for a decade while the FRA and railroads ignored calls for better braking systems, tracks, passenger safety and working conditions from Congress and the National Traffic Safety Board.

The FRA's user-friendly Web site has .dbf files ready for download, as well as record layouts and code lists. There are four files. One is accidents, which are available from 1991 through February 1996. Reports provide rich detail, including railroad, time, weather, place, cause, type of track, number killed, number injured, number evacuated and number of cars with hazardous materials. The one catch is that there is a record for each railroad and track owner involved in the accident, making for as many as three records for one accident. The FAQ file explains that the unique identifier is year3, month3, rr3 and incdno3. The data also has a field with the same state-county code used by MapInfo. I mapped the serious accidents for December 1995, which made a compelling graphic.

There are also separate tables on casualties, grade-crossing accidents and number of miles traveled by each railroad. Casualties contains injuries to employees, trespassers, passengers and others. Grade crossings has accidents at railroad crossings. A few of these will be reported in the accident file, if damages to the train were more than $6,300. The grade crossing database also has multiple entries for some accidents, so you need to group by the same four fields.

The FRA has put the entire, very detailed 1994 Accidents/Incidents bulletin on the Web site for downloading. You can spot check your calculations against those in the report.

There are several weaknesses in the data. First, it is self-reported by railroads. Second, the accidents database only has those with damages over $6,300. The damage threshold omits more than 90 percent of the accidents. CSX, one of the largest railroads, had 104 reportable accidents in 1995, but the total number was 1,755. I found a number of small spills of hazardous materials on the Hazmat database, but missing from FRA.

We got total numbers of inspections and violations from the FRA, but did not get the data. It is available on nine-track tape, but it is difficult to use because the inspection system changed in the past two years, as did some of the data. The FRA database administrator, Robert Finkelstein, is very helpful, though.

If you are contemplating a story on railroads in your area, check NTSB accident reports and recent Congressional testimony for detailed background. You will need to develop sources among rail workers, who can guide you to the worst problems. Smaller railroads often have higher accident rates.

Here are some useful Web sites:

- Railroad Definitions: http://pavel.physics.sunysb.edu/rr/railroaddefinitions.html

This includes a glossary of railroad terms, such as FREDs, blue flags, bottling air.

- Stone Railroad Switching: http://www.fortnet.org/%7fdk/uprrhome.htm

This links to many railroad sites.

- Track Warrants: http://www.aimnet.com/~steves/n/twar/twmenu.htm

This is a personal Internet publication with information on some crashes.

Penny Loeb can be reached at (202) 955-2640, or send e-mail to ploeb@usnews.com

The Federal Railroad Administration offers downloadable databases (in ASCII or dbase format), including railroad incidents and accidents, casualties, rail grade-crossing accidents and railroad operations, at gopher:// gopher.dot.gov:70/11/ fra/safety/rrsafety NICAR offers a cleaned-up version of this data with additional documentation for $40 to $60, depending on size of news organization. Call (573)882-0684 to order.

Continued from page two: Air crashes

List, published in Great Britain, and confirms information with the airplane's makers.

- Track ownership and major changes to an aircraft. The AP's Drew Sullivan checked FAA ownership and airworthiness certification records in Oklahoma City. Major repair records also include information for developing sources, including the name of the engineer who signed off on the work.

Andy Lehren can be reached at (573) 882-0684, or send e-mail to andy@nicar.org

Continued from page one:

Technical tips ...

forms, typos and just plain wrong information. But even flawed databases can yield a wealth of information. The key is to understand and quantify the limits of the database. A database must be checked for internal integrity, or how clean the database is kept, and external integrity, or how inclusive is the database.

Improving internal integrity

The first step in processing any database is to "scrub" the data to remove inconsistencies. The user should convert all lower case letters to uppercase and should trim all unwanted spaces. These steps should be done even if the data appear to be uniform; it is not good enough to assume that since the first 1,000 records are upper case all the records will be.

In FoxPro, these steps are easily done with the following statements:

```
REPLACE ALL fieldname WITH;
ALLTRIM(fieldname)

REPLACE ALL fieldname WITH;
UPPER(fieldname)

REPLACE ALL fieldname WITH;
STRTRAN(fieldname,space(2),space(1))
```

The first command removes all spaces before and after the data. The second command converts everything to upper case. The final command replaces every occurrence of two spaces with one space. Run the last command a few times to convert all multiple spaces.

