Uplink

January • February 2004
Volume 16, Number 1

Published bimonthly by the National Institute for Computer-Assisted Reporting

www.nicar.org

GIS DATA

Finding files for mapping

By Jaimi Dowdell
IRE and NICAR

In the sea of resources on the Web, a great deal of geographic information system (GIS) data is available free for download. The catch is you just have to know how and where to find it.

While general government agency Web sites like the U.S. Geological Survey (*www.usgs.gov*) and the U.S. Census Bureau (*www.census.gov/geo/www/maps*) contain plenty of GIS data ready for download, much of the data available on the Web is so widely scattered and hidden so deep within government agency holdings that it is difficult to know where to begin.

To help remedy this problem, the federal government has developed *www.geodata.gov*. As part of the Geospatial One-Stop initiative, this site is intended to serve as a portal for access to federal, state and local GIS data.

The intent of the portal is to

continued on page 20

SPOTLIGHT: CENSUS REPORTING

Data shows woes for middle-class blacks

By David Mendell and Darnell Little, *Chicago Tribune*

Almost any demographics reporter can attest to this: Poring over spreadsheet after spreadsheet of U.S. Census Bureau data can be a mind-numbing experience. And even if done with the keenest demographic eye, it often leads to a numbers-heavy story with little potential for readership.

So as the two-man, tag team of demographics/Census reporters at the *Tribune*, we decided to take the opposite tack: Look at a fascinating human phenomenon that seems crystal clear to us in the real world, and then turn to the numbers to see if they

continued on page 12

SPOTLIGHT:

For more on the Census see:

- A guide to the SF4 data, p. 4
- PUMS analysis finds Indiana "brain drain," p. 8
- Using American Factfinder to find data, p. 11
- The January/February 2004 *IRE Journal*

CRIMINAL JUSTICE

White-collar convicts get lighter sentences

By John Archibald, *The Birmingham News*

The decision to take a long look at white-collar sentencing in North Alabama's federal courts was easy.

The late U.S. District Judge Edwin Nelson pretty much made the decision for us when he set Mari Sanders free.

Though Sanders in 2001 was convicted of using an 11-month scheme to steal more than $187,000 from her bank employer and, though evidence revealed she had stolen more than $33,000 from a previous employer, Nelson sentenced her to eight hours in custody.

The 11th U.S. Circuit Court of Appeals ordered the judge to reconsider the sentence. But Nelson stuck to his guns, again sentencing her to eight hours.

Federal courts reporter Val Walton had seen drug defendants go to prison for years. She also had repeatedly seen white-collar thieves like Sanders go

continued on page 22

Bits & Bytes

Editor training

Journalists from around the globe come to Columbia, Mo., to learn computer-assisted reporting at boot camps sponsored by IRE and NICAR several times each year.

Now editors and other newsroom managers can get into the CAR game on April 16-18 with an abbreviated version of the regular boot camp that is specifically tailored to their needs.

Managers will learn how to make CAR succeed in their newsrooms. They'll get just enough hands-on experience to understand what their reporters and producers are doing, and a taste for what more is possible. They'll hear from others who have experience editing and guiding computer-assisted reporting. They'll also learn how to negotiate for electronic records.

In just three days, editors will learn the skills that will allow them to enhance news coverage back home.

For additional information and a registration form visit *www.ire.org/training/bootcamps.html.*

Looking for training that's more specific to your newsroom? IRE and NICAR can arrange specialized training for news organizations inside or outside the United States. For more information e-mail training @ire.org or call (573) 882-2042.

continued on page 4

INSIDE NICAR

Getting ready for the big show

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

This is a busy time of year for us here at IRE and NICAR, as we get ready for the Annual Computer-Assisted Reporting Conference. This year we're meeting in Cincinnati, where former Reds star Pete Rose, who is back in the news, barreled his way to countless head-first stolen bases.

That means, by the time you read this, the staff around the IRE and NICAR offices will be charging just as hard to put together the best possible program for the journalists and journalism educators from around the globe who attend this event.

We're also putting together a roster of hands-on classes – everything from spreadsheets to industrial-strength database managers – taught by some of the best journalists in the business. There's a beginner's track for newcomers who want to learn CAR that will include special panels and a mini-boot camp of hands-on classes. Another track, for broadcast journalists, will address the unique challenges of getting CAR stories on the air.

On Thursday afternoon there will be pre-conference sessions for journalists with advanced CAR skills.

There's a lot on tap for everyone, beginner to expert. If you haven't registered for the conference yet, now's the time. You can register by using the form on page 3.

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

Hands-on CAR training

IRE and NICAR have a number of training opportunities coming up for journalists who want to learn more about computer-assisted reporting.

IRE and NICAR will conduct intensive CAR Boot Camps for journalists who want to learn how to use spreadsheets and database manager to produce high-impact news stories. The dates are March 21-26, May 16-21 and Aug. 1-6 in Columbia, Mo.

Journalists can also get hands-on CAR training at some of the Better Watchdog Workshops offered by IRE and the Society of Professional Journalists. Upcoming workshops with optional CAR sessions include Feb. 7-8 in Fullerton, Calif., and Feb. 21-22 in Austin, Texas.

The IRE Annual Conference, June 17-20 in Atlanta, also will have plenty of CAR offerings. The optional CAR day (June 17) will include panels for beginners and experts alike, and IRE and NICAR will offer hands-on classes in spreadsheets and database managers throughout the conference.

2004 Computer-Assisted Reporting Conference

Cincinnati, Ohio • March 12-14, 2004

Don't miss panels by the best in the business on every beat and topic including Census, campaign finance, crime, education, business, local and state government, freedom of information, transportation and more. Hands-on classes will be offered featuring spreadsheets, database managers, mapping, statistics, social network analysis and more.

PLUS! Exhibitors, networking, software demonstrations and special tracks for beginners, educators, broadcasters and advanced users.

For more information visit www.ire.org/cincy04

Conference Location and Hotel:

Hilton Cincinnati Netherland Plaza Hotel
35 West Fifth Street, Cincinnati, OH 45202

To make a hotel room reservation, call 513-421-9100 or 1-800-HILTONS and ask for the CAR room block to receive the discounted rate of $99 single/double occupancy plus tax.

Last day to make reservations at the discounted rate is Feb. 13.

Local Host: *The Cincinnati Enquirer*

Sponsors: *The Cincinnati Post/The Kentucky Post*, WCPO-Cincinnati

Presented by Investigative Reporters and Editors, Inc. and the National Institute for Computer-Assisted Reporting

---

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

Name: _____

Employer/Affiliation/School: _____

Address: _____

_____

City, State: _____ Zip: _____–_____

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

Home Phone: _____ Fax: _____

To register, mail this form and a check to IRE, 138 Neff Annex, Missouri School of Journalism, Columbia, MO 65211. To register by credit card, you must have a VISA or MasterCard. We cannot accept American Express. You may fax your credit card registration to 573-882-5431 or register online at **www.ire.org/training/cincy04**.

Cancellations need to be sent via e-mail to jgreen@ire.org. There is a $50 processing fee for all cancellations until March 11, 2004. Refunds will not be given for cancellations after March 11.

Card Number: _____ Expiration Date: _____

Card Holder Name: _____

Card Holder Signature: _____

To attend this conference, you must be a current IRE member through April 1, 2004. Memberships are nonrefundable.

MEMBERSHIP

____ I am a current member of IRE through April 1, 2004.

____**$50** I need to join/renew my U.S. or international ❑Professional ❑Academic ❑Associate or ❑Retiree membership.

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

REGISTRATION

____**$150** I would like to register for the conference as a ❑Professional ❑Academic ❑Associate or ❑Retiree.

____**$100** I would like to register for the conference as a student.

LATE FEE

____**$ 25** For registrations postmarked or faxed after Feb. 27.

TOTAL $ _____

Bits & Bytes

continued from page 2

Lovesick teachers?

Does something about Valentine's Day make public school teachers ill? Sounds a bit unromantic, but producer Ron Regan of WEWS-Cleveland – with the help of IRE and NICAR Database Library analysts – found that Feb. 14 was a day many public school teachers called in sick.

Based on the analysis of how public school teachers used their sick days, WEWS reported on patterns of abuse in its local school system and estimated the costs to taxpayers in the Cleveland area.

Check out the full story at *www.newsnet5. com/news/2656080/detail.html.*

Loans updated

For those who want to know more about home mortgage lending practices – including demographics of loan applicants – in their area, IRE and NICAR offer the 2002 update of the Home Mortgage Disclosure Act database.

The IRE Beat Book "Home Mortgage Lending" by Jo Craven McGinty is a great guide for journalists who want to learn how to analyze the data and detect compelling news stories. Order the book online at *www.ire.org/store/books* or call (573) 882-3364.

For more information about the HMDA data or to order contact the Database Library at (573) 884-7711 or (573) 884-7332.

SPOTLIGHT: CENSUS REPORTING

Delving into ancestry with SF4 data

By Paul Overberg, *USA Today*

Think of the Census 2000 summary files (SF1 through SF4) as a choir that sings the story of your community.

The full choir is SF1, all the voices in simple melodies. SF3 is smaller, chosen from the full choir to simplify the director's task of blending representative voices in more complex melodies.

By design, some voices emerge at times from each choir to sing solo, or in counterpoint. These are the separate tabulations by race and Hispanic ethnicity. Their stories ring clearly.

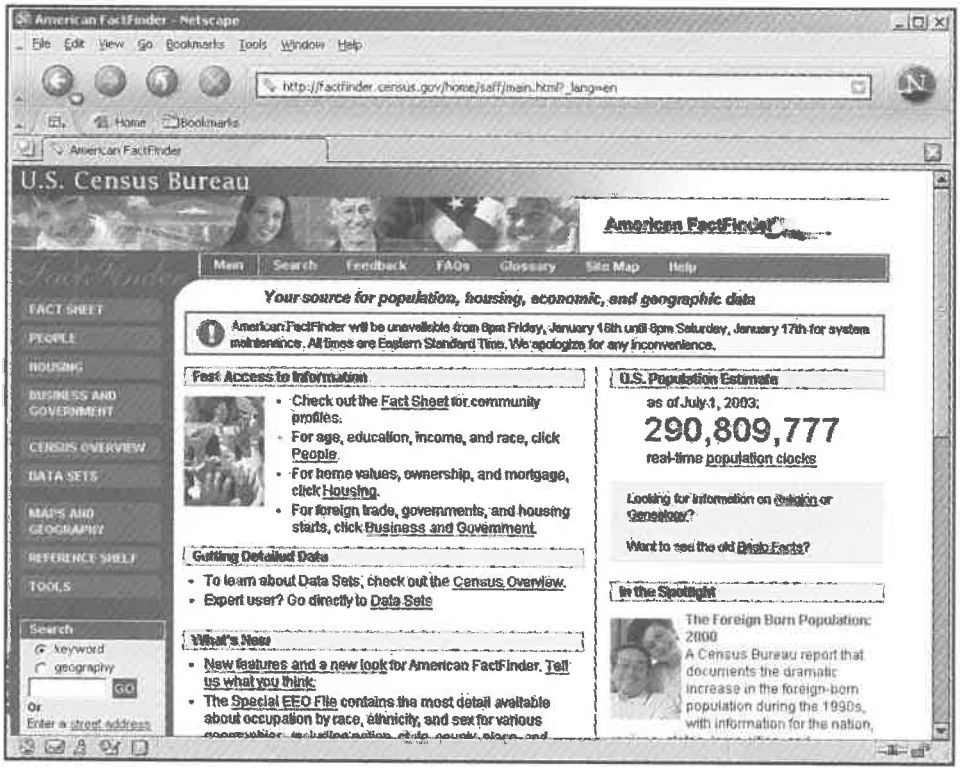
But we know the choir has more distinct voices embedded in its harmony. If we could hear them alone, we would hear their stories clearly. We could even hear the subtle differences between the melodies of Welsh and Scots, Kiowa and Comanche, Bulgarians and Romanians, Lebanese and Syrians.

