Uplink

March • April 2003
Volume 15, Number 2

Published bimonthly by the National Institute for Computer-Assisted Reporting

www.nicar.org

DATA ACCESS

Obtaining restricted files

By Megan Christensen
IRE & NICAR

You think you have found the perfect data: thorough, current, topical, ASCII text. Then, the fine print. "This data is provided for statistical research purposes only" or "Authorized for release to qualified research personnel or institutions."

The good news is that there is no special data club only for people with Ph.Ds. Journalists, too, can sign research agreements and gain access to databases. These arrangements can open the door to a wide variety of data for computer-assisted reporting.

But before accepting data with legal strings attached, journalists who've considered these agreements recommend exploring: Is there any other way to get the story? Can I report the story the way I want under the terms of the agreement?

Agencies restrict access to many of these data sets be-

continued on page 15

SPOTLIGHT: CAR FOR WARTIME

Key data, resources for covering conflict

By Brant Houston, *IRE & NICAR*

Databases and resources abound for covering war-related stories, but organizing them in a useful way can be a challenge. At IRE's recent Annual Computer-Assisted Reporting Conference in Charlotte, several speakers slotted the resources into these categories.

Military and security

• The U.S. Census Bureau has an enormous amount of data on the military, veterans, educational background of military and so on. Summary File 3 of the Census 2000 release is jam-packed with information. You can use American Factfinder *(http://factfinder.census.gov)* to zero in on what you need.

• National Guard and reserves deployment data can be obtained from the Defense Department DefenseLink Web site in Adobe Acrobat format at *www.defenselink.mil/news/Apr2003/d20030409ngr.pdf.* Check the DefenseLink news releases section for updated lists.

• Defense contract data is in the Federal Procurement Database or on the Web at *www.defenselink.mil/news/contracts.html.* The DefenseLink Web site also has reports with html tables on a variety of subjects. The IRE and NICAR Database Library can provide the procurement database to journalists.

continued on page 18

SPOTLIGHT:

FOR MORE ON CAR FOR WARTIME SEE:

• Using INS data to track deportations by country of origin, p. 3

• GIS helps map out the business of war, p. 7

Help from IRE & NICAR

By David Herzog
Missouri School of Journalism and NICAR

The U.S.-led war with Iraq is over but key questions remain unanswered. How long will coalition forces remain in the country? Will troops continue to battle pockets of resistance? Who will profit from rebuilding Iraq? No one has the answers, but you can count on IRE and NICAR to provide the vital information resources and assistance that you'll need for your own wartime news coverage.

For starters, you'll find timely articles and resources for covering the military, terrorism and cross-border issues in

continued on page 2

Bits & Bytes

New audit data

How are government agencies and nonprofit groups spending money they received from Uncle Sam? A new database from the IRE and NICAR Database Library can help journalists find out. The Federal Audit Clearinghouse database lists all audits of state and local governments, and nonprofits that spent $300,000 or more a year in federal money. The database contains audits from 1997, when the federal government started requiring them, through February 2003. There are 170,900 records in all.

Unlike other databases available from the Database Library – namely, the Consolidated Federal Funds Report (CFFR) and the Federal Assistance Award Data System (FAADS) — the Federal Audit database looks at the spending of federal assistance. It also includes audits of sub-recipients of federal aid. So if a state government gives some of its federal award money to local agencies, audits of those local agencies will also appear in the database.

The data contains information about whether the audits have revealed any problems, plus valuable contact information, including names, addresses and phone numbers for the audited organizations and the auditors. To order the database, call the Database Library at 573-884-7711 or go online to *www.ire.org/datalibrary/databases/fedaudit.*

continued on page 4

Help

continued from page 1

this issue of *Uplink*. Brant Houston, executive director of IRE and NICAR, writes here about the must-have databases. Mark Bixler of *The Atlanta Journal-Constitution* shares how he used Excel and INS data to show how the federal government stepped up deportations of visitors from primarily Muslim countries after the terrorist attacks of Sept. 11, 2001. We'll have even more in the May-June edition, including an in-depth look at reporting about military contractors in your region.

The Annual Computer-Assisted Reporting Conference in Charlotte, N.C., in early March featured a number of panels about using CAR for wartime coverage, everything from obtaining data about the military to the battle over access to information. We'll post any tipsheets from those sessions on the IRE War Resources Web page, *www.ire.org/resourcecenter/war_resources.html.* Check back regularly for updates.

We'll offer even more opportunities to get up to speed on covering those topics at the IRE Annual Conference June 5-8 in Washington, D.C. The conference kicks off with the optional CAR day and includes hands-on classes where you can learn how to analyze data with spreadsheets and database managers.

Looking for a multi-national perspective? You can join journalists from around the world at the second Global Investigative Journalism Conference this May 1-4 in Copenhagen and discuss strategies for covering war, arms smuggling and terrorism. There will also be hands-on CAR classes. You can find more information at *www.ire.org/training/globalconference.* Program organizers include IRE and the Danish Institute for Computer Assisted Reporting.

The IRE and NICAR Database Library has been assisting journalists covering the military and terrorism by providing U.S. government databases regarding federal contracting and immigration. See the IRE War Resources Web page for more.

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

Hands-on CAR training upcoming

IRE and NICAR has numerous training opportunities in the coming months for journalists who want to learn computer-assisted reporting. For a complete list of training events visit *www.ire.org/training/otr.html.*

Here are some of the highlights:

There are two CAR Boot Camps in Columbia, Mo., in the coming months for journalists who want hands-on training in how to use spreadsheets and database managers to analyze data for high-impact stories. The sessions are May 18-23 and Aug. 3-8.

The IRE Annual Conference, June 5-8 in Washington, D.C., will feature demonstrations of CAR techniques and hands-on classes. Members of the IRE and NICAR staff will be there to help you develop your stories and answer questions about computer-assisted reporting. Journalists can count on stellar panels and special presentations about national security, public safety, courts, the military, business, education and more.