Once the data has been scrubbed, test the important fields to see how populated they are and whether the data meet expected norms. Consider a national database of train accidents (TRAIN). A simple SQL command to test the integrity of the field that reports the state in which the accident occurred (ACCSTATE) would be:

```
SELECT ACCSTATE, COUNT(ACCSTATE);
FROM TRAIN;
GROUP BY ACCSTATE;
ORDER BY COUNT(ACCSTATE)
```

The result would be a list of states and how many accidents occurred in each:

```
IL    23,908
CA    18,991
TX    11,444
       9,773
NY     8,663
```

The result reveals that 9,773 fields — a sizable part of the database — do not contain state names. In addition, it indicates that Illinois has the most accidents, which might be expected because Illinois has a large number of railroad hubs. The absence of a state or the presence of very few records from a state might indicate that data may not have been collected by a state. There is a serious problem if Hawaii leads the nation in railroad accidents.

This same procedure can be used for all geographical entities, including counties, cities, ZIP codes, and congressional districts, as well as to check any field where there is an expected range of values, such as companies, gender, race, candidates for office, or any coded item.

A Minnesota TV station used this method to discover that no rapes had been reported for a particular large police agency. The absence was not due to low crime but because the FBI coded rapes differently than that police agency and, rather than resolving the difference, the FBI simply left out the data.

Also look for codes that do not match the expected code values. Gender codes from one to nine might indicate that the data was improperly imported, data entry was careless, or you're dealing with a very strange group of people.

Standardizing data

Often data does not conform to your needs. An AP reporter wanted to list the top metropolitan areas represented by runners in the Boston Marathon. He found a database in which runners reported their hometowns as Brooklyn and Queens, rather than as New York City. The data had to be standardized to get an accurate count. In this case, the command was:

```
REPLACE ALL cityname WITH;
STRTRAN(cityname,'BROOKLYN',;
'NEW YORK CITY')
```

This same procedure can be use to correct for other data errors. For instance, if a database uses the code CO for both Colorado and Connecticut,

Continued on page seven

Get them while they last — all 12 Uplink 1995 issues are available in a bound edition for $10 plus shipping.

With the bound edition, you can review macros devised by the *St. Louis Post-Dispatch's* George Landau for cleaning data (January 1995); revisit advice from *U.S. News and World Report's* Penny Loeb on covering schools (April 1995) and tracking nonprofits (December 1995); and remind yourself of how other reporters have used OSHA, HMDA and more.

To order, call (573) 882-0684.

Continued from page six:

Checking data integrity

a simple statement can correct the error. Rather than guessing whether Groton is in Connecticut, use the ZIP code in the database as follows:

```
REPLACE ALL state WITH
STRTRAN(state,'CO','CT') for zip='07'
```

This statement will change all the codes of CO to CT if the ZIP code starts with a 07.

For city names, state names and ZIP codes, the data can be easily checked by joining the data with ZIP code databases. This will standardize all the data in one step. (NICAR keeps such a database on its Web site.)

Relating the data

When using relational tables, check the uniqueness of key fields and the referential integrity of the data.

"I don't believe anyone who tells me it's a unique field," says Andy Lehren, database administrator for NICAR. "It's really important to double-check, otherwise all your joins get screwed up."

If a key field is not unique, records may be lost or duplicated, or unrelated records may be joined, depending on the problem. In some databases, a combination of fields may make a record unique. You can use the above SQL COUNT statement with the key field to test uniqueness. A value above one in the count field indicates a non-unique key.

After checking the uniqueness of the key field, check the referential integrity of the database. If a related table has a one-to-one relationship with the master table (i.e., there is one record in the master for each record in the related table), you should end up with the same number of records after the join as before. An outer join (sometimes called a union join) would be empty. Fewer records indicate poor referential integrity; i.e., records deleted from the master were not deleted from the related tables.

A one-to-many join should produce the same number of records as is in the "many" database.

Null values

Another common database problem involves null values. Say you have a field that lists state workplace violations fines. If you import the field as a character data, you might find that many of the records list fines equal to zero. Other records might be blank. Can you assume that a blank is equal to a zero? Absolutely not. It's time to call the agency and ask what it means if the field is left blank.

In FoxPro, if you convert the field to numeric to calculate average fine, those blanks will convert to zero, and your numbers will be skewed. Be careful. The lesson may be to import all fields as character fields and convert them to numeric fields later.

One warning: You cannot do this with packed fields. Packed fields must be unpacked as they are described in the record layouts, otherwise you will get bad data.