Hence SF2 and SF4. In these files, each voice can emerge to sing solo. Are you a good enough director to lead them?

Profiling 336 groups

SF4 is simply a cut-down version of the tables released as SF3 but repeated, again and again and again, for each of up to 336 racial, ethnic, tribal and ancestry groups.

For journalists, these files are challenging for their size and structure alone, even with the help of American FactFinder (see page 11). Then there's the conceptual hurdle. More than one puzzled editor has asked: "It's the same stuff that we just got a few months ago, but cut up differently? Why do we care?"



A handful of journalists in the know have soldiered on, producing insightful stories about their communities.

• Robert Gebeloff and Brian Donohue of *The Star Ledger* examined New Jersey's Filipinos, whose median household income had climbed to the highest ranks. "We had not really written much about Filipinos as a group, and so this was a perfect excuse to do a profile," he said. The newspaper also printed charts profiling each group.

• Janet Roberts at the *St. Paul Pioneer Press* focused on the state's Hmong, a group of Laotians resettled mostly in the Twin Cities of Minnesota after the Vietnam War. With Kermit Pattison, she profiled a group whose ranks grew to 43,000, more than doubling in the decade. Hmong made big strides in catching up to state averages in homeownership, literacy and income.

• In a more modest and lighthearted effort, my colleague Haya El Nasser and I used SF4 data on Austrians and Greeks living in California to paint a quick profile of two groups with ancestral ties to candidates in the state's recall election for governor: Arnold Schwarzenegger and Arianna Huffington.

Remember that the rules of sampling error apply...

SF4 doesn't isolate every group. Roberts said that she had to scrap plans for a story on local Somalis, a newer refugee population, after discovering that most of the Minnesotans who listed their ancestry as "Subsaharan Africa" were actually U.S. born.

Broadening the focus

And there's a threshold to prevent disclosure of data about individuals. If the Census Bureau figures that there aren't at least 100 of a group in an area – whether it's a Census tract, city, metro or state – no SF4 tables are produced for that group in that area. Since SF4 is based on the long form, which was not distributed to all households, no tables are produced for any area where the bureau received fewer than 50 forms on which someone from that group was checked off.

This means you may have to shift your focus to a broader scale than you'd prefer to profile small groups. It also means that you need to remember that the rules about sampling error apply, just as they do with SF3.

continued on page 10

Recent *Uplink* Census stories

Uplink regularly features articles that provide tips from journalists about using Census data. It also highlights stories that have used Census data analysis as a key element in the reporting. Here are some recent articles:

"Ticket data shows search rate gaps." Karisa King of the *San Antonio Express-News* recounts how she used city police traffic-stop and Census data to uncover disparities in search rates by race. (May-June 2003).

"Mapping the Census" shows how journalists used geographic information system software to map Census data and discover patterns in poverty levels, non-English speaking areas and commutes. (November-December 2002).

"Children: tracking quality of life." Dave Davis of *The* (Cleveland) *Plain Dealer* explains how he used Census data along with Ohio health and environmental data to reveal how Cleveland children are living in worse conditions than other areas in Ohio and other major cities nationwide. (September-October 2003).

"Using Census files to link 1990 and 2000 geography." Dan Keating of *The Washington Post* explains one way to deal with Census geography shifts using tract-to-tract relationship files. Keating provides step-by-step guidance. (March-April 2003).

"Segregation in black, white, shades of gray." Holly Hacker of IRE and NICAR offers advice on using Census data to measure segregation and explore the racial make-up of cities, counties and other geographic areas. (May-June 2003).

"Plotting race by population." In the "Mapping It Out" feature, *Uplink* illustrates three journalists' stories that use Census data and GIS to map patterns of blacks and whites, looking for segregation and migration changes. (May-June 2002).

"Covering income, commuting with SF3." Jeff Porter, director of the IRE and NICAR Database Library, provides a guide to the Summary File 3 release from the 2000 Census and highlights some stories that used the data. (November-December 2002).

Compiled by Megan Clarke, *IRE and NICAR Database Library.*

MAPPING IT OUT

The latest uses of mapping in news reporting.

Showing deadly crashes

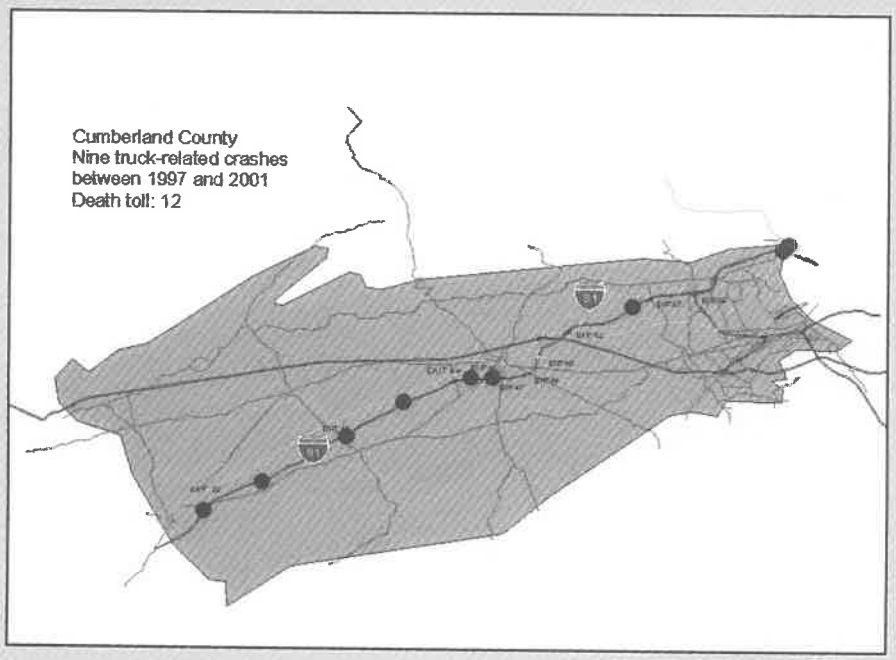
By Garry Lenton
The (Harrisburg, Pa.) Patriot-News

Pennsylvania isn't called the Keystone State for nothing. It's hard to reach any location in the northeastern United States without crossing the state.

And nowhere is this phenomenon more apparent than in the state's capital city, Harrisburg. This town is like the hub of a wheel, criss-crossed by three major highways, and railroads heading off in all directions.

This fact is not lost on the trucking and warehousing industry, which has become one of the largest employers in the region. Truckers can reach every major market in the eastern United States within 10 hours. But, along with the high-paying jobs that the industry brought, is an unwelcome consequence: fatal crashes with 80,000-pound trucks at almost three times the national average.

My editors wanted to get a handle on how our region compared to others when it came to fatal crashes with large trucks. So we turned to the U.S. Department of Transportation's Fatality Analysis Reporting System (FARS) database – purchased from the IRE and NICAR Database Library – for help.



Finding death counts

It took a while to get familiar with the tables, but Database Library staffers and the DOT helped us sift out the data we needed – the number of accidents involving tractor-trailer trucks that occurred in our region, and the death count.

But the FARS data could only take us so far. Our editors were impressed, but not really sure what they were looking at. Frankly, I wasn't so sure myself. We needed to focus the story on something readers could understand. That's when George Ginter, the assistant city editor overseeing the project, chipped in: What's the most dangerous highway we have? The answer was easy: Interstate 81. The question then became, just how dangerous is it? And how do we communicate that to readers?

The answer to the second question turned out to be a map.

But before we could turn to ESRI ArcView 3.2, our geographic information system software (GIS), we had to go back to the FARS data.

We aggregated the data in Microsoft Access to get a table that I could map. That took grouping the deaths along I-81 by the county of the accident. The result was the hook for the story.

From 1997-2001 no county along the 855-mile highway had more deaths from accidents with tractor-trailers than Cumberland County, which is the heart of our circulation area.

In our area, 40 percent of all people killed in crashes with a truck died on I-81. In the process of investigating the story we also found that the proportion of trucks on the highway was approaching 43 percent in some regions. That meant that the slightest error in judgment by a driver could result in a fatal accident with a truck.

Motorists who used the highway told us that they hated it. Too congested, too many trucks. But they also told us something else: It was the passenger-car drivers they most often saw driving erratically, not the truckers. The data verified that and we interviewed the authors of studies showing that car drivers accounted for the 80 percent of all crashes between trucks and cars.

Maps verify problems

The map helped us to bring that message home to readers most directly.

The first step in the GIS analysis was to chart I-81's 855-mile course from the New York-Canada line into Tennessee.

I used only three layers to build the map: roads, counties, and state boundaries – all files that came with ArcView.

First, I mapped all the U.S. counties. This was overkill, but I later pared this down to just the counties that I-81 passes through.

Next I added the roads shapefile. This turned the map nearly black with a web of lines representing everything from the simplest two-lane highway up to the interstates. Using the theme properties options, I filtered out everything but I-81.

I learned that it helps to spend some time scrutinizing the road layer's data table first to see how the highway is identified in each state.

Inconsistent names

Sometimes local authorities rename portions of highways, rendering the Name field useless as a filter. I tried filtering by Name and got just a dotted line for my trouble.

But I-81 was named consistently in the Sign1 and Sign2 fields of the table. So I used this syntax to get the entire highway:

```
([sign1]="I81") or
([Sign2]="I81")
```

Now I had a single red line stretching across a mosaic of multi-colored counties. But what we wanted was a map of the highway and just the counties it passed through.

I identified those counties by using the Theme\Select by Theme menu item. I activated the roads layer and then used Select by Theme to have ArcView highlight each county feature that I-81 intersected.

Then I opened the table for the county layer and hit the promote button on the tool bar. This moved all of the selected features to the top of the table.

Now it was time to create a layer of the selected counties. I closed the data table and selected Theme\Convert to Shapefile from the menu and added the new file to my view.

Then I added the deaths data table to my project, joined it to the counties map and, using the legend editor, shaded each county along I-81 by its death count.

For reference, I added the state shapefile and, in the theme properties, selected the six states that I-81 passes through.