SPOTLIGHT: CAR FOR WARTIME

Tracking deportations by country of origin

By Mark Bixler, *Atlanta Journal-Constitution*

Not long ago, computer-assisted reporting was a language that other reporters spoke, as foreign to me as Serbo-Croatian. I had a hard enough time opening e-mail attachments, so slicing and dicing numbers by computer seemed beyond my grasp. Then my newspaper sent me to a four-day workshop that it sponsored with the National Institute for Computer-Assisted Reporting. An instructor demystified spreadsheets and databases. I saw the possibilities.

I cover immigrants and immigration for the *Atlanta Journal-Constitution* and have written extensively about changes affecting non-citizens since Sept. 11, 2001. The Immigration and Naturalization Service detained hundreds of people from Arab and Muslim nations for violations of immigration law that were once considered minor, such as overstaying a visa. It has held many of them in secret and displayed great reluctance to divulge information about them.

Wouldn't it be great, I thought, to know the nationality of people deported after Sept. 11?

A simple phone call to the INS public affairs office in Washington got a Microsoft Excel spreadsheet in my e-mail inbox that showed the nationality of all people removed from the United States for the past six budget years. It listed countries alphabetically; the nations were not ranked by the number deported in any particular year.

The spreadsheet had more than 200 rows for the countries, from Afghanistan to Zimbabwe. The 18 columns showed for each of the past six years the number of people expelled, the number of criminal expulsions and the number of noncriminal expulsions.

Lucky for me, the federal budget year runs from Oct. 1 through Sept. 30. The most recent figures were for the 2002 budget year, which corresponded closely to the 12 months after Sept. 11, 2001. Any changes from previous years would be easy to spot.

My knowledge of immigration law came in handy. I knew people are deported for two main reasons. They either commit a crime, such as breaking into a house or selling drugs, or they violate immigration law by entering the country without permission or overstaying a visa.

TOTAL CHANGES 00-01-02.xls

	A	B	C	D	E	F	
1		FY 00 TOTAL	FY 01 TOTAL	FY 02 TOTAL	% change 02-01	rank change 02-01	% ch
2	Kazakhstan	6	1	14	1300.0%	1	
3	Malawi	3	1	5	400.0%	2	
4	Tunisia	8	13	52	300.0%	3	
5	Congo, Democratic Republic	20	6	22	266.7%	4	
6	Morocco	48	32	109	240.6%	5	
7	Mauritania	4	4	13	225.0%	6	
8	Anguilla	1	1	3	200.0%	7	
9	Azerbaijan	0	1	3	200.0%	7	
10	Bahrain	0	1	3	200.0%	7	
11	Egypt	98	92	273	196.7%	10	
12	Yemen	26	18	49	172.2%	11	
13	Mongolia	4	7	18	157.1%	12	
14	Guyana	239	130	311	139.2%	13	
15	Sudan	6	8	19	137.5%	14	
16	Jordan	101	82	193	135.4%	15	
17	Burkina Faso	1	3	7	133.3%	16	
18	Lebanon	66	49	113	130.6%	17	
19	Pakistan	289	335	767	129.0%	18	
20	Eritrea	2	5	11	120.0%	19	
21	Samoa	13	10	22	120.0%	19	
22	Algeria	34	18	38	111.1%	21	
23	Saudi Arabia	9	16	33	106.3%	22	
24	Cameroon	25	13	26	100.0%	23	
25	Guadeloupe	1	1	2	100.0%	23	
26	Kyrgyzstan	4	1	2	100.0%	23	
27	Libya	2	2	4	100.0%	23	
28	Montserrat	3	2	4	100.0%	23	
29	Argentina	138	247	481	94.7%	28	
30	Slovak Republic	19	20	39	90.0%	29	
31	Unknown	33	16	30	87.5%	30	

Sheet1 / 02-01 / 01-00

Slicing the data

To determine whether – and to what degree – the INS had shifted its focus, I decided to analyze all the expulsion categories. The last category, expulsions for immigration law violations, was important. There are 2,000 INS agents and about 8 million illegal immigrants in the U.S., so the agency has great discretion in how it uses its limited resources.

Armed with my new, elementary spreadsheet skills, I ranked all countries by the percentage change in citizens deported from the previous year. I calculated percentage change in deportations from year to year and then used the Excel sort function to put them in descending order. The countries with the largest increases of citizens deported appeared at the top of my table. What I found would make its way into the paper soon enough:

The number of foreign nationals expelled to North Africa, the Middle East and South Asia multiplied faster than for citizens of nearly all other nations in the year after Sept. 11, 2001. Eight of the 10 countries with the most dramatic increases in deportation were Muslim nations. The top 10, in order, were Tunisia (300 percent), Morocco, Egypt, Yemen, Guyana, Jordan, Lebanon, Pakistan, Samoa and Algeria (111 percent). By contrast, overall deportations went down 18 percent. The number of immigrants de-

continued on page 4

Bits & Bytes

continued from page 2

Extra! Extra!

For some inspiration, check out Extra! Extra!, an online guide to investigative and computer-assisted reporting stories. The Web page, sponsored by IRE, contains summaries of recent investigative pieces and links to the stories. Journalists can also search for stories by category, from health, homeland security and politics to nonprofits. The Web page is at *www.ire.org/extraextra.* To find older stories through the Resource Center, go to *www.ire.org/resourcecenter* or call 573-882-3364.

KBS boot camp

Journalists come from all over the world to learn computer-assisted reporting in boot camps sponsored by IRE and NICAR. In November, visiting journalists from Korea came to the University of Missouri School of Journalism to learn CAR. For a week the reporters from the Korean Broadcasting System, South Korea's national broadcasting service, learned the basics of spreadsheets, database managers and importing data from the Internet, and returned home eager to apply their new skills.

IRE and NICAR can arrange similar specialized training for news organizations, inside or outside the United States. For more information, contact Ron Nixon, IRE and NICAR training director, ron@ire.org or 410-995-0343.