A word of caution when you modify fields: Any substantial changes increase the possibility that you are introducing errors into the database. To be safe, work only on a copy of the data, preserving the original. And, even in the working copy, it is wise to duplicate fields you wish to modify, leaving the original fields alone.

External integrity

After processing data, a number of external checks are recommended. Make sure the record count matches the expected number. Many databases can be checked against a report. The FBI Uniform Crime Report databases, for example, can be checked against the FBI's "Crime in the US" report, which comes out a few months earlier. If there are differences — and there often are — make sure you can explain them.

Ask the agency for reports based on your database. If they don't have them, ask if you can send them some results to look at. They know the data better than anyone and will likely see egregious errors. It's amazing how cooperative they become once they've seen you have done the work. Many times, they will recreate your results to test whether your work is correct.

Jennifer LaFleur, database editor of the *San Jose Mercury News*, checks demographic data in any database she gets against Census data.

The final step is simply to make sure things are reasonable. "Beware of the 'Wow!'," warns LaFleur. "If you see something in a database that's really cool, it's probably an error."

Drew Sullivan can be reached at (800) 845-8450, ext. 7639, or send e-mail to drew@ap.org

Miss something? Check out the IRE-L and NICAR-L mailing list archives on our websites at http://www.ire.org and http://www.nicar.org. You can see posts to both lists organized by thread, author and date. The list archives are available in html or in plain text format, which is useful for downloading a particular month.

Continued from page one:

Beware of data perils

cluded a violence rate per 100 students. Under the category labeled "students" was what looked like the number of students per school. Actually, the number stood not for the head count, but for the number of students for which the school was designed.

• In looking at data on bridges, the *Asbury Park Press* found the names of the towns in which the bridges were located were missing about 20 percent of the time. Reporters noticed the problem when they tried to match town names with Census data. In addition, sometimes the town names were just wrong.

• The *St. Paul Pioneer Press* was looking at crimes on buses when it found that some route numbers were the same for separate routes in St. Paul and Minneapolis. By checking the route number with intersections, the newspaper was able to figure out which route was which and avoid an erroneous analysis.

Data problems

From these specific examples and others, we can draw up a partial list of data problems:

• Incomplete data. Categories of information may not be filled out. In the widely-used Federal Election Commission database, the occupation of the contributor is frequently not given.

• Non-standardized data. The same street name, city, company name, or interstate may appear in several ways. If you want to count all the accidents on Interstate 90, you may have to count those for I-90, Interstate 90 and I 90. If you don't, you will be wrong.

• Data entry errors. It's not hard at all to type 10000 instead of 100000. It happens all the time.

• Wrong codes. To save time and space, database designers code categories, such as race, as numbers. They then put together a code sheet for translating these codes. It's simple for an overworked data entry clerk to type the wrong code.

• Misunderstood codes. A database is an evolving beast. Codes are frequently added. Things happen for which there is no code, like "community service" in the *Herald* example. Codes exist for activities that no longer take place, and activities happen and get new codes, but the codes are not added to the code sheet.

Avoiding data potholes

As you can see, numerous data potholes await the unwary CAR reporter. But you can avoid them by using a little caution.

The Associated Press' Drew Sullivan, NICAR's former database library administrator, gives helpful technical and specific advice in Tech Tips in this issue, but here are some general rules for preventing errors.

• Anytime you enter data, double- or triple-check your entries. This means taking each newly-created electronic record and comparing it to the hard copy from which the information was gleaned.

• Never double-check check data by yourself. Use the buddy system.

• If you are looking at annual data, look at last year's report. Take the Atlanta mistake. A glance at last year's crime reports would have shown that some of the top cities were no longer near the top or even on the list.

• If you have a category that can be summed, such as loan amounts, then always compare your total with the total in a hard copy report.

• Always make sure you have the correct total number of records for the database. If there are supposed to be 111 cities in your data, make sure you have 111 cities. If there are supposed to be 50 states, then see how many states you have. If there are supposed to be 2.5 million records, then make sure you have them.

• Use standards to filter out bad records or to find stories. If the limit on campaign contributions is $2,000, then try to find all contributions greater than $2,000. If there are supposed to be codes 1 through 5 for ethnicity, then look for all codes greater than 5.

• Contact someone else who is analyzing the data and run the findings by them. That person could be a government official, a social researcher, or another journalist at a non-competitive news organization.

• Trust your journalistic instincts. If results from a database analysis seem too good or bizarre to be true, then they are probably wrong.