I exported the map to a PostScript file and e-mailed it to the newspaper's graphics editor. The end product was a centerpiece map graphic for the story, which attracted a lot of positive attention from readers.

Contact Garry Lenton by e-mail at glenton@patriot-news.com

Would you be willing to share a mapping example with fellow journalists? Send an electronic copy of the map along with details to David Herzog at dherzog@nicar.org

SPOTLIGHT: CENSUS REPORTING

Measuring migration with PUMS data

By Ron Shawgo, *The (Fort Wayne) Journal Gazette*

If someone would have asked me a year ago what PUMS stood for I might have said a "Pretty Ugly Mess of Statistics," with the meaning of the "s-word" negotiable.

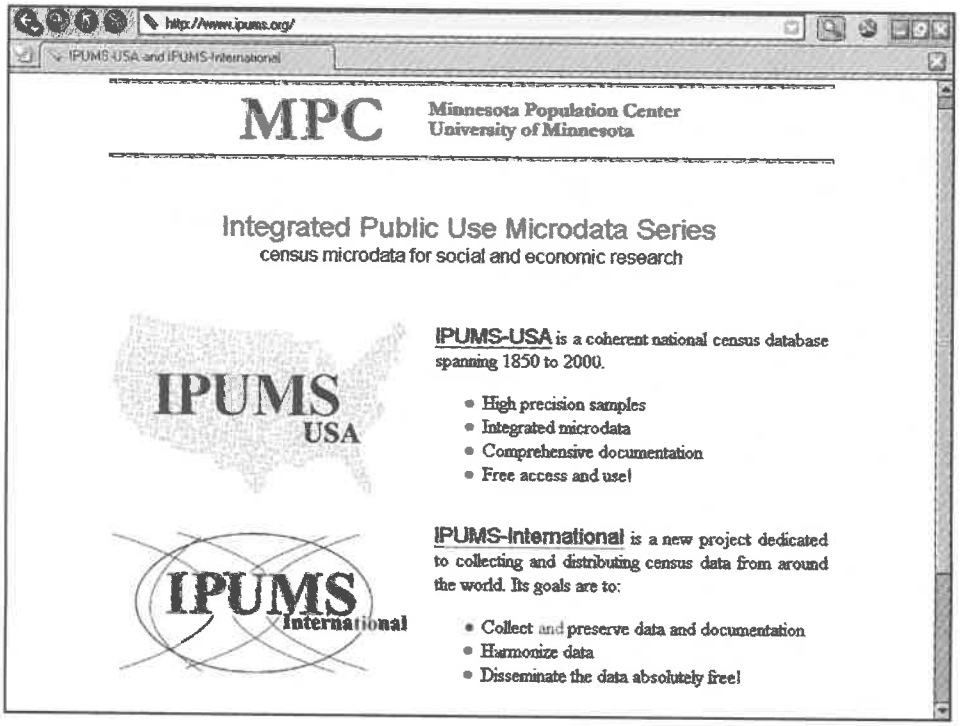
I've known for some time that the U.S. Census Bureau's Public Use Microdata Sample allowed users to dive deeper into the seemingly bottomless pit of Census data. But after three years as my newspaper's Census coordinator and writing numerous Census stories, I was ready to relegate PUMS to the "Isn't That Nice" pile. What more could be written?

That changed when journalists like Robert Gebeloff of *The* (Newark N.J.) *Star Ledger*, Mark Houser of Pittsburgh's *Tribune-Review* and *The Charleston Daily Mail*'s Brian Bowling mentioned stories using PUMS on the IRE and NICAR listservs. All three wrote about population migration, in particular about the loss of people to other states.

Borrowing from their ideas, I later told our readers that Indiana had the largest estimated net loss of college-educated 25- to 34-year-olds of any state between 1995 and 2000. It would have been difficult to do the story without the Integrated Public Use Microdata Series (IPUMS) maintained on a user-friendly Web site by the University of Minnesota. More on that later.

In barest terms, PUMS data contains responses from individual Census returns with identifying information such as names and addresses removed. The results are weighted so the data from a single return might represent 100 people.

Journalists and other researchers unsatisfied with the tabulations the Census Bureau routinely releases can pick and choose the information they want for their own tabulations. Simply sum the number of people in the weighted column and you're on your way.



PUMS comes in two flavors: 1 percent and 5 percent samples of the population. The 5 percent files are much larger and were just being released as I began my research. I chose the 1 percent sample for that reason and because it was available on the IPUMS Web site. Mindful of sampling error in the data I went no smaller than the state level for my analysis.

Downloading PUMS

The Web site, *www.ipums.org*, allows users to download extracts containing just the information they need, including figures for all states. This ability to compare states is important. Otherwise, journalists looking at state trends would have to download large files for all 50 states (51 including Washington D.C.) from the Census Bureau Web site. I don't have that kind of computer storage space.

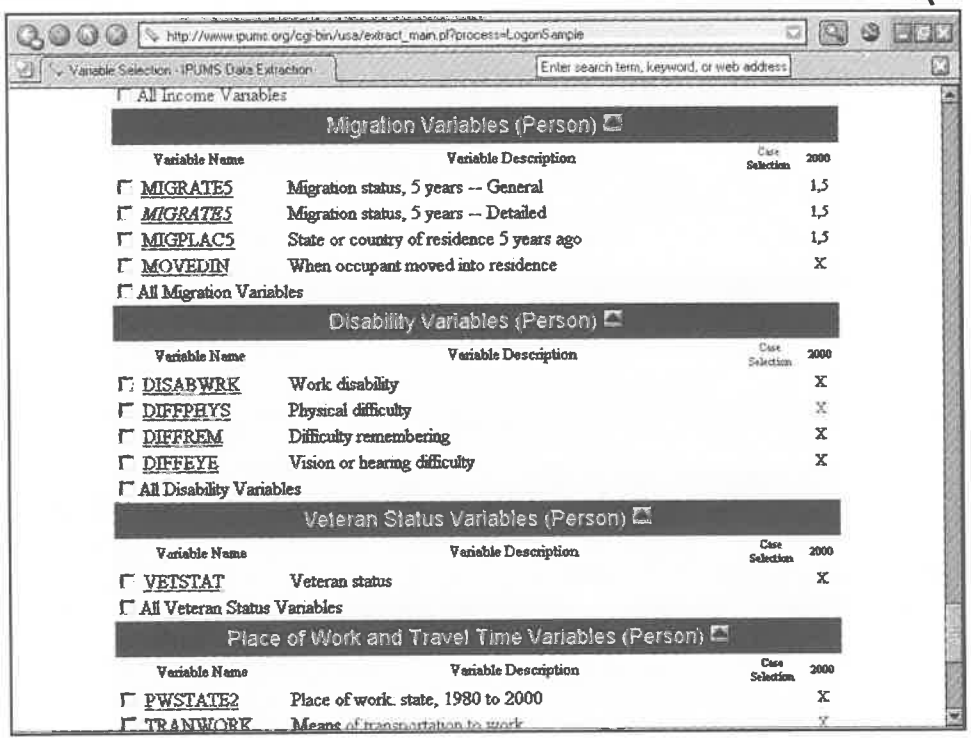
One caveat: IPUMS data file options are SPSS, SAS and Stata, all statistical programs. Anyone using the site should read the online documentation.

IPUMS is a free service, but registration is required. The IPUMS folks ask for proper attribution and a copy of any story published, and require those who download information to "Use it for GOOD — never for EVIL." A fair deal, I think.

After registering, you start the IPUMS download at the "Create New Extract" link. The page asks for your e-mail address and lets you chose the type of file you want (regular density for the 1 percent sample). Selecting "Only the most commonly requested samples," will serve up the 2000 PUMS data as an option.

Click "Continue to Sample Selection." On this page I chose a rectangular file in SPSS format and checked the box for "2000 1% Census PUMS."
Click "Continue to Variable Selection." This is where you'll choose the information in which you are interested.

For my statewide migration story I chose STATEFIP (state FIPS code), PERWT (the weighted number of people for a given variable), AGE (I also checked the "case selection" box to the right), BPL (birthplace – general), EDUC99 (educational attainment), MIGRATE5 general (migration status five years ago) and MIGPLAC5 (State or country FIPS code of residence five years ago). I also was interested in income so I chose INCTOT (total personal income) and INCWAGE (wage and salary income).



Click "Continue to Case & Flag Selection." Here I chose "Include only those persons meeting case selection criteria." Because I marked the case selection box for AGE on the previous screen, the Web site now gave me the option of selecting the ages I wanted. I chose 25 through 34.

Click "Continue to Extract Request Summary." Users are given the chance to make changes and write a brief description of the data before clicking "Submit Extract Request." The data will be delivered to your e-mail account. It's been my experience that the data usually is delivered within 10 minutes. A modem connection might be slower.

The e-mail includes a link to where the data can be downloaded. The files come zipped, so you'll need a program to expand them. I easily opened my data in SPSS. Because I'm not yet proficient in that program I then saved the data as dBASE file, which I imported into Microsoft Access.

Filtering and querying

I ran a query to extract from the EDUC99 field those with at least a bachelor's degree. For this and other operations you'll need to know IPUMS codes (14 = bachelor's degree, 15 = master's degree, etc.). Read the online documentation.

To find how many college-educated 25- to 34-year-olds left Indiana between 1995 and 2000, I filtered out Indiana in the STATEFIPS column but used it as the criterion in the MIGPLAC5 column. In other words, I asked to see those folks who lived in Indiana in 1995 but left by 2000. Summing the PERWT column gave me 58,887. Another caveat: It's impossible to get the number who moved abroad.

To find how many entered the state, I did just the opposite. I used Indiana as the criterion in the STATEFIPS column but filtered it out of the MIGPLAC5 column. Still, this didn't give me an accurate answer. People who lived in the same house in Indiana (those recognized as not applicable to my query criteria) also popped up. MIGPLAC5 assigns "0" to those not applicable folks. After filtering them out, summing the PERWT column gave me 41,782.

I did this for every state, placed the results in Microsoft Excel and sorted on the difference between those who entered and left. Indiana's estimated net loss of 17,105 was the largest.

I used STATEFIPS (where people lived in 2000) and MIGPLAC5 (where they lived in 1995) and BPL (birthplace) to track the migration of native Hoosiers for a sidebar, an idea I stole from Robert Gebeloff, who had looked at New Jerseyans. I was able to compare education and income levels of native Hoosiers who still lived in Indiana with those who left the state. I also found that some of Indiana's largest salaries go to residents who were born elsewhere.

With the numbers crunched I was able to find native Hoosiers living in other states by tapping into Purdue University and Indiana University alumni clubs listed online. Club presidents in other states were usually willing to talk or forward me to someone who would.

Contact Ron Shawgo by e-mail at rshawgo@jg.net.

