Uplink

July • August 2004
Volume 16, Number 4

Published bimonthly by the National Institute for Computer-Assisted Reporting
www.nicar.org

ELECTIONS

Targeting voting wrongs

By Benjamin Lesser
The (Hackensack, N.J.) Record

Last year reporters at our newspaper learned that the U.S. Department of Justice was investigating possible voting rights violations in Teaneck, a small New Jersey suburb of New York. This January one of our reporters, Scott Fallon, approached me about doing something more substantive than the typical he said-she said story about the probe.

After some discussions with our editors we decided to give it a go. Our plan was to conduct a parallel investigation to the Justice Department's. A few months later we published a story that detailed for our readers what the Justice Department investigation involved, what conclusions the investigators may draw and ultimately what the investigation could mean for the town.

The Justice Department began its investigation after receiving an anonymous complaint from

continued on page 14

MODELING REALITY

Examining evacuation during nightclub fire

By Paul Edward Parker, *The Providence (R.I.) Journal*

Computer simulations, or models, are commonplace in the scientific world. They are used to predict hurricanes, forecast the spread of epidemics and calculate how long it would take people to leave a crowded building. Computer models are considered scientifically reliable, yet they are a rarity in the world of computer-assisted reporting.

In the course of covering the Rhode Island nightclub fire that killed 100 people in February 2003, *The Providence Journal* developed a database of people who were inside the building when it caught fire. This database was used for stories that documented how many people were inside the building, and for a narrative of the six minutes from when the fire started until the last survivor escaped. We also used it for a computer model that examined how the crowd fled the nightclub and how things could have been different with a few minor changes.

continued on page 20

SPOTLIGHT: REAL ESTATE

Property data analysis finds extra tax credits

By Kevin Corcoran, *The Indianapolis Star*

Thousands of Indianapolis taxpayers cashed in last year on multiple homestead property tax credits even though this valuable break was intended to reduce taxes only on a homeowner's primary residence.

After analyzing computer records for nearly 210,000 properties that received the break for owner-occupied homes we reported that more than 9,300 people benefited from as many as 11,465 homestead tax credits they weren't entitled to. These credits cut taxes on homes taxpayers no longer lived in – or had never lived in. Some taxpayers got five, 10 – even 20 – credits at once.

continued on page 12

SPOTLIGHT:
For more about reporting on homes see:

- Mapping finds Tampa Bay values surging, p. 6
- Defects uncovered in central Florida new homes, p. 8
- Analysis shows unequal taxation in New Orleans, p. 10

Bits & Bytes

Data updates

The IRE and NICAR Database Library now has Multiple Cause-of-Death Public Use (Mortality) data covering 1988-2001 and Occupational Safety and Health Administration (OSHA) data covering 1972-2004.

The 2001 mortality data contains over two million death certificate records from the United States and its territories. Mortality data has generated several stories about death trends in the United States. Combined with U.S. Census data, the mortality database can help reporters investigate how the economic health of a particular area contributes to mortality and life expectancy. See *www.ire.org/datalibrary/databases/health* for more information.

The latest OSHA data contains over three million inspection records and is current as of January 2004. OSHA data can be used to investigate which organizations violate employee safety standards, as *The New York Times* and Frontline recently did in an award-winning joint investigation of the McWane Corp. See *www.nicar.org/data/osha* for more information about the database.

A sure bet

Ever wonder how to turn complicated stories into interesting television journalism? Do you want to learn investigative skills that keep government and business accountable? Then join some of the best journalists from around the country at the **IRE Regional Conference** in **Las Vegas**, Sept. 10-12.

Sessions include presentations taken from the Better Watchdog Workshops sponsored by IRE

continued on page 4

INSIDE NICAR

New training director making mark

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

David Donald is the newest member of the team at IRE and NICAR. As the training director, he's been on the road during the past few months. Maybe you've seen him this year at the annual computer-assisted reporting and investigative reporting conferences, at a Better Watchdog Workshop or another training event.

Donald joined IRE and NICAR early this year and brings a wealth of experience from the newsroom. He oversaw the CAR and research program at the *Savannah Morning News* in Georgia, where he was precision editor. Before that, he covered education and served on the newspaper's projects team. He was the lead organizer of IRE's regional conference in Savannah in 2002, and many of you have seen him speak on panels at the Annual CAR Conferences, sharing his knowledge of statistics, SPSS, surveys and building a CAR program. Donald also brings his experience teaching in high school and university classrooms to his work for IRE and NICAR.