Brant Houston can be reached at (573) 882-1984, or send e-mail to brant@nicar.org

A couple of handy search resources are allinone, an all-inclusive list of search engines on the Internet, located at http://www.albany.net/allinone/ and Beaucoup Search Engines at http://pacific.discover.net/~dansyr/engines.html Also, check out Internet Sleuth, the most comprehensive of the meta search engines with more than 1,500 searchable databases covering a variety of topics, at http://www.isleuth.com/

Continued from page four:

Erratic enforcement

Long, a professor of quantitative methods at Syracuse, and Burnham, a former investigative reporter with *The New York Times*, started TRAC in 1989 to collect, develop and disseminate detailed information about the activities of federal agencies.

Johnston, who has covered tax and pension issues at the *Times* since February 1995, began working on the story after Burnham told him TRAC was about to mount an IRS criminal enforcement page on the World Wide Web after three months of extensive work on the sight.

The data was mounted March 28 under an arrangement that required a password and a two-week embargo to allow for analysis. The sight, http://trac.syr.edu/tracirs/, was visited by more than 165 news agencies during the two-week period, and more than 30 articles were generated in a variety of newspapers.

"We wanted to do several things when we decided to mount the sight," Burnham said. "We wanted to teach reporters and show the erratic enforcement levels of the IRS," he said.

Discovering patterns

Long orchestrated the analysis of the data, which resulted in 91 databases all connected with hyperlinks. "We used a variety of software programs including SAS, Data Desk, and Geo Query," she said. Long likened the project to an epidemiological study where the intent is to summarize complex data and understand the data using geography and visualization to discover patterns.

Long noted that the heavy data crunching was done on large work stations with SAS, which made cross-platform analysis possible. "We were dealing with large databases from the Justice Department and merging those with Census data, federal employee data, mapping data and many others," Long said. "The data required a lot of clean-up and confirmation by cross-checking different databases in a variety of software languages," she said. Long then configured the data to reflect rates, rankings and percentages for each city, state and district.

Burnham and Long said the project is intended to arm reporters with sophisticated and powerful data made bite-size by TRAC and allow them to do what they do best, which is to ask informed questions of their local officials on how their agencies function.

John Sullivan can be reached at (573)882-0684, or send e-mail to john@nicar.org

For an in-depth look at problems with databases, check out "Computer-Assisted Reporting: A Practical Guide" by Brant Houston. It can be ordered from NICAR or Investigative Reporters & Editors for $26 plus shipping.

What's on TRAC's IRS Web site

TRAC's IRS Web sight for criminal enforcement allows you to choose from four main menu choices. The first is a roadmap to the sight, which includes information about TRAC.

The second, Findings and Data Graphics on IRS, includes graphs, rankings and tables on the United States and each district, including a ranking by district.

The third is Data Sets on IRS for downloading information on each criminal referal recommending prosecution in a district, or summary information on all districts, including comparative counts, percents, rates and ranks for districts and the nation.

Finally, you can also navigate by topic, district or by comparative rankings.

If you choose to navigate by topic you may view a series of graphs and tables that give an overview of IRS referrals for criminal prosecution. It also provides information about conviction and prison terms, length of time to prosecute or decline a referral, and odds of referral.

Navigating by district provides essentially the same information, but allows you to choose from 90 separate federal districts and view analyses from those districts.

Navigating by comparative rankings provides a series of maps displaying comparative district rankings and tables used to tabulate the rankings.

TRAC's IRS Web site is located at http://trac.syr.edu/tracirs/

— John Sullivan

On the Internet

Data you can surf

Bruce Maxwell's new edition of "How to Access the Federal Government on the Internet," by Congressional Quarterly Books, is available. The 455-page book contains detailed descriptions of more than 400 Internet sites, mailing lists, and related resources that offer federal government information. Cost is $28.95 and can be ordered from the publisher by (800) 638-1710 or (202) 822-1475 or by sending e-mail to bookhelp@cqalert.com. The ISBN is 1-56802-185-2.

By David A. Millirón
Gannett News Service

When the federal building was bombed in downtown Oklahoma City, *Detroit News* reporter Dave Farrell turned to the Internet to show readers how bomb books are just a phone call away.

Farrell had found one of the most complete encyclopedias on destruction — "The Big Book of Mischief," a 40,000-word electronic tome on bomb-building, explosives and weapons of terror.

When ValuJet Flight 592 crashed in the Florida Everglades, killing all 110 people aboard, *Tennessean* reporter Heather Newman turned to the Internet for a listing of the aircraft's Service Difficulty Reports. Newman was able to document specific problems experienced by the downed plane in the months prior to the crash.