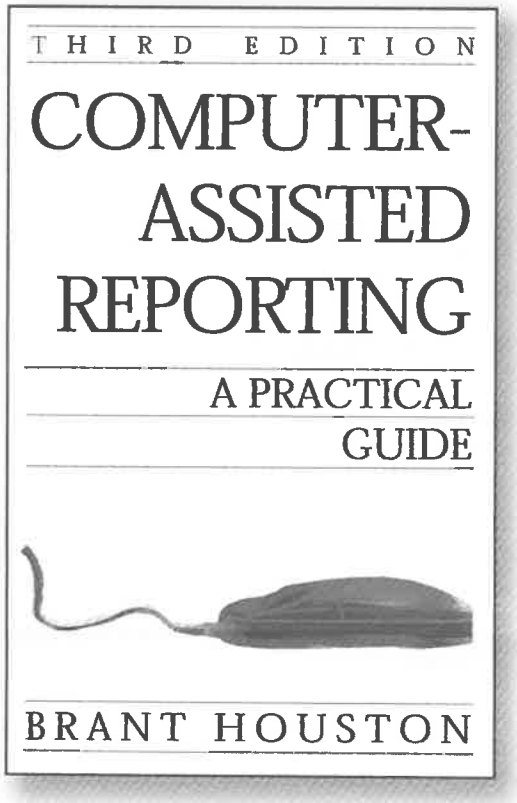
readme.txt

For a detailed introduction to PUMS read Tipsheet No. 1614 by Richard O'Reilly of the *Los Angeles Times*, available to members at no cost through the IRE Resource Center Web site at **www.ire.org/resourcecenter**. Others can order a copy of the tipsheet by contacting the Resource Center at 573-882-3364 or rescntr@ire.org.

COMPUTER-ASSISTED REPORTING:

A PRACTICAL GUIDE

by Brant Houston



NEW EDITION!

From surfing the Internet for data to number crunching in spreadsheets and database managers, "Computer-Assisted Reporting" is the most practical introduction to the data analysis skills that journalists need to succeed in today's newsrooms. This edition expands the highly regarded basic skills instruction and offers more advice on how to write better news stories.

This edition provides four chapters of tutorials, advice, and practice on spreadsheets and database managers, more than any other CAR book.

"Computer-Assisted Reporting" shows journalists how to find data on the Internet, use spreadsheets and database managers, build databases, and clean "dirty data" to analyze information faster and better.

Features

- Supplementary exercises and links to related Web sites
- Comprehensive and current technical coverage
- Field-tested reporting and writing advice
- Suggestions for using CAR skills in daily and beat reporting
- Illustrated tutorials, real-life examples, clear writing and more than 250 screenshots
- End-of-chapter and online exercises reinforce the skills demonstrated in the text.

New to the Third Edition

- New "CAR Wars" boxes showcase working journalists using CAR skills.
- New supplemental chapters on spreadsheets and database managers
- More instruction on how to accurately and ethically turn data into sources, story ideas, and finished stories
- Appendices offer introductions to hardware and software selection

ORDER NOW

By phone: Call 573-882-3364 with your VISA or MasterCard
By Web: Visit www.ire.org/store to order online or download an order form.
By mail: Send check to IRE, 138 Neff Annex, Missouri School of Journalism, Columbia, MO 65211

Plus postage: Media rate — $4 for the first book, $2 for each additional book

$32 members • $39 nonmembers

www.ire.org/carbook

SF4

continued from page 5

In general, don't make a big deal out of small differences.

But that still leaves lots of possibilities. Some brainstorming can help. Here are some starter ideas:

- You want to compare the social and economic status of members of an American Indian tribe who live on a reservation, which is tabulated separately in SF3, from members of that tribe who live across your state.

- Half a century ago, one neighborhood in your city was predominantly Italian-American. Today, the grandchildren have scattered to suburbs across three counties. SF4 would let you take a statistical group portrait, or contrast them with the Italian-Americans who still live in the old neighborhood.

- You have a hunch that members of the local Korean community are splitting into two groups, one that is intermarrying and one that is not. You'd like to look at their rates of linguistic isolation side by side.

- Immigrants from three Caribbean countries live in your city. You'd like to contrast their rates of homeownership and poverty.

- You'd like to contrast your city's new wave of Irish immigrants from the descendants of the great wave of Irish immigration that occurred in the middle of the 19th century.

Sometimes it's better to brainstorm the other way, by starting like this: Which groups in my community are the most important? What do we know about them? What more do we want to know? Could this data help?

Whichever approach you use, SF4 can probably provide deeper insight.

Contact Paul Overberg by e-mail at poverberg@usatoday.com.

Download SF4 data easily

By Paul Overberg, *USA Today*

The best place to start learning about the potential of SF2 and SF4 is American FactFinder. Even if you've downloaded and tamed the hundreds of tables for your state, FactFinder is sometimes the best way to get a quick profile of a group from SF2 or SF4. This example uses SF4, the data and groups tabulated from the long form, but you can adapt it to SF2.

Start at FactFinder's home page, *http://factfinder.census.gov.* Click on the box labeled "Data Sets". On the page that loads click on the radio button for SF4, and from the options that appear to the right, click on "List all tables."

In the page that loads, you'll see profiles (DP), quick tables (QT), geographic comparison tables (GCT), then a long list of population (PCT) and housing (HCT) tables. For simplicity, choose DP-4 and click on "Next."

When the page reloads, use the top drop-down list to choose the geography you're interested in. For this example, choose "Place."

When the page reloads, choose "California" from the state drop-down list, wait for the page to reload and pick "Los Angeles city" from the place list.

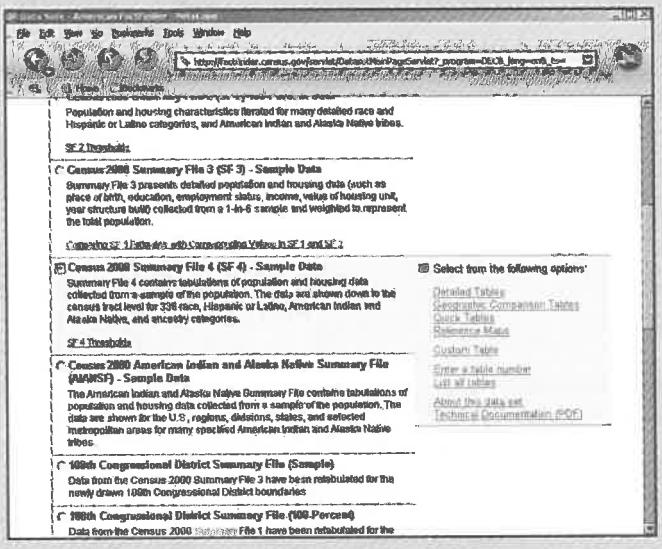
Click the "Add" button below to push it into the selection box below. Then click on the "Show result" box.

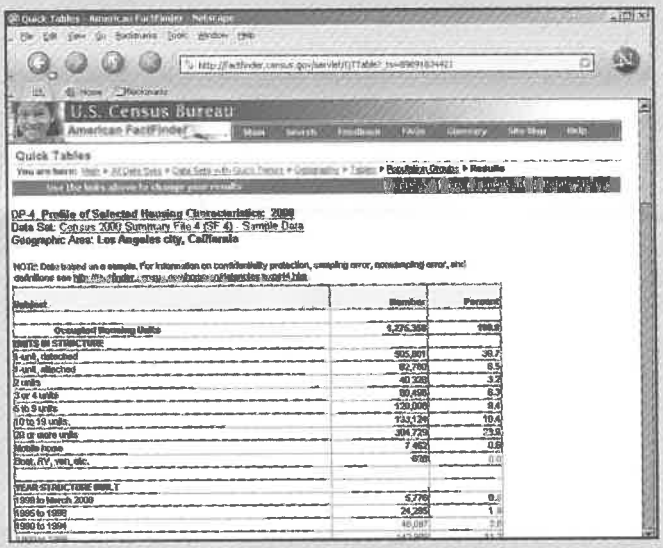
When the page reloads, you see Table DP-4 for Los Angeles. Ignore it and look across the top of the page, where you'll see "You are here:" followed by a series of hyperlinks back into each step we've just taken. The last live link is "Population groups." Click on it and a two-tab page loads.

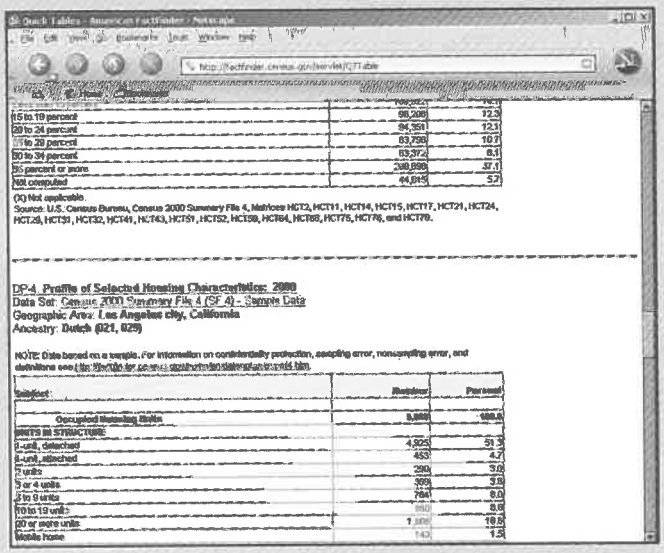
We made it! These are the groups for which every table of SF4 is separately produced, from the whole nation down to just your census tract, as long as there are 100 of that group.

Click on the "Ancestry groups" tab, then click on "Dutch," "Add," and "Show result."

When the page reloads, it redisplays DP-4 for Los Angeles and below, DP-4 just for the 16,156 Angelenos who filled in "Dutch" on the ancestry question. From this point, use the Print/Download button at the top of the page to shape your output.







Woes

continued from page 1

match our observations.

The phenomenon: If you are a member of the black middle class in Chicago, you are more likely to live closer to poverty and crime than someone in the white middle class.

Lo and behold, the numbers said we were right.

We discovered that while they fell into the same economic bracket, blacks and whites in the middle class still lived in vastly different social worlds. And that discovery led to the following nut graphs:

"The analysis showed that, for blacks, just rising into the fiscal middle class does not transport them into the idealized American dream: a home neighborhood well insulated from poverty, blight, crime and other societal ills.

"In seemingly stable middle-class black neighborhoods like Roseland, Chatham, South Shore and Austin, such social disarray is usually just a few blocks away in either direction. But in many predominantly white middle-class neighborhoods, particularly on the Northwest Side, such disorder can be miles away and largely confined to an isolated pocket.

"The new census data allow for refined demographic snapshots of areas as small as a handful of blocks averaging 1,175 residents. The *Tribune* analysis showed that 78 percent of such black middle-class block groups in Chicago were within half a mile of block groups where at least one-third of residents lived in poverty. By contrast, fewer than 25 percent of white middle-class block groups in Chicago were within half a mile of areas with poverty levels that high."

Key questions

Coming to this conclusion had its share of challenges. Our first hurdle in examining the socio-economic conditions of the black middle class was arriving at a specific definition of "middle class."