Tracking

continued from page 3

ported to Mexico, which sends the vast majority of immigrants to the United States, went down 24 percent.

Further spreadsheet calculations put my findings in context.

In previous years, none of the Muslim nations ranked among the top countries in surges in citizens deported. Those nations that did have the biggest increases, from Armenia in 1998 to Bolivia in 2001, had more modest jumps than did many of the Muslim countries in the year after Sept. 11. The number of Bolivians deported from October 2000 through September 2001, for example, increased 170 percent, more than for any other country with at least 10 citizens deported the previous year. Four countries (Yemen, Egypt, Morocco and Tunisia) had more dramatic increases in the following year.

The changes were even more dramatic in the deportation of what the INS calls "non-criminal immigrants" – people who were kicked out the country for overstaying a visa of some other violation of U.S. immigration laws. The top eight were Muslim nations: Morocco (407 percent), Tunisia, Yemen, Jordan, Saudi Arabia, Sudan, Algeria and Egypt (166 percent). Meanwhile, the number of Mexican nationals deported for immigration offenses, such as overstaying a visa, went down 38 percent. The total number of "non-criminal immigrants" who were removed declined 28 percent.

Calculate, sort

I arrived at all of these findings with basic spreadsheet methods: Write a formula to calculate the percentage change, then sort descending.

Stepping up the focus on illegal immigrants from Muslim nations struck many as common sense given that all 19 hijackers who struck on Sept. 11, 2001, were from Muslim nations (only two had overstayed their visa and were eligible for deportation at the time of the attacks). Others raised questions about racial profiling and selective enforcement: Why throw a Pakistani in jail for overstaying a visa while a Mexican who did the same thing goes free?

Doris Meissner, INS commissioner from 1993 to 2000, told me that increased scrutiny of foreign nationals from Muslim nations "was inevitable and predictable right after 9-11, but I think it became clear pretty quickly that focusing that way was not unearthing terrorists."

The authorities have charged at least 13 U.S. citizens with crimes relating to terrorism. None of the roughly 900 non-citizens held for immigration violations after the terrorists attacks has been charged as a terrorist.

My story pointed out that the deportees included an Arab student in New York who was expelled for working seven hours a week beyond what his visa allowed and a Jordanian in New Jersey who violated terms of a tourist visa by working at a Dunkin' Donuts; evidence, immigrant advocates say, that the focus on Arab and Muslim immigrants has ensnared petty violators, not hard-core terrorists.

I also interviewed people such as Carma Said of Bethlehem, Pa. She said the INS deported her husband to Egypt even though it had approved his request for legal residency.

My first computer analysis came with a mini-crisis or two. I forgot a formula one day and couldn't remember how to sort columns the next. More computer-savvy reporters in the office answered my questions about columns and rows and sorts and filters. Those terms might sound foreign to the uninitiated, but spend a few hours with someone who knows and you'll discover what I did: It's not that hard after all.

Contact Mark Bixler by e-mail at mbixler@ajc.com.

FEDERAL SPENDING

Getting a handle on shuttle contractors

By Jeff Porter, *IRE and NICAR*

Space shuttle Columbia crashed in pieces Feb. 1, and all the NASA shuttle program money going to contractors quickly became part of the story.

Journalists covering that angle of the story found that a database of federal contracts of $25,000 or more – called the Federal Procurement Data System – is a great source of information for looking at program spending.

The IRE and NICAR Database Library maintains copies of the data from the General Services Administration back to 1992, covering about a half-million federal contracts each fiscal year. The day of the shuttle disaster, the Database Library culled all of the NASA contracts for each of those years, then posted the data and documentation on its FTP server for a quick download.

Two dozen news organizations obtained the data. In some instances, it allowed journalists to show a local connection to the space program, because the data could show exactly how much money NASA gave to specific contractors in a given state, city or county.

The database is fairly straightforward with a few wrinkles.

There's just one data table, although IRE and NICAR provide a few lookup tables with geographic information and details about the contractors. The primary data table includes the agency, name and address of the company, the amount of the contract, the type of work and where the work was performed.

The type of work listed is not specific enough for certain news reporting. For example, it doesn't say the company worked on the left wing of the space shuttle Columbia. Instead, it might say the company's product or service involved "space vehicles" or "space vehicle components."

Next, contracts can be modified over time, and it's important to understand how the changes are reflected in the data. The GSA data system has a separate record for each of these modifications with each record bearing the same contract number. Take, for instance, contract NAS850000 (with the Boeing Co. for space vehicles), modified 104 times from fiscal year 1992 through 2001.

The records show modification No 135 in 2001 and an additional obligation of $16.7 million. The next modification resulted in a reduction of the obligation by $1.2 million. So if you add all the numbers together (sometimes over several fiscal years), you can find the total.

Complicating matters further, the GSA has added some fields over time. For fiscal year 2001, the IRE and NICAR Database Library created a program that processes the data, which will smooth some consistency with field names and how dollar amounts are handled over the long run. All those are reflected in the documentation the Database Library provides. Those documents should be examined carefully to account for any of those changes.

It's possible to turn the 10 yearly data files into one. The Database Library kept them separate so it could deliver the data more quickly to journalists the day of the shuttle disaster.

Now, let's turn to the data. We'll go over a basic series of Visual FoxPro commands to choose the fields we're interested in and string the years together, then make some basic queries of the data. Any number of other database managers – Access, SQL Server, for example – could do the trick, too.

First, we studied the documentation and, for our purposes, decided to go with 18 fields, chosen because of their consistency year to year and their importance. They include the contracting agency, the number identifying the contract and any modification, the amount of money, the four fields describing the type of work, the three different fields that could list the company name, plus the locating information on where the company is based and where the work was performed. (See the record layout at *www.ire.org/datalibrary/databases/fedcontacts/layout.txt*).