These are just two examples of how the Internet can help reporters cover major news events.

Computer-assisted reporting is quickly becoming as valuable to reporters as notepads and the telephone. However, although the Internet can be a valuable reporting tool, the information gained is no more accurate than it is from other sources.

Some reporters and their news operations were embarrassed in the aftermath of the Oklahoma City bombing when they reported as factual two online hoaxes.

Someone on America Online identifying himself as "Timothy McVeigh" posted several messages that tabloid TV shows — and later some newspapers — picked up and used without confirmation. McVeigh had been behind bars two days prior to the postings.

On an Internet news group, a man identifying himself as a militia member from Montana posted inflammatory messages that the U.S. government was behind the bombing.

The man later admitted he had posted the messages as a joke, but not before some newspapers already had reported the remarks as factual.

When surfing the Internet, always be leery of dirty data. Almost every piece of data you will work with is certain to have some quirks. So, bottom line online: Check it out!

Here's a list of free searchable databases any newsroom with Internet access can access:

- FACSNET is a service developed by the Foundation for American Communications for print and broadcast journalists. Not just another homepage with lists of Internet links, FACSNET goes deep into economics, science, business, technology and other areas that journalists cover every day. Databases include reporting tools, sources online, internet resources for journalists and top issues. http://www.facsnet.org
- The Poynter Institute for Media Studies' "Hot News/Hot Research" homepage will feature research sources that will help you understand and cover top stories. They will be looking at a number of links, but only list the ones with some useful content. http://www.poynter.org/poynter/hrintro.html
- U.S. Federal and State Courts Finder is a service of the Emory School of Law. This point-and-click map to federal courts provides hypertext links to all on-line federal court sites.

  http://www.law.emory.edu/FEDCTS/
  http://www.law.cornell.edu/opinions.html
- Supreme Court opinions and related documents in WordPerfect 5.1 or html format. From the HERMES project, these documents are transmitted by the court upon release and uploaded to The Federal Bulletin Board daily. http://fedbbs.access.gpo.gov/court01.htm

  http://www.law.cornell.edu/supct/
  http://www.law.cornell.edu/syllabi
  ftp://ftp.cwru.edu/hermes/
- One of the most complete all-in-one sites with access to several dozen state constitutions, statutes, codes, legislation, is found at http://www.law.cornell.edu/statutes.html
- United U.S. Post Office Database that actually corrects addresses and gives you ZIP+4.

  http://www.cedar.buffalo.edu/adserv.html
  http://www.ualberta.ca/~slis/guides/direct/phone.htm
  http://www.usps.gov/ncsc/aq-zip.html
  gopher://odie.niaid.nih.gov:70/77/deskref/.zipcodes/index

  And, you can even look up ZIP+4 Codes for United States vessels at http://www.semaphorecorp.com/cgi/navy.html
- Search college telephone books from around the world through a vast collection of searchable

Continued on page eleven

Growing collection of federal databases

From the NICAR library

NICAR offers a number of federal government databases. Here is a list of our growing collection:

NEW • A monthly CD subscription for all 1995-96 Federal Election Commission campaign contributions by individuals and political action committees, plus all presidential matching fund requests.

NEW • The Health Care Financing Administion's 1995 database of all Medicare-funded inpatient work in U.S. hospitals.

• Federal Railroad Administration data for accidents, casualties, and highway crossings. 1991-1995.

• Coast Guard boating accidents, 1969-1994.

• Federal Aviation Administration data, including airplane maintenance work documented in the service difficulty report, pilot licenses and grades, and aircraft registration.

• Home Mortgage Disclosure Act records, for tracking who gets loans and who gets turned down, and finding redlining patterns.

• Federal procurement data, 1992-1994, includes breakdowns by agency.

• Alcohol, Tobacco and Firearms gun dealer records.

• National Bridge Inventory System data, includes inspection grades.

• FBI Uniform Crime Reports, a detailed compilation of crime data that includes statistical breakdowns of individual murders. This includes the new 1994 data.

• Social Security death records, by name and social security number, going back to 1937.

• Occupational Safety and Health Administration violation data includes worker accidents and exposures to hazardous chemicals by companies.

• U.S. Department of Transportation truck accident and census data. It includes accidents by company and road.

• U.S. Small Business Administration loan guarantees, 1989-1995. This includes the name of the business, address, amount covered by the SBA, and status, including whether the loan went bad.