We contacted the Census Bureau, whose experts advised against using a local measure of median income to define middle class for Chicago-area residents. The bureau recommended using the figure of $40,000, approximately the median income for the U.S. population and thus a good figure to define the economic middle class. (The actual median household income for the U.S. is $41,994). Actually, our story would look at the middle- and upper classes, since we did not define an upper limit for income.

The data used for the analysis came primarily from the Summary File 3 and, to a smaller extent, Summary File 4 releases from Census 2000.

The smallest geographic unit reported in the SF3 data is the block group. After consulting several Chicago-area demographers, we decided to use block groups for the analysis. To reduce statistical error in dealing with the sampled SF3 data, we examined only block groups with at least 100 households.

We defined a black or white block group as an area where at least 70 percent of the population is of that particular race. The figure of 70 percent was arrived at after, once again, consulting with local demographers. Chicago is highly segregated between blacks and whites. Blacks are equally segregated from Latinos and Asians.

We found that the bulk of the black middle class block groups we looked at was at least 90 percent black, so defining a "black" neighborhood was fairly easy. Whites, who make up 31 percent of Chicago's population, are not quite as segregated from Latinos and Asians. After some discussion with demographers, we decided on 70 percent as good measure of a "white" neighborhood.

From the Census Bureau's FTP site, we downloaded population, income and poverty data for Illinois. Using Microsoft Access, we calculated the racial makeup of each block group. We then looked at the median household income for each racial group within that block group. Using these figures, we identified all of Chicago's black and white middle-class block groups.

We also calculated the poverty rate for every block group in the city. We identified all block groups where the percent of individuals living in poverty was greater than or equal to 33 percent. A block group where at least a third of the population lives in poverty struck us as an area easily identifiable by readers as "high-poverty." After identifying all of the middle-class and high-poverty block groups, we were then ready to examine the results spatially.

Mapping neighbors

In ESRI ArcView 8.2, we displayed the black and white middle-class block groups, as well as the high-poverty block groups. We then created a half-mile buffer around the boundary of each middle-class block group. ArcView could then identify each block group whose buffer intersected a high poverty block group. From these spatial queries, we could count the number of black and white middle-class block groups that are within a half-mile of high-poverty block groups.

The release of Summary File 4 allowed us to provide an even more detailed portrait of Chicago's poor. With the newly released data, we could calculate the percentage of Chicago's poor whose income is less than half of the poverty level, the percentage of poor people without a high school diploma and the percentage of poor people residing in female-headed households.

Once we harvested the data, it was time to look inside the black community for people who matched what we found in the data – to put a face on the story, as we like to say in this business. As with most sweeping demographic patterns, the people who produce those patterns

are not all that hard to locate.

With the reporters already having some sources on Chicago's Far South Side, the Roseland neighborhood seemed like the perfect place to base the story. Roseland was a formerly white neighborhood that became nearly all black during Chicago's white flight era of the 1960s and 1970s.

Today, the neighborhood possesses blocks and blocks of stable, middle-class homes. But it also has occasional streets of blight and poverty.

We interviewed Roseland residents who were involved in their local block clubs and seemed to have a good sense of their community. When one of the subjects told us about the nicknames of local streets, the lede for the story was sealed:

"Adelaide Randall, who for 37 years has lived on a serene street in the Far South Side neighborhood of Roseland, refers to several less-than-serene neighboring blocks by their local nicknames: Death Row, Skid Row and Thriller.

" 'That block has quieted down,' she said of Thriller, officially known as 101st Place. 'But we still need to get more active on the others.'

"Randall believes that neighbors have made enough phone calls to police and absentee landlords that the street is showing fewer signs of troublesome activity recently: fewer idle young men hanging around, fewer overt signals of drug peddling, less trash in vacant lots.

"But the fact that Roseland's middle-class residents must be so vigilant about maintaining their surroundings illustrates a disquieting fact of life for Chicago's black middle class: Financial security does not necessarily buy neighborhood tranquility."

Contact David Mendell by e-mail at dmendell@tribune.com

Contact Darnell Little by e-mail at dlittle@tribune.com

Upcoming Workshops

Advanced Statistics Workshop

Feb. 13-15 • Arizona State University

Pulitzer Prize-winning journalists Steve Doig, interim director of the Walter Cronkite School of Journalism at Arizona State University, and Sarah Cohen, database editor of *The Washington Post*, will team to teach this workshop. This session is aimed at strengthening the skills of reporters who want to move beyond basic computer-assisted reporting and use statistical analysis in their work. Reporters should know spreadsheet and database manager applications and have experience in computer-assisted reporting.

Fellowships

A limited number of minority and small-news organization fellowships are available for IRE and NICAR workshops. Visit www.ire.org/training/fellowships.html for more information and an application.

Better Watchdog Workshops

Investigative Reporting on the Beat

Investigative Reporters and Editors Inc. and the Society of Professional Journalists, with funding from the Sigma Delta Chi Foundation, have joined forces to offer a series of workshops focused on doing investigative reporting while covering a beat.

The workshops, specifically for journalists at small- to medium-sized news organizations and those in bureaus of larger organizations, will emphasize the use of freedom-of-information laws and address juggling a beat while producing investigative and enterprise pieces.

"You'll learn enough in the first 15 minutes to keep you busy for a month."
Kevin McGrath, *The Wichita Eagle*

Workshops are scheduled for:
Feb. 7-8 – Fullerton, Calif.
Feb. 21-22 – Austin, Texas

For more information, visit www.ire.org/training/betterwatchdog

To request a workshop for your area, send an e-mail to watchdog@ire.org.

More information is available at www.ire.org/training

Tech tip...

Cell phone lookup

By James E. Wilkerson, *The (Allentown, Pa.) Morning Call*

A stack of paper records is a funny thing to computer-assisted reporting types – it can evoke feelings of longing, fear and loathing at the same time.

You look at it and think, "Imagine what great stuff is in there," but recoil at the thought of getting all of the information on them onto your hard drive.

At least, that's how I felt when our city reporter dumped photocopied pages of phone records on my desk, after getting the billing records for all city-owned cell phones.

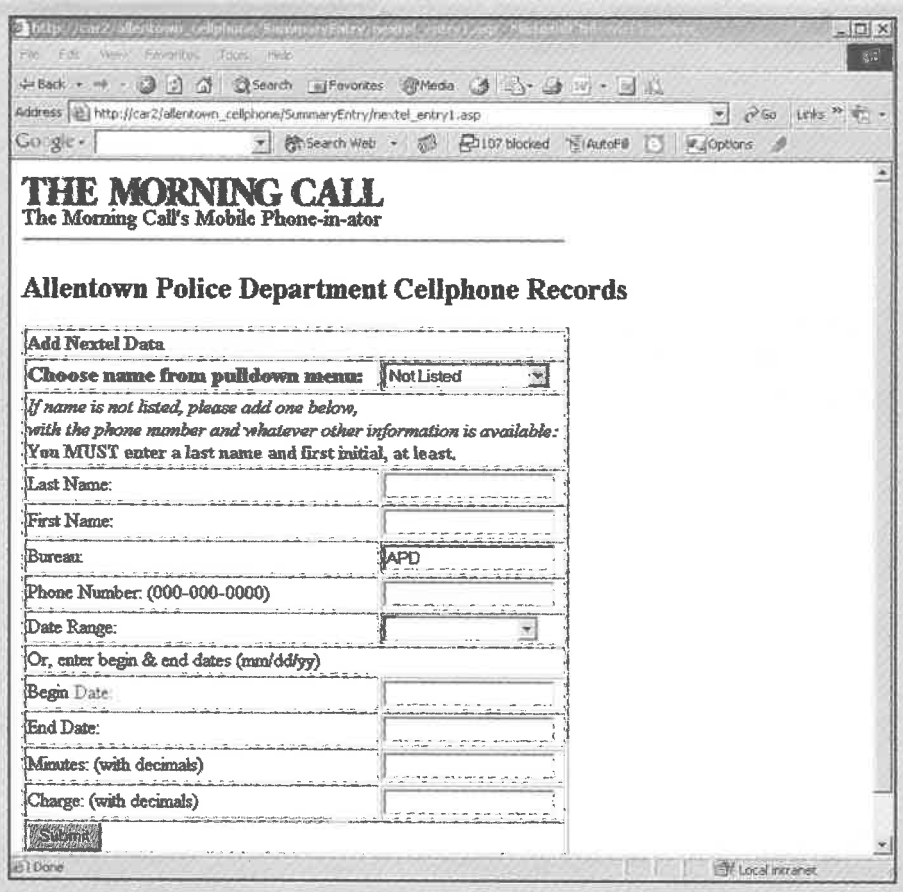
We hadn't looked closely at government phone records in a long time, so I was excited about what we might find. But as I started looking through the summary records, I had trouble figuring out how to make sense of them.

First, the records were from two different phone companies, and in very different formats: administrative numbers were listed by cell phone number, and police numbers were listed by the name of the user. The city had also blacked out some of the police numbers and written in department names instead to hide potentially sensitive information.

Those and other complications meant that I would have to incorporate several different approaches in cataloging the information. I decided to do it through an internal Web site.

I've been increasingly choosing internal Web sites over Microsoft Excel or Access, because they give me more flexibility to create custom forms that can easily be accessed by anyone in the newsroom or our bureaus.

I display the data on the Web sites using Microsoft's Active Server Pages (ASP) language. I have a Windows 2000 server that houses my intranet, and program my Web sites using Microsoft Visual Studio. (You can get started with ASP using a much simpler setup. We'll talk more about that later.)



After playing around with a few different approaches, I decided to build separate entry pages for each cell phone account. I tweaked them so that names or phone numbers that had already been entered could be chosen through a drop-down menu. That saved a lot of data-entry time.

It took about a week for a news assistant to enter the data. While she worked, I developed increasingly grandiose plans to use the site for analyzing records from other government agencies, and built a few simple display features to allow reporters easy access to the data.

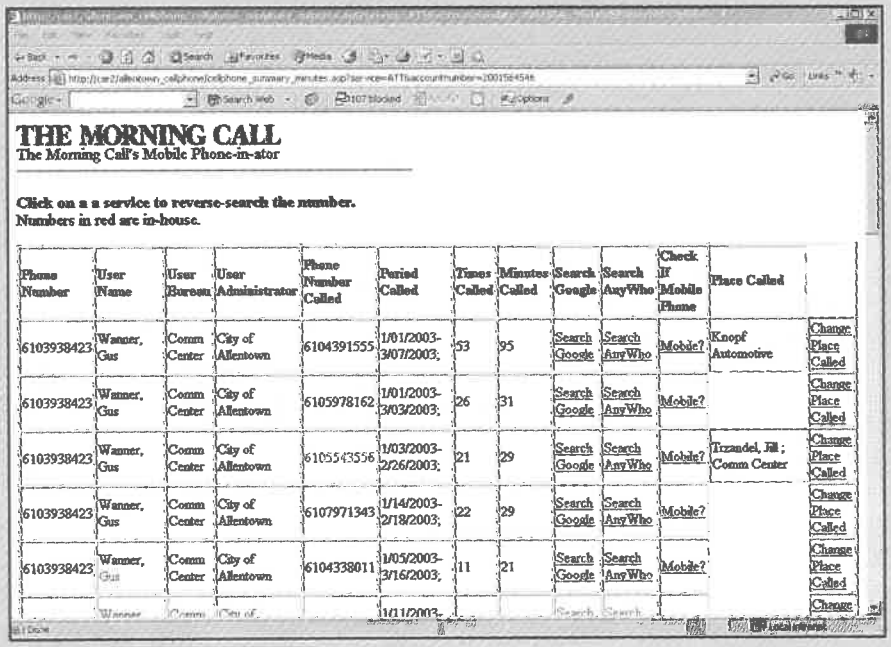
Once the Allentown data was entered, I found the top users in both accounts, and we requested detailed records for them. The city quickly gave us the bills for administrators' phones, but balked at providing police phone details,

saying that investigations could be compromised. We continue negotiating for that data.