If you'd like to arrange on-the-road training – anything from an open workshop to a private seminar at your news organization – contact him at ddonald@ire.org.

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

Upcoming hands-on CAR training

IRE and NICAR have several training opportunities for journalists seeking hands-on instruction in using computer-assisted reporting.

There are four weeklong Boot Camps in Columbia, Mo., for journalists who want to learn how to acquire electronic information, use spreadsheets and databases to analyze the information and to translate that information into high-impact stories. IRE and NICAR provide follow-up help after participants return to their news organizations. The 2005 Boot Camps are Jan. 9-14, March 20-25, May 15-20 and Aug. 7-12.

Journalists interested in learning how to map data for news stories can take advantage of a mini-Boot Camp in using geographic information systems (GIS). The training, Aug. 20-22 in Columbia, will use ArcView 8.3 GIS. Participants will learn how to create basic maps, work with projections and geocode data. All participants will be eligible to purchase GIS software at discount. Additional GIS training is being planned for Washington, D.C.

Hands-on classes in basic CAR skills will be offered at some of the Better Watchdog Workshops sponsored by IRE and the Society for Professional Journalists. Two upcoming workshops with optional CAR training are Sept. 18-19 in Richmond, Va., and Nov. 6-7 in Oklahoma City.

FIRST VENTURE

New-teacher data boosts narrative project

By Tara McLain, *Statesman Journal (Salem, Ore.)*

I was several weeks into my narrative project that involved following a college student through 10 weeks of student teaching.

The student was 54 years old and teaching was his third career. Not a normal situation — so I thought. The data told me otherwise.

After attending an IRE and NICAR Boot Camp this year, I figured I could bolster my story with some statistics about how rare it is for older people to go into teaching. I called the Oregon Teaching and Standards Practices Commission.

Just as I learned in Boot Camp, the spokeswoman didn't know the data and the data guy didn't know policy.

After a lot of conversation and getting a few sets of data with the wrong categories, the data guy e-mailed a Microsoft Excel file containing information about the 10,000 teacher licenses given out last year. The columns of data included the teacher's name, school district, date of birth, date the license was issued and the type of license.

Then I got back on the phone with the spokeswoman and found out just a few of the dozens of types of licenses are given to new classroom teachers. The rest were renewals or went to principals or many other variations.

I used Excel to filter and sort my way to the new teachers. Then I needed to calculate their ages when they got the license.

Here's where a second lesson from Boot Camp came true: The program doesn't know the difference between dates, numbers and text unless you tell it the difference. I wrote a formula for Excel to subtract the date of license from the date of birth. It gave me a huge number. Excel thinks in days, so I divided the number by 365.

Then I noticed there were some doubled, tripled, and even quadrupled, entries. I started deleting a few then scrolled down to see many more.

That's where a third lesson from Boot Camp helped: There's always a shortcut.

There's always a shortcut.

This was my first CAR story and I couldn't think of a way to search and destroy the duplicate rows. So I e-mailed Jeff Porter, the IRE and NICAR Database Library Director. He provided, with lightning speed, the perfect formula. He said he stole it from someone but couldn't remember whom. With apologies to Jeff and to the person he stole it from, here it is:

```
=IF(AND(E2=E3,F2=F3,A2=A3,D2=D3),
"dup","not")
```

The formula tells Excel to compare information in each column by row. If a last name, first name, date of birth and date of license were all the same in one row as the row before, it would answer "dup". Then I filtered and sorted the names by duplicates and deleted them.

So a fourth lesson from boot camp came true: Cleaning the data can take longer than crunching the data.

Now, down to the analysis.

I made a pivot table of the count of teachers by age. I got one count of each teacher. Remember the dividing by 365? Excel was looking at the age of each teacher as a decimal, although I had rounded it up, and created a category for each.

I ran another formula (with the help of the research coordinator in the marketing department of our newspaper) that stripped everything off the number except the first two digits. So 35.4332355 in column H became 35 in column I after I entered the formula `=LEFT(H2,2)`.

I ran the pivot table again. I summed the counts into easy-to-digest age groups: 20 to 29, 30 to 39, and so forth.

Here's where the fifth lesson came in: Editors love pie charts.

The pie chart showed that half of the new teachers in Oregon last year were over 30. Not exactly your fresh-faced college undergrads.