• U.S. Small Business Administration disaster loan guarantees, 1989-1994. This includes individuals and businesses, the amount covered by the SBA, and the status, including whether the loan went bad.

• U.S. Small Business Administration's list of minority companies certified for SBA assistance in seeking federal contracts. It includes the name of the company, its address, the owner, type of business and phone number.

• U.S. Department of Transportation hazardous materials accidents database, a collection of roadway, rail, air and waterway accidents from 1971 to 1995.

• U.S. Department of Transportation fatal accident reporting system. It includes all roadway accidents from 1988 to 1994.

• U.S. Coast Guard directory of U.S. merchant vessels. It includes the name of the ship, the managing owner, home port and various descriptive information.

• National Endowment for the Arts, grants, 1989-1993.

For up-to-date prices and more information, call (573) 882-0684, or send e-mail to nicar@muccmail.missouri.edu.

NICAR's week-long bootcamps offer hands-on training in computer-assisted reporting skills, including the use of spreadsheets and database managers, accessing data in various media, such as nine-track tapes, and negotiating for data. The next open session is Jan. 5-10 in Columbia, Mo. For more information, call NICAR (573) 882-0684, or send e-mail to nicar@ muccmail.misssouri.edu

Continued from page ten:

On the Internet

telephone books and other directories.

gopher://gopher.nd.edu:70/11/

Click on "Non-Notre Dame Information Sources"

Click on "Phone Books—Other Institutions"

• Reporters and editors can calculate everything from foreign exchange rates, local tides, distances between cities, taxes, loan and mortgage rates, the amount of seed needed to plant an acre of different crops and even a quarterback's passer rating on the Calculators On-line Web page.

http://www-sci.lib.uci.edu/HSG/RefCalculators.html

We'll look at some more sites next month.

David Milliron can be reached at (703) 276-5805, or send e-mail to dmilliro@gci l.gannett.com

Bits, Bytes and Barks

New and Improved reporter.org

The new version of the www.reporter.org web site is up. It's located at http://www.reporter.org/index.html

Among the professional journalism organizations and related mailing lists hosted by the site are:

www.ire.org (Investigative Reporters and Editors)

www.nicar.org (National Institute for Computer-Assisted Reporting)

www.aaja.org (Asian-American Journalists Association)

www.nabj.org (National Association of Black Journalists)

www.nyabj.org (New York Association of Black Journalists)

www.ewa.org (Education Writer's Association)

www.jaws.org (Journalism and Women Symposium)

To subscribe to a mailing list for announcements regarding the www.reporter.org web site, send mail to: majordomo@reporter.org

In the body of the message, type:

subscribe reporter-announce

You'll receive a confirmation of your subscription shortly thereafter. This information, as well as the archived postings to the list, is available at: http://www.reporter.org/lists/reporter-announce

If you have any comments regarding the new site or would like to become part of the reporter.org web project, send e-mail to Web Master Wallace Winfrey at wally@nicar.org

Producers wanted

WDIV-TV, the Post-Newsweek-owned NBC affiliate in Detroit, is looking for three producers, including one each for its I-Team, Consumer Unit and Medical team. All should have several years television news experience in writing and producing.

For the I-Team, the candidate must have done investigative reporting in the past. Ideally, the person would also be familiar with CAR techniques and the Internet.

Send tapes and resumes to: News Director Carol Rueppel, WDIV-TV, 550 W. Lafayette Blvd., WDIV-TV, Detroit, Mich., 48226. For questions, call Mike Wendland at (313) 222-0532, or send e-mail to mikew@wdiv.com

Miami Herald changes, job opening

After more than 19 years at *The Miami Herald*, Associate Editor and Researcher Steve Doig is leaving to become Knight Professor of Journalism at Arizona State University, Tempe. Dan Keating will move from the *Herald's* Broward office to replace Doig, and the Herald is searching for a reporter with CAR skills to fill the Broward opening.

Send resume and clips to: Rick Hirsch, Broward City Editor, The Miami Herald, 3325 Hollywood Blvd, Suite #102, Hollywood, Fla., 33021.

New IRE-L and NICAR-L addresses

Effective immediately, IRE-L and NICAR-L have changed addresses. They are located on lists.missouri.edu, a dedicated Unix system designed specifically to service the onver 200 discussion groups run out of the University of Missouri. All posts to the lists should now be sent to: ire-l@lists.missouri.edu or to nicar-l@lists.missouri.edu

All commands for subscription changes should be sent to: listproc@lists.missouri.edu

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