I scanned in the detailed bills they did provide, and after cleaning up the information in Excel, imported the lot into SQL Server. Then, it was a piece of cake to display the most-called numbers.

All that was left was figuring out to whom they belonged.

I turned to ASP again, and built a page that allows users to display the most-called numbers for each individual. But I added a twist – I built links to online reverse directories into the page, with the phone numbers embedded in the link, so that searches could be run on a **click of the mouse.**



THE MORNING CALL
The Morning Call's Mobile Phone-m-ator

Click on a a service to reverse-search the number.
Numbers in red are in-house.

Phone Number	User Name	User Bureau	User Administrator	Phone Number Called	Period Called	Times Called	Minutes Called	Search Google	Search AnyWho	Check If Mobile Phone	Place Called	
6103938423	Wanner, Gus	Comm Center	City of Allentown	6104391555	1/01/2003-3/07/2003;	53	95	Search Google	Search AnyWho	Mobile?	Knopf Automotive	Change Place Called
6103938423	Wanner, Gus	Comm Center	City of Allentown	6105978162	1/01/2003-3/03/2003;	26	31	Search Google	Search AnyWho	Mobile?		Change Place Called
6103938423	Wanner, Gus	Comm Center	City of Allentown	6105543556	1/03/2003-2/26/2003;	21	29	Search Google	Search AnyWho	Mobile?	Trzandel, Jill ; Comm Center	Change Place Called
6103938423	Wanner, Gus	Comm Center	City of Allentown	6107971343	1/14/2003-2/18/2003;	22	29	Search Google	Search AnyWho	Mobile?		Change Place Called
6103938423	Wanner, Gus	Comm Center	City of Allentown	6104338011	1/05/2003-3/16/2003;	11	21	Search Google	Search AnyWho	Mobile?		Change Place Called
	Wanner	Comm	City of		1/11/2003-			Search	Search			Change

And you don't have to build a Web page to do the same thing. Most spreadsheet and database programs will do just fine.

Go to your favorite reverse directory, run a search and see if the code for the search appears in the subsequent address.

I like AnyWho, which displays a reverse search of my office number like this: (Area code and phone number in bold for emphasis.)

```
http://www.anywho.com/qry/
wp_rl?npa=610&telephone=
8206770&btnsubmit.x=34&btnsubmit.y=5
```

Many programs, such as Excel and Access, can be used to build links using that model. Just replace the area code and phone number with the new number, and you're ready to go.

In Excel, with the area code in column A and the rest of the number in column B, it would look something like this:

```
="http://www.anywho.com/qry/
wp_rl?npa=" & A2 & "&telephone=" &
B2 & "&btnsubmit.x=34&btnsubmit.y=5"
```

Apply that to all the phone numbers, and you've got an easy-to-use research tool.

To cap off that research, I added a field to my page that allowed users to cut-and-paste the results of the searches. I was able to identify more than half of the top numbers in less than an hour using the site, which was much easier than looking each up by hand in our city directory.

For those interested in ASP, I highly suggest *http://4guysfromrolla.com*, which provides a bunch of information for newbies and experts alike. "Active Server Pages 2.0 for Dummies," $17.49 at Amazon, also provides very simple, step-by-step instructions on building ASP pages.

And you don't have to have complicated, expensive programs like SQL Server or Visual Studio to serve data or build pages – Access works just fine, and basic ASP (Not ASP.net, which is the newest version and has a steeper learning curve) can be written alongside HTML using any text editor or Web authoring program. Tutorials for creating ASP connections to Access and other database programs are available all over the Internet.

As for server space, check with your tech department to see whether you can piggyback on an existing setup. But if a server isn't available to you, Microsoft does offer a scaled-back server that will run on its desktop software. Search for "Personal Web Server" on Microsoft's Web site for more information.

Contact James E. Wilkerson by e-mail at james.wilkerson@mcall.com

CAR TOOL

Cluster analysis finds what's nearby

By Holly Hacker, *IRE and NICAR*

A health reporter might examine a town near a chemical plant where several residents have contracted cancer. A crime reporter might search for patterns in car thefts. A demographics reporter might look into new pockets of Latino neighborhoods in the region.

In any of these cases, journalists could use a geographic information system to map features – cancer cases, crimes, Census blocks – to better understand and explain the story. With help from a GIS the crime reporter might write, "Most cars were stolen within a half-mile of City Hall."

That's a good story, but it leaves questions unanswered. Do those thefts form a distinct, meaningful pattern? Or could they have happened where they did by chance?

A statistical tool called cluster analysis could fill in those information gaps. In the most general terms, cluster analysis is a way of placing people, places or events into groups. In geographic terms, then, cluster analysis searches for patterns in space.

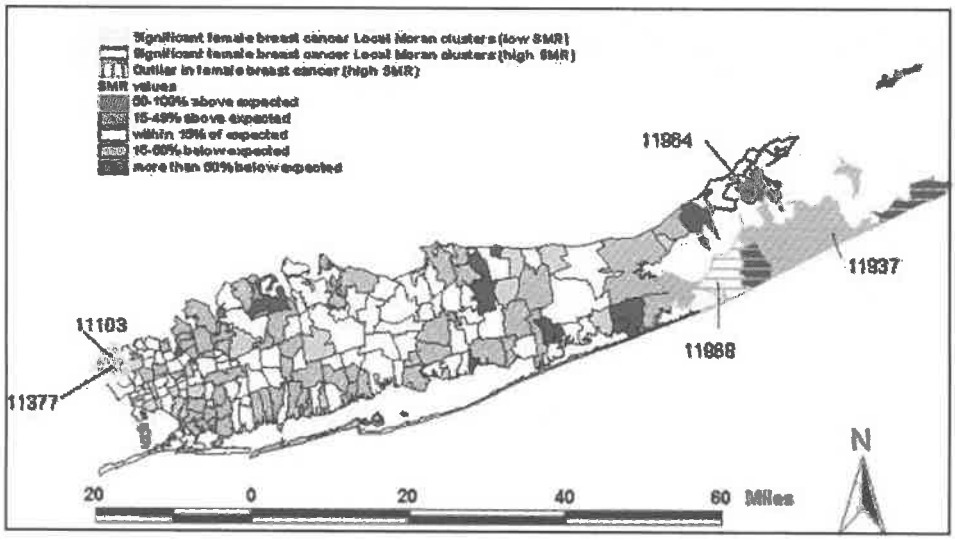
ClusterSeer, a software program from Michigan-based TerraSeer Inc. (*www.terraseer.com*), finds spatial patterns in data and determines their significance.

ClusterSeer helped *Newsday* make sense of cancer clusters on Long Island for a series in 2002 (read "Tattered Hopes" online at *www.newsday.com/ny-canceronlongisland0728.htmlstory*).

Clusters in time, space

For that series, researchers at TerraSeer used the software and performed the analysis. Journalists who want to do their own cluster research need to be ready to learn some sophisticated statistics – and have a boss willing to shell out $2,100 for the software.

ClusterSeer uses statistics to find and evaluate clusters in space, time or both. It was originally created with grants from the National Cancer Institute to study disease clusters. That's still a primary use, though not the only one. The same type of analysis could be done with demographic, crime or business data, to name a few areas.



It is a stand-alone program, but it accepts files from GIS, text and database sources. It also saves projects, exports images and loads data to supplement other maps.

While a GIS presents geographic data, ClusterSeer interprets it, said Nicholas Jacquez, TerraSeer's president and chief executive officer. Map crimes, liquor stores or other events in a GIS, and likely they will be concentrated in some areas. But, Jacquez said, "Unless you test it statistically, all you're looking at is a coincidence in space."

In cluster analysis, the items of a given group should be as similar as possible to each other, and as different as possible from items in other groups.

The criteria for grouping items vary. One could group drugs by their chemical properties, or consumers by their income and shopping habits. Used with GIS, cluster analysis involves classifying features based on distances, so points or other features close to each other end up in the same cluster. It also allows items to be categorized by their distance to a source of risk, such as cancer cases in relation to a nuclear power plant.

A look at cancer

For the *Newsday* report, TerraSeer researchers studied for free the distribution of breast, lung and colorectal cancers in Long Island and Queens between 1993 and 1997. New York state has identified a few possible clusters including one in Suffolk County.

Newsday environmental reporter Dan Fagin wrote that the state Health Department's methods "are so ineffective that they have very little chance of distinguishing a true cluster from a statistical fluke, and almost no chance at all of ever identifying a cluster's real cause."

Fagin said he wanted to show readers that there are better methods for study-

ing cancer clusters. Checking with sources led him to Geoffrey Jacquez, the chairman and chief scientist of TerraSeer (and Nicholas' brother).

"The more I talked to him the more I thought, well, this is a relationship that could actually work," Fagin said. "He was all about not only understanding the science, but also applying it."

Using state Health Department data, Geoffrey Jacquez and fellow scientist Dunrie Greiling found nine clusters. In a story, Fagin described them as "groupings from three to 50 ZIP codes where rates of one of those cancers, in men or women, were so consistently high between 1993 and 1997 that they were almost certainly not caused by chance" (read the story about the analysis at *www.newsday.com/news/health/ny-licanc132821454aug13.story*).

Geoffrey Jacquez and Greiling also tried to pinpoint possible causes of the cancer clusters. They obtained data on air pollution sources from the National Air Toxics Assessment, gathered by the U.S. Environmental Protection Agency. They used another TerraSeer product, BoundarySeer, to compare the boundaries of the cancer clusters with the location of pollution sources. The result: a strong correlation between air pollution sources and lung cancer, but not breast or colorectal cancer.

Not proof

Fagin stressed that the analysis did not prove that air pollution causes lung cancer. But it did lay the groundwork for future research. Fagin said the analysis done in ClusterSeer had sound scientific basis, unlike politically motivated studies by public agencies.

The *Newsday* series won the 2003 Science Journalism Award from the American Association for the Advancement of Science, and the Science in Society Award from the National Association of Science Writers.

Fagin said he wanted to show that valid research could be done on cancer clusters. But he doesn't recommend that journalists simply jump into this kind of research themselves.

"Definitely, people should use these amazing tools. But they should be very careful setting themselves up as an armchair epidemiologist," he said. Fagin said he prefers to collaborate, as he did with Geoffrey Jacquez and Greiling, on such complicated issues.