I got the data from five years ago, went through all the cleaning again (much faster the second time) and found that the proportion of older people (30 plus) being hired to teach had grown from to 52 percent last year from 41 percent in 1999. That blew my assumption that all new teachers were young.

The officials said the growth probably was because of the stagnant Oregon economy in the past few years and loads of school budget cuts. Districts had their pick of new hires and were choosing older people with more life experience whom they could pay the same as 23-year-olds, they observed.

I found out that my 54-year-old student teacher isn't as rare in Oregon as I had assumed. My data-driven sidebar complemented my narrative feature.

I also found an 84-year-old teacher in my local district who just had her licensed renewed.

Contact Tara McLain by e-mail at
tmclain@statesmanjournal.com.

Bits & Bytes

continued from page 2

and the Society for Professional Journalists, in addition to several sessions for broadcast journalists. IRE and NICAR are also offering computer-assisted reporting classes on a first-come, first-serve basis.

The conference costs $50 for IRE members and $90 for nonmembers (includes $40 discounted membership). Students may attend for $25.

For more information, including hotel rates, please see *www.ire.org/training/vegas04.*

Campus crime

The 1990 Jeanne Clery Disclosure of Campus Security Policy and Campus Crime Statistics Act mandated all colleges report the number of crimes occurring on their campuses. But it wasn't until 2000 that these figures were centrally collected.

Senior producer Lora Johnson LeSage of WMAQ-Chicago recently enlisted NICAR data analysts to examine the data and rank Illinois universities by crime totals. WMAQ reported "the University of Illinois at Urbana-Champaign reported 390 arrests for alcohol-related infractions and 93 for drug offenses on or near its campus in 2002, ranking UIUC at the top among schools analyzed." The report also offers reasons for the large number of assaults on or near Northwestern University's Evanston campus.

Read the story on the Web at *www.nbc5.com/education/3279237/detail.html.* To purchase the campus crime database from IRE and NICAR or to learn more about it, see *www.ire.org/datalibrary/databases/campus.*

AIR SAFETY

Data shows maintenance problems with planes

By Ted Mellnik, *The Charlotte Observer*

Soon after a heavily loaded U.S. Airways Express Flight 5481 crashed in Charlotte shortly after takeoff, airline maintenance appeared as a likely cause.

A mechanic at a small West Virginia repair station had recently adjusted the plane's elevator control cables, a job he had never done on a Beech 1900D. His instructor was also his inspector. He and other mechanics worked nights; the manager usually worked days. And the FAA officer in charge of monitoring the site had visited only once since mechanics began working there.

The Charlotte Observer set out to find how the Jan. 8, 2003, crash that killed all 21 aboard fit with industry trends. A four-part series Dec. 7-10 reported that airlines are spending less to maintain their planes. Mechanics are checking them less often. Federal oversight is stretched thin. And maintenance is increasingly a factor in fatal crashes.

The *Observer* found that since 1994, maintenance problems have contributed to 42 percent of fatal airline accidents in the U.S., up from 16 percent the previous decade. Faulty maintenance contributed to three of the past five fatal airline crashes in the U.S., and likely played a role in a fourth, now under investigation.

Airlines have invested millions to fix other serious problems such as pilot error, and overall, crashes are declining. But faulty maintenance, an equally preventable problem, has never received the attention it deserves, experts told the newspaper.

Sources of data for the series included:

- **National Transportation Safety Board accidents.** This relational database, available for download in Microsoft Access format or for online query from the NTSB, includes facts about crashes, the planes involved, events and contributing factors. At the time of our use, the database held details on more than 50,000 events; it's updated monthly. It was the primary source for statements about accident trends.

We started with a list of U.S. fatal air carrier accidents since 1994 in which the aircraft was substantially damaged or destroyed. An intranet application built using Active Server Pages/VBScript and Microsoft SQL Server allowed reporters to browse the list, filter it with a Structured Query Language "Where" statement, follow a link to investigative reports, as well as add and edit crash entries. The additions were for new crashes.

The air accident data is referred to as ADMSPUB data and documentation can be downloaded at *www.ntsb.gov/ntsb/query.asp.* The documentation is key to understanding and using the relationships among accidents, aircraft, people, events and contributing factors. Staff in the NTSB Public Inquiries branch answered almost all data questions.