The complete Visual FoxPro program can be found at *www.nicar.org/techtips.html*. Here are some key parts of the program.

continued on page 6

Shuttle

continued from page 5

```
use nasa01

select contractor,;
connum,;
modnum,;
entrytype,;
contaction,;
dollaramt,;
prodcd,;
prodnomen,;
naicscd,;
naics,;
contnmicar,;
contnmcif1,;
contnmcif2,;
contcity,;
contstcd,;
contzip,;
stateperf,;
perfcity from nasa01;
into table m:\dataprob\nasa
```

We continued to make separate tables of each year. If a field was named differently in a previous year, we told the program to change the name, simply like this:

```
… siccd naicscd,;
sicnomen naics,;…
```

The GSA had changed its reference from the SIC (Standard Industrial Classification) to NAICS (North American Industry Classification System) from one year to the next. The language simply tells Fox to select the fields refer to SIC and call them the same field names as the 2001 data calls them.

Before 2001, the data implies three zeroes after the dollar amount. So to make it consistent, we multiplied all the dollar amount fields by 1,000 like this:

```
…dollars*1000 dollaramt,;…
```

That line simply does the math and names the field the same as the 2001 data. The 1999 data had a slight variation for the dollar amounts. Instead of indicating a negative number, the data included a field, called NEGATIVE, showing whether the number is negative or positive. So for the 1999 data file, before we ran our main program (although we could have incorporated it into that program), we created a new dollar amount field, then ran a couple of replacements:

Replace all dollar2 with dollar
Replace all dollar2 with 0-dollar for negative = '-'

Before we stitched all those years together, we needed a way to tell which year is which, so we added a field to each table and told Fox to fill that in with the appropriate year:

```
use nasa98
alter table nasa98;
add column year c(4)
replace all year with '1998'
close databases
```

Finally, we reopened the 2001 table we simply called NASA, then appended all the years together.

There are other ways around this. For example, one could have added the years, then run a UNION ALL query to string the tables together.

With the data file covering 1992-2001, containing 114,126 records, a few queries could trigger some potential thoughts. Here's a very simple one:

```
select year, sum(dollaramt)
total;
from nasa;
where prodcd like '18%';
group by 1;
order by 1
```

It shows a 10-year timeline of the money related to space vehicles. Interested in a particular state?

```
select year, sum(dollaramt)
total;
from nasa;
where prodcd like '18%' and
(contstcd = 'TX' or stateperf
= 'TX');
group by 1;
order by 1
```

That includes the same timeline, but limiting it to contracts where the company is located in Texas or the work was performed in Texas. Locate the biggest contractors for Texas, on space vehicles work?

```
select contestcd, contnmicar,
contnmcif1, sum(dollaramt)
total;
from nasa;
where prodcd like '18%' and
(contstcd = 'TX' or stateperf
= 'TX');
group by 1;
order by 4 desc
```

This query groups by the Dun and Bradstreet number for the company, the field called CONTESTCD. That's often useful because the data is somewhat inconsistent in how the company is named. It's important to note, though, that some companies have multiple Dun and Bradstreet numbers, so you might need to do a little data checking and cleaning for accuracy and consistency.

Nancy Amons of WSMV-Nashville, was able to localize the shuttle disaster with federal contract data from IRE and NICAR, reporting "NASA pumps millions into our local economy," identifying companies and finding people to interview, guided by the data.

"In all," Amons reported, "NASA awarded nearly $21 million in contracts to Tennessee companies in 2001. Saturday's disaster could mean even more."

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

readme.txt

For more information about the Federal Procurement Data System or to purchase it from the Database Library, contact NICAR at 573-884-7711 or go to *www.ire.org/datalibrary/databases/fedcontacts*. The Database Library can provide the data from 1992 through 2001.

MAPPING IT OUT

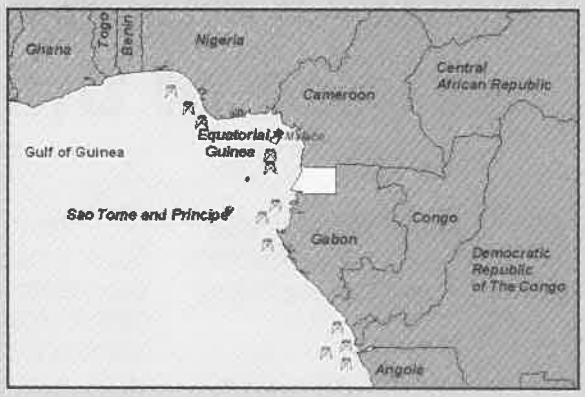
The latest uses of mapping in news reporting.

The Business of War

By Ron Nixon, *IRE and NICAR*

Wars and ethnic conflicts have wreaked havoc on the economies and internal stability of developing countries around the world. Many of these are countries in name only, as the strife has fractured them into sections. For a small number of people around the world, however, wars and conflicts have been profitable.

Many operate out of places and countries that few Americans have ever heard of, so when the Center for Public Integrity's International Consortium of Investigative Journalists began its special report called "Making a Killing: The Business of War," it used ArcView 3.2 to create maps that illustrated the global connections of the weapons trade.



The center used ArcView to show places like Ostend, Belgium, an operating base of key weapons dealer Victor Bout. It also used the mapping software to show the location of the oil-rich Gulf of Guinea, where several states employ private military companies (once called mercenaries) and the location of the rich diamond deposits in the Congo.

The maps were largely locator maps and none required much advanced GIS work, but they did require research to locate information. The mapping work consisted of overlaying country shapefiles, national capitals and resource features from ArcAtlas: Our Earth, an ESRI data set. ArcAtlas lacked some of the resources that the Center wanted to show. For example, the location of oil fields in most cases was included, but not for those offshore. Additionally, resources such as uranium tailings in Angola were not included in the shapefiles.

To get the locations of other resources, the Center turned to other sources. One was the CIA World Factbook (*www.cia.gov/cia/publications/factbook/*), with maps and information about every country.