"I would never feel comfortable doing that on my own, without working with a qualified professional."

Indeed, the user guides for ClusterSeer can be intimidating to statistical novices. Chapters tackle techniques such as Diggle's Method, Kulldorff's Scan and Getis-Ord Local G. Many journalists, meanwhile, continue to grapple with basics like percent change.

Then again, other journalists use ArcView or SPSS without understanding the math behind a Kernel density function or Pearson's product-moment correlation. They just need to know how to use those tools. So, while it helps to have a basic understanding of spatial autocorrelation (the degree to which neighboring areas are alike) one needs not know all the mathematical calculations behind it. Plus, ClusterSeer has an online help feature that says if, for instance, you're searching for disease patterns on a local level, you should use the local Moran test (a measure of spatial autocorrelation).

On a related note, journalists need to recognize what the software can't do. Just because cancer cases are clustered near an oil refinery, that does not prove cause. In addition, cancer can take years to develop and require prolonged exposure to carcinogens. People can move or change jobs in that time, making spatial relationships harder to find. Without such awareness, a journalist using fancy statistical software could come to faulty conclusions.

Another issue is, even with the statistical knowledge, whether newsrooms will invest $2,100 for a commercial license – especially if they already have invested in a GIS. The two clearly have different purposes. A GIS presents and analyzes geographic data, while ClusterSeer detects and analyzes patterns in time and space. More advanced analysis in a GIS often requires add-on programs that can run into the thousands of dollars. So either way, news organizations that want to do more than make pretty maps may have to invest serious dollars.

Contact Holly Hacker by e-mail at holly@ire.org.

readme.txt

Have a favorite CAR tool that *Uplink* readers should know about? Drop us a note at uplink@nicar.org and tell us about it.

MOVING?

Please send us your new address so you don't miss one issue of *Uplink*.

Drop us a note at
jgreen@ire.org

Investigative Reporters and Editors, Inc.

OPEN SOURCE

Programming a zip with PHP

By Aron Pilhofer, *The Center for Public Integrity*

I love Perl. It's powerful, free and incredibly flexible – truly the Swiss Army knife of programming languages. But frankly, Perl is a pain to learn.

I've tried many times, but I've never quite gotten over the hump. I'll be clipping along just fine when I run into something like this:

```
$_~=/^\.*?\$$/;
```

Crystal clear, right? Not to the beginner, and that's the problem with Perl. It's a quirky language loaded with shortcuts, abbreviations, special characters and more exceptions than rules.

It's not that you can't write clean, readable code in Perl; it's that so few developers actually do. Perl programmers seem to value brevity over readability. And for those of us who learn programming by taking examples and deconstructing them, that's a problem.

More often than not, I find myself falling back to another language I've learned to love because it makes complex tasks so simple: PHP.

It's a programming language that offers all but a handful of the features of Perl without the steep learning curve. The reason is that PHP was designed from the ground up to be clear and concise, with as little ambiguity as possible.

More than any language I am familiar with PHP exemplifies a tenet of the open source community: simple tasks should be simple to accomplish, and complex tasks should be only as complicated as needed.

Why scripting?

Play around with data long enough, and eventually you'll run into a problem you can't solve with structured query language (SQL) alone.

Maybe you've got a directory full of inconsistently formatted text documents you'd like to clean up. Or possibly you have data stored in a series of spreadsheets and dBASE files you want to import into a single file. Perhaps you'd like to mine a number of Web sites and pull data from them into a single location.

All are tasks easily accomplished in PHP.

Recently I decided to automate the tedious process of pulling weekly updates to a federal agency database down from the Internet, importing it into an Microsoft SQL Server database, cleaning and parsing several inconsistently ordered name fields, then finally coding the data based on our internal scheme.

I certainly could have written a Data Transformation Services (DTS) package in SQL Server to handle some of the process, but my work would have been limited to that particular server and that specific platform. And all that work would be lost if we switched to another database back-end or, as many corporations are doing, to a non-Microsoft platform.

That's the advantage to languages like PHP (and Perl, for that matter): you can write a program once and run it anywhere.

Why PHP?

PHP was originally designed to make programming the Web easier, and in that capacity it excels.

That's one reason its popularity has skyrocketed. In just the past five years or so, PHP has come from relative obscurity to be one of the most popular scripting languages in the world. It has done so because it is simple, powerful and totally free.

Like Perl, Java and Microsoft's C#, PHP's operators and control structures are derived largely from the C programming language. Anyone familiar with C or its derived languages (such as C++) would have no trouble picking up the basics of PHP. (In fact, most of the basic language structures – control loops, operators, variables – are virtually identical to Perl's.)

The difference comes in the way PHP handles functions. Unlike most newer languages, they are all built into the base language rather than residing in separate libraries and included into applications where needed.

And while that structure is a bit of a throwback, it does make programming in PHP a whole lot less confusing. Most languages today require at least some level of understanding of object-oriented programming, which is in my opinion extremely confusing for beginners.

PHP does not require applications to be written this way though it does allow it. Developers are given the option of writing in the Perl, C#, Java style, with objects and classes, or they may code in the more intuitive (at least initially) C style, called procedural programming.

For me, it was not until I started using PHP that I began to see the advantages of object-oriented programming, and why it is so popular. But it was a gentle ramp-up, rather than an abrupt shove.

Perhaps PHP's greatest asset for newer programmers is the fact that it is purposely more verbose than most languages in its naming of functions. Even if you run into a particular function or operator you are unfamiliar with, PHP almost always gives a clue about what it is and how it's used.

Functions related to manipulating external files, for example, all begin with the letter "f." To open a file you would

use the fopen() function. To read a file into memory, use fread(). To write to a file, use –you guessed it– fwrite().

PHP's native functions that communicate with a SQL Server database server all begin with "mssql," so it isn't hard to figure out what the function mssql_connect() does, or mssql_fetch_row() for that matter.

Take the following example:

```
$url="http://www.claptrap.org/
fec.html";
$fec = fopen($url, "r");
while ($line = @fgets($fec, 1024))
```

Even if you've never seen a line of PHP code in your life before now you can probably guess what's going on here. This code opens a Web site, and reads it line-by-line into a variable.

What's missing?

PHP has nothing yet that matches CPAN, the Comprehensive Perl Archive Network. CPAN is an enormous collection of useful code available for free download. Just about anything you can think of is in CPAN. There are, for example, more than 200 classes, or sets of subroutines, dealing with calendar functions.

Just because something is written for Perl doesn't mean you are totally out of luck; it is not hard to write a short Perl script to use a particular CPAN module, and then call that script from PHP. But doing so does have some limitations and will tend to slow things down.

PHP does have something like CPAN, but nowhere near as comprehensive. The PHP Extension and Application Repository, or PEAR, now contains some 240 packages – a fraction of what's available in CPAN. But every year, the PEAR archive is growing broader and deeper. Another resource is *www.phpclasses.org*, which has an even more impressive collection of user-donated classes.

PHP surpasses Perl in area: in some of the more popular database classes called ADODB. (Available at: *http://php.weblogs.com/ADODB*.) It is a single class that allows developers to write one application that will, with a change of just one line, work with dozens of databases – from SQL Server, to Microsoft Access, to Postgres, to Oracle. It is written to mimic Microsoft's ADO format, so anyone familiar with ASP or VB should have no problem figuring it out.

Conclusion

PHP is a simple yet powerful scripting language that can be used to easily automate many of the mundane tasks CAR folks do day to day. If you are interested in learning more, a place to start might be a tutorial, *www.oscar.org/oscar/tutorials/php/advanced%20scripting%201-0.pdf,* I wrote to import and parse a Federal Elections Commission Web page, or the PHP home page (*www.php.net*).

If you are looking for a scripting language that maximizes your productivity with a minimal amount of initial effort, PHP is by far your best choice.

Contact Aron Pilhofer by e-mail at apilhofer@publicintegrity.org

readme.txt

Uplink started reviewing open source software last year. Here are the programs we examined and when.

OpenOffice productivity suite (January/February 2003)

MySQL database manager (March/April 2003)

Front-end programs for MYSQL (May/June 2003)

Red Hat Linux 9 operating system (September/October 2003)

R statistical analysis program (November/December 2003)

Computer - Assisted Reporting

BOOT CAMPS

These unique seminars give journalists a jumpstart in computer-assisted reporting techniques. Participants come to Columbia, Mo., where they are trained in how to acquire electronic information, use spreadsheets and databases to analyze the information and to translate that information into high-impact stories. The National Institute of Computer-Assisted Reporting provides follow-up help.

- March 21-26
- May 16-21
- Aug. 1-6

More information is available at www.ire.org/training/bootcamps.html

Mapping

continued from page 1

make access to GIS data easier and faster for all people, not just GIS experts, said Hank Garie, executive director of the site. Along with development of the portal, the goals of Geospatial One-Stop include developing GIS data standards, forging partnerships with state data clearinghouses, and promoting cooperation among all levels of government – something that could save a lot of time and frustration for many journalists.

Let's say you are reporter in North Dakota and you want to find a map of state inland water boundaries for a story you're doing. Instead of using a search engine to sift through endless content on the Web for one geographic file, you could go to *www.geodata.gov*, click on the "Inland Water Resources" then click on the state link. In the list of data available, you'll see North Dakota's offerings. There you'll find a preview map, the metadata (or information describing the GIS data) and a direct link to the Web site where the layers originate.

While this data for North Dakota is currently available via the portal, there's a lot that's unavailable. Garie says this a shortcoming right now, but that eventually the site will have more links to data.

So, until that day comes, how can journalists find GIS data quickly? The same way they've been doing it for years — with a little bit of creativity and a lot of persistence.

Some journalists say GIS data is easier to find on the Web than other types of data.

"GIS people are far more likely — in my opinion — to post their data online (usually buried in deep, deep places on a governmental agency's Web site) than other data people. GIS people share more. It's part of their nature," said Matt Waite of the *St. Petersburg Times*.

The trick to finding the data is for journalists to jump in and get their feet wet. Use a favorite online search site and tweak searches to find the GIS data.

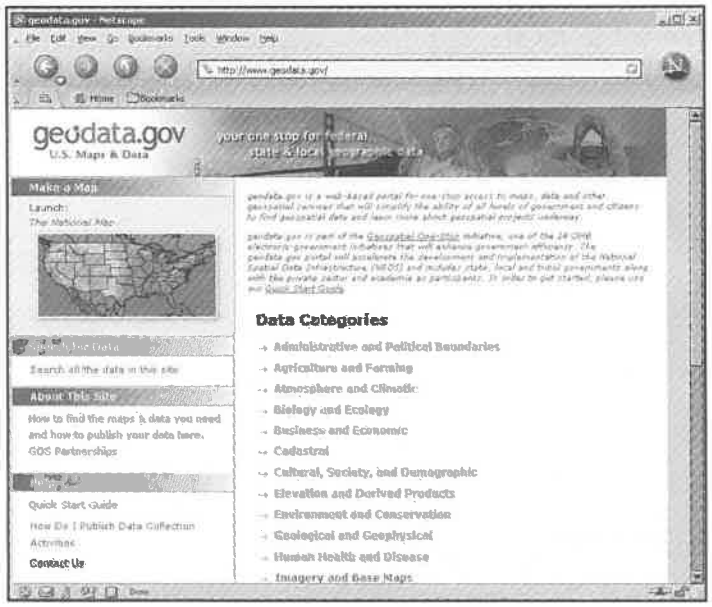
"I've been surprised at how often just punching .shp (or other file extensions) with a few other search terms into a Google search has brought up all kinds of interesting links for me," Waite said. (The .shp extension denotes a shapefile, which is used in ESRI's ArcView GIS.)

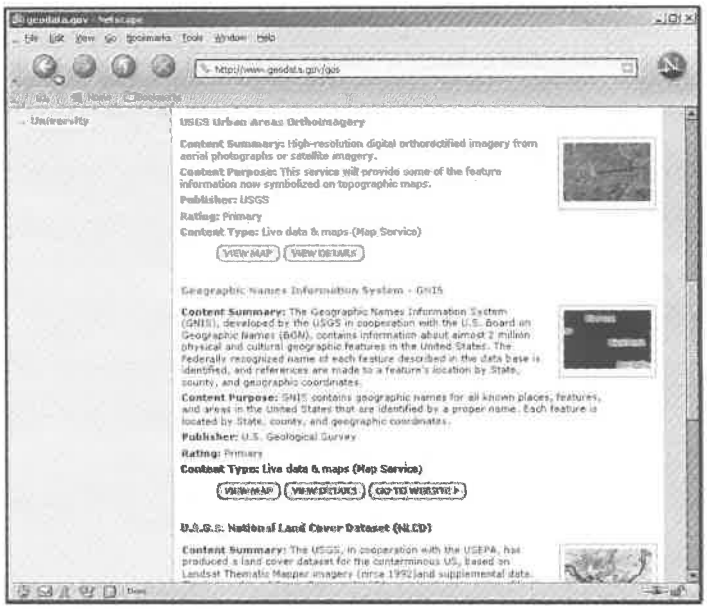
Another method Waite suggests is very similar. "Punch in 'GIS' and whatever it is you're interested in along with some geographic identifiers (GIS, housing, Florida, for instance) and see which agencies have a GIS department." This can help you locate hidden GIS data and the departments creating it. Then you can call up the departments and get the data yourself.

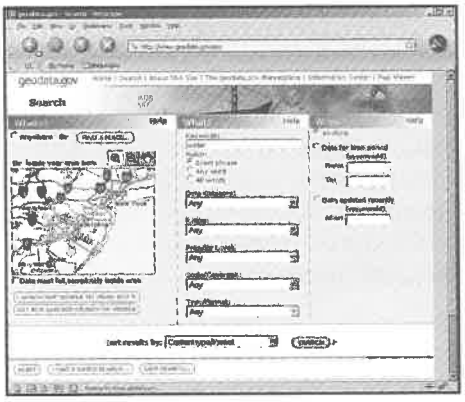
For journalists, preparation is key. Find out what GIS data is available and obtain it. Peruse government agency Web sites. Check out their Web site maps, looking for data or GIS-specific terms. Whenever you find GIS data, download it and save it for later. You never know when a map layer you downloaded on an off day might come in handy for breaking news.

While there are many places on the Web you might stumble upon or search for GIS data, there are plenty of other Web sites that were created to ease the process. For example, many states maintain GIS data clearinghouse Web sites. These sites typically compile GIS data from throughout the state and many have links to the data. Some of the sites require users to login; others are anonymous. More-developed clearinghouses have search options to help link users with the GIS data they are seeking.

The National Oceanic and Atmo-







spheric Administration Coastal Center has compiled a list of state clearinghouses that offer GIS data at *www.csc.noaa.gov/products/nchaz/htm/dinfo_6.htm*. A similar list can be found through the Massachusetts Institute of Technology libraries Web site at *http://libraries.mit.edu/gis/data/statecenter.html*.

Universities and other educational institutions are often a great source for obtaining GIS data. For example, the Michigan State University library provides a Web site (*www.lib.msu.edu/coll/main/maps)* complete with some boundary data, historical data and more. Downloads and metadata are available as well as a preview of the file. The school of architecture at the University of Texas-Austin also offers a Web site, (*www.ar.utexas.edu/Planning/data*), with a long list of free GIS data for the Austin area.

To find sites like these from your area, simply go to that university's Web site. Most educational institution sites have some sort of search engine. Use it and type in the words "GIS data" or any other terms you are interested in finding. This will usually direct you to a site where you can either find GIS data or the people from the university who are specialists in that area.

For example, if you're looking for GIS data and you're a journalist in Connecticut, you might try hunting down data at the University of Connecticut Web site, (*www.uconn.edu*). If you type "GIS data" in the search window you will be directed to a long list of sites, including a link from the University of Connecticut's library. The Map and Geographic Information Center, (*http://magic.lib.uconn.edu)*, offers shapefile downloads that include data at the state, county and town levels.

You also might want to try some federal government search engines. One of these sites is the Clearinghouse Search Wizard, (*http://130.11.52.184/servlet/FGDCWizard)*. This site takes users through a step-by-step search for GIS data. The data includes everything from political boundaries to tourism and recreation. While the interface is a bit archaic and clunky, this site might be useful for journalists wary of searching for GIS data all over the Web. A map is also available to help select the geographic location where you are interested in finding the data. International searches are available.

If you're interested in international GIS data, Dan Keating, database editor at *The Washington Post,* suggests checking out GIS Data Depot. This commercial site (*http://data.geocomm.com*) provides some free GIS data downloads while others are available for what the site claims is the "lowest price around, guaranteed!"

If all of this doesn't work and you're about to beat your head against the wall, don't. Pick up the phone and give the government agency a call. Better yet, make an appointment to meet with the agency's GIS guru – you might just walk home with a CD-ROM full of shapefiles and other useful data. GIS professionals are typically more than happy to talk to someone interested in their work.

As for the quality of GIS data on the Web, it's just like any other data. Some of it is of high quality, some is not. Check the source of the data and read the metadata. Just as with any other data you might use, do integrity checks to spot shortcomings.

Contact Jaimi Dowdell by e-mail at jaimi@nicar.org

SAVE THE DATE

ATLANTA

2004 IRE Annual Conference

JUNE 17-20, 2004

The best in the business will gather for more than 100 panels, workshops and special presentations about covering civil rights, public safety, courts, national security, the military, business, education, local government and much more. Don't miss IRE's premier event!

REGISTER NOW:
www.ire.org/atlanta04

Conference Hotel:
Atlanta Marriott Marquis
265 Peachtree Center Ave.
Atlanta, GA 30303

Hotel Reservations:
• $119 plus tax
• Hotel reservations must be made before May 21.

To make a reservation, call 404-521-0000 or 800-228-9290 and ask for the "IRE 2004" room block. To make reservations online, follow the Marriott link from IRE 's conference Web page and use online code IREIREA.

More information:
www.ire.org/training/atlanta04

Sentences

continued from page 1

free or serve little time. As scandals at Enron, Mississippi-based WorldCom and Birmingham's HealthSouth shook the financial world, we wanted to quantify that sentencing disparity.

It seemed especially relevant for Birmingham, where more than a dozen HealthSouth workers have pleaded guilty to fraud and await sentencing while federal agents continue their probe.

We immediately turned to TRAC, the Transactional Records Access Clearinghouse at Syracuse University (*http://trac.syr.edu*), which maintains federal prosecution and sentencing data and offers a subscription query service.

We began the computer work early in 2003, querying the TRAC databases for sentencing information on all categories of crime. We decided to look in a 15-year window. How many crimes were prosecuted in each of the nation's 90 federal districts since 1987? How many of those prosecutions resulted in a conviction, and of those, how many went to prison and for how long? We thought we could get by with a couple hundred queries from TRAC, so we entered into a special media rate agreement for about $300.

It took a while to learn the TRAC system and then the computer holding our data died. I had failed to back up the TRAC data so I had to start the queries anew. Then our contract ran out and we had to renew that for another $300.

(If I had to do it all over again I would likely purchase an annual license from TRAC, which would cost more than $1,000 but would provide more help and support.) We learned by trial, and plenty of error. But in the end we got the hang of it.

We copied page after page of query results, and pasted them into Microsoft Excel for further analysis. We focused on drug possession, bank fraud, official corruption and related offenses. We could tell immediately that our hypothesis had been on target.

White-collar criminals in North Alabama repeatedly walked away from convictions with little or no prison time during the past 15 years, the analysis showed. The average prison sentence for white-collar criminals in North Alabama was 10 months during that period, while the average drug sentence was eight times that. Thanks largely to federal mandatory minimum sentence guidelines, a charge of simple drug possession resulted in an average 68-month sentence, while bank fraud brought an average of eight months.

From a national perspective, North Alabama federal judges were particularly tough on drug crime and immigration violations, but light on white-collar crime, weapons charges and civil rights issues.

Finding cases

We began to look for examples, and found them with some difficulty. Using TRAC's Going Deeper query system we pulled individual case details for bank fraud and drug charges in the Northern District.

We were dismayed to find that the details lacked the case number and defendant's name. The data did, however, provide dates, sentences, the judges names and other pertinent information.

We could immediately tell that our hypothesis had been on target.

After compiling a list of relevant cases – drug possession charges with unusually long sentences or white-collar crime cases with little time served – we went to the federal courthouse to try to identify them. We had no luck. From the clerk's office to judge's chambers and on to data processing court workers told us the same thing: It can't be done without the case number or the defendant's name.

We didn't believe that, and it became clear we could figure out how to identify the cases. We turned to PACER, the online federal court docket index (*http://pacer.psc.uscourts.gov*) to "reverse engineer" the answers.

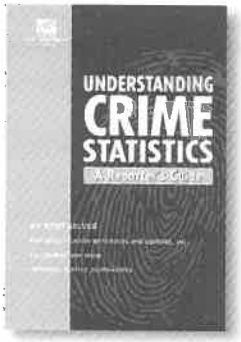
We took the cases one by one. Using the PACER search, we entered a date range that included only the date the case was filed and a couple of days after. In the required party name field we entered "USA." The search brought up every case in the district that began on the right day, usually several screens full. We scanned the list until we found the right one. We ultimately found dozens of cases in which defendants convicted of bank fraud and embezzlement spent less than a month in prison. It was a clumsy and frustrating process, but it worked: we found subjects for our story.

REQUIRED READING For Your Newsroom

Understanding Crime Statistics

A Reporter's Guide, by Kurt Silver

Know the data and reports available to add perspective and authority to your crime reporting. Learn how to use the data and avoid common pitfalls while writing responsible crime statistics stories.



ORDER TODAY!

Call 573-882-3364 with your VISA or MasterCard

-OR-

Visit our Web site at **www.ire.org/store** for online ordering or order form downloads

IRE MEMBERS: $15 each • NONMEMBERS: $25 each