Google searches found the Internet sites of global oil companies and information about their offshore sites.

Searches of the French newspaper *Le Monde* were used to locate some of the natural resources, such as diamond mines and oil fields. The Center used the National Geographic Web site, located at *www.nationalgeographic.com,* to find many of the cities not included in the ArcView shapefile of major world cities.

The Center used the ArcView legend editor and symbol palette to create unique symbols for the oil fields and diamond mines. The Center also changed the background color of the view from the default of none to blue, so the oceans appeared as blue.

Though the maps were a small part of the project, they helped by showing the locations of the places and key natural resources discussed in the project.

"Making a Killing: The Business of War" can be found online at *www.icij.org/dtaweb/icij_bow.asp.*

Contact Ron Nixon by e-mail at ron@ire.org.

FIRST VENTURE

Simple queries find prince of the city

By Michael Lafleur, *The (Lowell, Mass.) Sun*

I had always thought that the most difficult part of any computer-assisted reporting project would be negotiating for the data.

But in my first real attempt at a CAR story, I found the opposite to be true. I got the data – a catalog of complaints about alleged building and health code violations at properties in the city of Lowell – after a fight, but not an intolerable one. The most difficult part of my project was figuring out what to do with the data.

My initial story idea turned out to be undoable with my time and budget constraints, forcing me to explore other avenues with the data. What began as an attempt to examine the effectiveness of Lowell's property inspection system ended up as a profile of an outspoken and flashy Nigerian prince who was one of the largest owners of run-down properties in the city.

The project started in late summer, after several gang-related shootings in one of the city's poorest neighborhoods. Politicians in Lowell, a former textile manufacturing boomtown, called for a police crackdown on the gangs. City administrators began focusing on the absentee landlords whose properties police said served as hangouts for the gangs.

With the new emphasis on enforcing state building and health codes, we wanted to quantify what had been done prior to the high-profile crackdown. We especially were interested in determining the frequency of property inspections and the city's overall success rate for abating problem properties. I had attended a NICAR Boot Camp a few months earlier and thought this project presented a good opportunity for me to incorporate what I had learned.

As I reported, I discovered that city building inspectors kept a Microsoft Access database of their activities dating back to 1998. I asked for it, making what seemed to be the first request for electronic records that the city of Lowell ever had. My interest in the data initially triggered consternation from city officials. The building commissioner accused me of targeting him for unfair scrutiny and said I would have to pay 50 cents a record for each of the 12,000 records in the database. City legal staff asked me to place stricter limits on my request. At the time, I wasn't exactly sure what the data would show, but I hoped that it would at least shed light on the effectiveness of building department inspections. Unfortunately, that did not turn out to be the case.

The database contained complaints about problem properties called into the building department and a description of the response to them, but I was told it was not comprehensive. I was beginning to get a bit discouraged. I soon learned that building inspectors enforce only state building codes, which encompass construction guidelines and minimum structural soundness rules. They become involved with properties when there are problems such as a lack of exits from an apartment or a balcony railing is unsafe. Cockroach and rodent infestations, broken plumbing, leaking sewage, heat and electricity shutoffs – almost any problem in the interior of an apartment – are health code violations. These account for the bulk of citations generated by the city and are handled by Lowell's Health Department.

Health and building inspectors kept records separately on paper, under a property's address, and neither did it very well. The state housing court for my region lists all the Lowell property citations that are heard by a judge, but court records only are available on paper at 50 cents a page and offer no insight on how many complaints city inspectors actually resolve. Reporting a story on the total number of inspections conducted in Lowell and their outcome would have required me to sift through large numbers of files and far exceeded the time I had available for the project and possibly my newspaper's budget.

I discovered many of these shortcomings about the same time I negotiated my public records request with the city legal staff. From the outset, the legal staff repeatedly asked me to explain specifically what I was looking for and to limit my request to just that. My insistence upon receiving all of the data really flustered the city attorney I dealt with. She was unfamiliar with the concept of a database and seemed very nervous about giving me something that could reflect badly upon her later. At one point, she said she would need months to review every bit of information in the database before she could release it to me. But we finally scheduled a formal meeting.

She brought along the city's chief information officer, who reinforced her demand that I explain exactly what I wanted by unleashing a stream of techno-jargon at me about how difficult it could be to fulfill my request. This was all very frustrating. It was now nearly two months after I had asked for the data and no one whom I was dealing with had even looked at it. Had I been unaware that the database was relatively simple and small, the protests would have sounded completely valid.

Luckily, the paper's executive editor, Kevin G. Keene, had agreed to accompany me to the meeting. His willingness to stand behind my request, I think, was the deciding factor in my success. He asked the two city officials to look at the database and report back to us on how difficult it would be to provide a copy. Otherwise, he told them, we were willing to pursue legal action. They agreed and two weeks later I had a copy on com-

pact disc for $20. The chief information officer, on instructions from the attorney, redacted only the names of people who filed complaints. She argued that releasing the complainants' names would violate their privacy rights under state law. I didn't challenge that because I was more interested in the number and focus of the complaints than who complained. (The public records attorney in the Massachusetts Secretary of State's office later told me he would have sided with the attorney's interpretation if I had appealed.)

The NICAR-L listserv proved to be invaluable during the negotiations. I was able to solicit the advice of veteran reporters who had already gone through what I was experiencing. Every time I dealt with the city attorney I asked journalists on the list about how to respond. I also relied upon the free legal defense hotline maintained by The Reporters Committee for Freedom of the Press (*www.rcfp.org*), which was another very handy resource.

Unfortunately, my suspicions about the shortcomings in the data proved true. The database held two main tables, a listing of building permits granted and the catalog of property complaints – each containing between 10,000 and 15,000 records dating to 1998. And then there were three smaller tables that referred to fields in one of the two larger tables. But there would be no way for me to quantify the efficiency of all city property inspections. While I could have done a story on how well building inspectors were addressing the complaints they received, I would not be able to provide a similar comprehensive look at health inspectors, who field many more property complaints.

Now the pressure was on. My boss had invested money for me to attend the Boot Camp and his time in helping me get the data. Figuring out what to do once I realized my initial idea would not work was probably the most intimidating part of my first experience with CAR.

Sifting through the data I came across several complaints logged against a man named Prince Charles Eweka, who piqued my interest. It turned out the prince had a fascinating story. The scion of a Nigerian royal family, Eweka was one of the largest property owners in Lowell's low-income neighborhoods. He insisted that he was trying to reinvigorate the area, but the prince's properties had myriad structural problems and a long history of building and health code violations, which he typically did not correct until receiving a court order. But I found all this through interviews and manually reviewing city and court records.

My data analysis was very simple, and Boot Camp had prepared me well for what I needed to know. Using Microsoft Access 97, I focused my attention on the complaint table. I performed a group by name and count query to see which names appeared most often. I didn't have time to thoroughly clean the data, but because the table had relatively few records I printed it and eyeballed it, looking for alternate spellings that had escaped my initial query. I corrected only the names of people who appeared several times. In that way, I came across Eweka and decided to investigate him further.

The story, "The Prince vs. the City," ran in December 2002 and detailed the city's complaints against Eweka.

Looking back, I think my nervousness about developing a story from the data was just first-time jitters. Now that I've incorporated CAR into a story once, I'm much more comfortable with doing it again. Next time, I won't be so surprised if my initial preconceptions aren't borne out by the data. Though I have yet to figure out a way to quantify the efficiency of Lowell's inspection departments, I still haven't exhausted all the story possibilities for the complaint database, and I'm currently using it in another project that involves more data analysis.

Contact Michael Lafleur by e-mail at maolafleur@attbi.com

EDUCATION

Analysis helps uncover college funding scheme

By Natalya Shulyakovskaya, *The Orange County Register*

Our "Phantom Classes" series started with a tip from a parent: His two sons had received community college credits for playing baseball in high school. Their college transcripts from Cypress College showed they had gotten A's – even though they had to transfer to another high school and did not even complete the course.

By the time I joined the project in late August 2002, my colleagues Scott Reid and Marla Jo Fisher had already found that Orange County community colleges enrolled high school athletes in droves in what they called college physical education courses. Their high school coaches rounded them up for the classes and got paid to run the phantom college courses. These classes, called Bridge P.E., are a variation of the legitimate "Bridge Program" that allows qualified high school students to "bridge" gaps in their educational programs by taking tougher community college classes that meet their educational needs.

After I joined the project, we turned to data analysis to get a clear picture of the extent of the program. In early September, we requested enrollment, course, student and faculty data from the California Community Colleges Chancellor's Office, using the California Public Records Law.

It took two months of negotiations — e-mails, letters and even appeals to state legislators – until we received some of the data we needed.

The Chancellor's Office argued that it could not release student-level data because the federal Family Educational Rights and Privacy Act prohibits the release of easily identifiable student records. Citing the same law, it argued that it could not strip all the identifiable information and replace the unique identification numbers that link the student and course tables. We accepted course-level summary data under protest.

The data came from the Chancellor's Office Management Information System (MIS), which gathers information from California's 108 community colleges. We were not charged for the data, which the office provided as ASCII fixed format files and allowed us to download from its Web site.

Tracking enrollment

The data covered 10 academic years starting from 1992-93 and contained records for 1.5 million courses, listing the number of students, the number of high school students and the number of full-time equivalent students, or FTEs, each course generated. The colleges use FTEs to track their workload and apply for funding.

The Register also obtained three additional tables from the MIS office. The tables listed more details for each course, including whether the courses were offered for credit. The state allocates a much larger chunk of funding based on the number of for-credit FTEs. In fact, more than 80 percent of taxpayer funding is allocated to community colleges on the basis of for-credit FTEs, and we focused our analysis on those.

Oddly, some of the totals the colleges reported for budget apportionment purposes were different from the numbers they had submitted for the MIS database. It took some tests, conversations with colleges and college finance experts to figure out how to handle the discrepancies – especially since the numbers fluctuated both ways.

Our analysis of the data showed that while the entire community college enrollment grew by 20 percent in the last 10 years, growth in the physical education classes outpaced most other courses, increasing seven-fold. Since the 1992-93 academic year, it ballooned from a mere 0.3 percent of the entire California community college system's enrollment to almost 2 percent.

Calculating costs

In order to figure out how much the phantom P.E. courses were worth, we waded though the state apportionment reports and built a database detailing funding allocations for each community college district and for the state, based on enrollment. The apportionment calculations are used to distribute money to community college districts. They are based on numbers the districts report to the state.

Five years worth of apportionment reports, in Adobe Acrobat format, was available on the Chancellor's Office Web site. Each contained at least 72 pages densely covered with numbers and text.

When I asked for the 10 years of the data in tables, I was told it would take about 12 weeks to compile. I saved the data in the Acrobat files as text. It then took me a day to write, test and modify a FoxPro program that parsed each report into a database.

Most of the muscle work, including rebuilding and condensing the tables, was done in Microsoft SQL Server. SPSS came in handy for the final analysis.

To understand the community college financing, I retraced steps used for calculating apportionment funding. I used three main sources for understanding how the complicated funding formulas worked: materials the Chancellor's Office provided to the college administrators during its annual statewide budget information workshops; Title V regulations that outlined funding algorithms, and spreadsheets used by the colleges to estimate their funding.

When my results came within 0.009

percent of the state's own numbers, I felt confident enough to start estimating the worth of different courses to community colleges.

Our most conservative analysis showed that in the 2001-02 academic year, high school athletes counted as community college students generated at least $56 million for the colleges. Our reporting showed that in their quest for higher enrollment and, ultimately, more taxpayer dollars, the colleges broke the rules defining what could be called a college course. The Bridge P.E. classes amounted to nothing more than regular high school sports practices. Some classes were not held at all. Colleges paid millions to high-school coaches to sign up and train the enrolled high school kids.

Doing the math

Here is how I calculated the key numbers:

• $56 million received for Bridge P.E. classes

Using the MIS database, I calculated a percentage of total credit FTEs that the colleges reported based on their Bridge P.E. enrollment. For example, in 2001-02, Bridge P.E. accounted for 1.8 percent of all for-credit FTEs. That year the state distributed $3.3 billion based on for-credit FTEs.

Because of budget growth limits established by the state, the state funded only 96 percent of all for-credit enrollments at community colleges in 2001-02. So we adjusted the final worth of Bridge P.E. to reflect that.

Some administrators argued that because Bridge P.E. represents roughly 2 percent of the entire enrollment these sports classes essentially are run at no cost to taxpayers. They base their argument on the fact that the funding shortfall is 4 percent. But the state doesn't calculate funding this way, records and interviews show. In reality, colleges apply for funding based on their total enrollment – including Bridge P.E.

• At least 77 of the 108 California community colleges conducted Bridge P.E. classes at high schools in spring 2002.

I counted only the colleges that had more than 100 high school athletes enrolled.

To come up with the most conservative estimate, *The Register* adjusted the numbers of students counted by the Chancellor's Office downward to reflect differences between those numbers and the numbers the colleges themselves report to the Chancellor's Office fiscal unit.

For some colleges, counts in the MIS data were lower than those reported to the fiscal unit. For others, they were higher. *The Register* adjusted down the numbers for over-reporting colleges. The adjustment was done only to analyze district-level data.

Statewide data was not adjusted because the state does not adjust its data. *The Register* also ran tests to make sure the differences between the two different state databases did not affect the overall trends.

• Typical P.E. class

The Register calculated the median for the 2001-02 Bridge P.E. courses: 166 high school students.

• Percent of state-funded growth for community colleges

The Register used the state budget apportionment data to calculate funded growth. Then the growth in Bridge P.E. was calculated using the MIS data. Because enrollment has not been funded fully, growth in Bridge P.E. was adjusted accordingly. Then I calculated percentages.

As a result of our questions and stories, the state announced it was withholding $80 million from the community colleges this year. The Chancellor's Office and the state Department of Finance announced they were going to conduct a system-wide audit. California legislators have called for an independent audit. Some colleges have canceled their planned P.E. classes for high school students. Our phones and mailboxes were flooded by messages from outraged readers and parents. The stories — especially in the midst of the California budget crisis — have really touched a nerve.

For broader look at the reporting behind the series, read the March/April 2003 *IRE Journal*.

You can read the stories online at *www.ocregister.com/features/phantom_classes.*

Contact Natalya Shulyakovskaya by e-mail at natalyas@ocregister.com

readme.txt

The IRE Resource Center has copies of education stories using computer-assisted reporting. Here are some from 2002.

Story No. 19715: "An uneven hand." *The Seattle Post-Intelligencer* examined the discipline gap in Seattle Public Schools, where black students were far more likely to be suspended or expelled than students of other races.

Story No. 19810: "A promise broken: Failing Indiana's schoolchildren." In two special sections, *The Indianapolis Star* examined school financing and found that Indiana has repeatedly asked schools to do more with diminished resources. State officials have failed teachers and students by denying them time, training and money to meet higher standards; failed taxpayers by not determining the costs to raise student achievement.

To order a story, call the Resource Center at 573-882-3364 or rescntr@nicar.org.

Tech tip...

Using Census files to link 1990 and 2000 geography

By Dan Keating, *The Washington Post*

No one inconvenience can be dubbed the bane of a Census reporter's existence since there is a whole list of banes, from shifting release schedules to questionnaire wording changes to margins of error. One nominee for the worst, however, is the shifts in Census geography that can make calculating changes over time a challenge. I once naively thought that tracts either split or combined. Oh, if it were only true. They shift, re-combine in parts and otherwise heave to and fro like a sailboat tossed at sea.

Census reporters have tried various approaches for reconciling the moving tracts. Richard O'Reilly at the *Los Angeles Times* devised a brilliant approach in SAS that read the raw text of the TIGER\Line mapping files, looked at the vertices of each polygon and divvyed them up. Others have chosen to make the comparisons right in their GIS programs by stacking the geographic layers and computing how much the polygons overlap.

I have chosen, however, to go with the Census Bureau's tract-to-tract relationship files. While the files can be tricky to explain and are befuddling to look at, the good news is that they're not difficult to use. For each 1990 tract, the files show how much of its population went into any 2000 tract. So if a tract was divided with 20 percent of its population going into one tract, 30 percent into a second tract and 50 percent into a third, the file would indicate that. You can then use those percentages to assign any 1990 count data into 2000 tracts and easily compare changes over time.

Inside the files

The 1990 to 2000 tract relationship files are available at the Census Web site at *www.census.gov/geo/www/relate/rel_tract.html*. The files I used are the Population-based Census Tract Relationship files. You can find files for each state and a national file, all in fixed-width ASCII, along with the record layouts you'll need to import them into a spreadsheet or database program.

There are 18 fields. Four fields make up the 1990 FIPS code (state, county, tract base and tract suffix), and another four make up the codes for 2000. There are additional fields for 2000 with the state abbreviation and county name.

I suggest importing the relationship tables into the same database program where you store your Census data. That way you can easily link to the tables. Make sure you format the tract FIPS codes the same way you have them in your data. So if you put periods between the tract base and the tract suffix, make sure you do the same in the relationship tables. You'll be linking these tables to your 1990 and 2000 Census data.

Each record describes the relationship between a 1990 tract and a 2000 tract. Tracts that do not overlap have no record. The key parts of each record are two fields that describe the relationship between the 1990 and 2000 tracts. One is labeled "Percentage of 1990 census tract population." The other is "Percentage of 2000 census tract population." If a tract didn't change, both of those numbers will be 100 percent, since none of the population has been split up.

There are also flags for 1990 and 2000 indicating whether the record represents part of a tract. If a 1990 tract was split over two 2000 tracts, there will be two records in the table for that 1990 tract – one record to describe the relationship to each of the 2000 tracts. In that case, each of the records will have a "P" on the 1990 tract FIPS code to show that the record only describes a portion of the tract.

There's one very important limit to using these files: You can use them only on counts, not on median values, averages or any other calculated fields. You can, however, use them to assign data and generate your own calculated fields.

An example with the files

I wrestled with the partial-tract flags and the two numbers describing "Percentage of 1990 census tract population" and "Percentage of 2000 census tract population" for a while until I realized that I could throw most of them away. If I want to assemble 1990 data in 2000 Census tracts the only number I need is the "Percentage of 1990 Census Tract population." I named that field "tract90mult," which means that is the multiplier that I use for the 1990 tract to get the proper share for the 2000 tract.

Let's look at a small example.

The relationship table tells us that 1990 Tract 0001 has 40 percent of its population in 2000 Tract 0002 and 60 percent of its population in

2000 Tract 0003. The pertinent parts of the table would have two records:

```
1990 Tract FIP (simplified)
2000 Tract FIP (simplified)
Tract90mult
0001  0002  40
0001  0003  60
```

So if I'm doing a story about African- American homeowners and I want to have a tract-to-tract comparison, I would divide the 1990 tract into the 2000 tracts using the tract90mult value. Note that the value is expressed as a percent. But if you multiply by those values, you'll get results 100 times too large. You have to divide the tract90mult by 100 first. I did that when I imported it. So if the tract90mult was 100, I turned it into a 1. In this case, the values become 0.4 and 0.6.

Here are the three steps needed to calculate:

(1) Link the tract relationship file to the 1990 data by tract FIPS and multiply the 1990 value by the tract90mult.
(2) Sum the new values grouping by the 2000 tract FIPS.
(3) Link the result to the 2000 data using the 2000 tract FIPS to do comparisons.

Here is how to calculate the rate of African-American homeowners using SQL. The 1990 data is in STF1 table H9. The 2000 data is in SF1 table H15B.

```
select  b.tractfip90,
b.tractfip00,
  b.tract90mult*a.H0090002
+ b.tract90mult*a.H0090007
as black_hhlds_share,
     b.tract90mult*a.H0090002
as black_own_share,
from tblSTF190 a join
lkpTractConvert90_00 b on
a.tractfip90 = b.tractfip90
where a.sumlev = '140'
```

The first step creates the values `black_hhlds_share` and `black_own_share` that have adjusted values. Now I have to aggregate those 1990 values into the 2000 tracts.

```
select    tractfip00,
sum(black_hhlds_share) as
black_hhlds90,
sum(black_own_share) as
black_own90,
from [the previous SQL
query]
group by tractfip00
```

Now I have a table with one record for each 2000 tract with a properly apportioned share of the 1990 data. All I have to do is link it to my 2000 data compare by tract.

```
select    h15bi1    as
black_hhlds00, h15bi2 as
black_own00,black_hhlds90,
black_own90, h15bi2 -
black_own90          as
black_owngain
from tblSF12000 a join [the
previous SQL query] b  on
a.tractfip00 = b.tractfip00
```

You can loop these three queries into one, which is how I did it. It looked something like this:

```
select    h15bi1    as
black_hhlds00, h15bi2 as
black_own00, black_hhlds90,
black_own90,black_rent90,
h15bi2 - black_own90 as
black_owngain
from tblSF12000 h00 join
   (select  tractfip00,
round(sum(black_hhlds_share),0)
as black_hhlds90,
round(sum(black_own_share),0)
as black_own90,
from          (select
tractfip90,tractfip00,
     tract90mult*h0090002
+ tract90mult*H0090007 as
black_hhlds_share,
tract90mult*h0090002 as
black_own_share,
from tblSTF190 t0 join
lkpTractConvert90_00 cv on
t0.tractfip00        =
cv.tractfip90
     where t0.sumlev =
'140' and localyn = 'Y' ) a
     group by tractfip00)
b on g00.stdfip00 =
b.tractfip00
```

Getting the stories

The first waves of Census stories have come and gone. Tools like the tract-to-tract relationship files can enable news organizations to delve deeper into patterns. *The Washington Post* used a version of the query above to help map home ownership for a story that looked at how African-Americans had fared over the 1990s in the Washington, D.C., region and elsewhere.

By using dot density representation of the counts at the tract level, we were able to display the 1990 values and additions in the decade on the same map. It showed that the old pattern of African-American homeowners concentrated in the District of Columbia had been replaced by a wide belt of suburban homeownership.

By the way, if tracts are too big for you, the Census has block relationship files that work the same as the tract relationship files. They're on the Census Web site at *www.census.gov/geo/www/relate/rel_blk.html*. There are not any block group relationship files.

Contact Dan Keating by e-mail at keatingd@washpost.com.

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

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

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

2074240001

Uplink Info

A newsletter of the National Institute for Computer-Assisted Reporting

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

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

Editorial
573-884-7711

Subscriptions
573-882-2772

Postmaster: Please send address changes to IRE-NICAR.

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

Subscriptions
IRE members $40, nonmembers $60

Director of Publications
Len Bruzzese
len@ire.org

Advertising Coordinator
Pia Christensen
pia@ire.org

Subscription Administrator
John Green
jgreen@ire.org

Editor
Brant Houston
brant@ire.org

Managing Editor
David Herzog
dherzog@nicar.org

Asst. Managing Editor
Jeff Porter
jeff@nicar.org

Art Director
Lisa Triefenbach

Copy Editor
Pia Christensen

Contributing Editors
Megan Christensen
Jaimi Dowdell
Holly Hacker
Stephanie Kang