- **PTRS inspections**. The FAA's Program Tracking and Reporting Subsystem manages information about inspections of contract repair stations and nonmajor airlines, and a portion of inspections of the nation's major carriers. It was obtained through a Freedom of Information Act request. The data was provided in yearly text data files, sometimes covering more than a million inspections. While the data structure was documented, the challenge was to learn how the inspection data was created and used so we could formulate queries. We focused on more than 450,000 closed, completed airworthiness and avionics inspections from 2000 to mid-2003. An inspection was considered to have an unfavorable result if the inspector reported that it led

continued on page 17

IRE AWARDS

Newspaper winners bank on CAR

By Andrea Lorenz, *IRE and NICAR*

The May-June issue of *Uplink* highlighted the television, special category and other media winners and finalists of IRE's annual investigative reporting contest that used computer-assisted reporting. Here we feature the daily and weekly newspaper winners and finalists that used CAR.

See the May-June issue of *The IRE Journal* for a full list of winners.

You can order copies of these stories and related conference tipsheets from the IRE Resource Center. Contact the center at 573-882-3364 or rescntr@ire.org and provide the number.

Some of the journalists who worked on these stories also recently spoke on panels at the IRE Annual Conference in Atlanta. You can order recordings of these panels from Sound Images Inc. Contact Sound Images at 888-649-1118 or service@soundimages.net.

Largest newspapers or wire service

(More than 500,000 circulation)

Finalists

- "Pharmaceutical Roulette," *The Washington Post*, Mary Pat Flaherty and Gilbert M. Gaul

What started as a story about Medicare drug prices turned into a widespread investigation of problems in the drug distribution system. The *Post* reporters uncovered an extensive network of players including doctors with sketchy histories prescribing the drugs online, felonious middlemen obtaining drugs fraudulently on the cheap, and Fortune 500 companies buying questionable drugs at bargain prices from illegal wholesalers.

The reporters built databases of Internet pharmacy sales from a variety of sources and obtained physician prescribing data.

The stories also won Best of Show in the Maryland-Delaware-D.C. Press Association Awards and first place in Investigative Reporting, Division A: Dailies over 75,000. (See Story No. 20870. To hear Flaherty speak on the Pharmaceuticals: Prescriptions for investigations panel order recording number IRE04-043)

- "Stealth Merger: Drug Companies and Government Medical Research," *Los Angeles Times*, David Willman

Willman investigated drug companies paying consulting fees to National Institute of Health scientists – a practice authorized by the head of the research institute. More digging revealed the consulting deals were kept private, as employees were not required to divulge payments from other sources.

The *Times* obtained data (13,784 pages from the NIH alone) for this series from financial-disclosure statements and other information about public employees. The information was entered by hand to create two databases.

The series also won the Worth Bingham Prize for investigative journalism. (See Story No. 20867)

- "Betrayal in the Ranks," *The Denver Post*, Amy Herdy and Miles Moffeit

This series uncovered military mismanagement of sexual assault and domestic violence cases. After battling the military, The *Post* obtained a database of domestic abuse and sexual assault cases from the U.S. Army's Criminal Investigation Command. Along with another database compiled from U.S. Air Force sexual assault records, the *Post* was able to profile more than 30 cases of abuse.

The series won a Denver Society of Professional Journalists award. (See Story Nos. 20697 and 20869)

Large Newspapers

(250,000-500,000 circulation)

Certificate

- "Crumbling Schools," *The Miami Herald*, Debbie Cenziper and Jason Grotto

In this series of articles about the Miami school district's botched construction efforts, *The Miami Herald* found overpriced building projects, delayed jobs, safety hazards and sketchy business deals. The reporters used a database of construction projects, the district's accounting ledger, county and city safety data, as well as paper records to build their own databases. Microsoft FoxPro, Access, Excel, ESRI ArcView and SAS helped them analyze construction costs and delays, contracts, safety violations and population statistics. (See Story No. 20155. To hear Grotto and Cenziper speak on Exposing school system failures panel order recording number IRE04-068)

Finalists

- "U.S. Olympians had failed drug tests," *The Orange County Register*, Scott M. Reid, William Heisel and Tony Saavedra

The reporters found that more than 100 American athletes were allowed to compete in the Olympics despite failed drug tests. Less than 40 percent of the U.S. Anti-Doping Agency's tests were given without notice though it is recommended that at least 70 percent be a surprise.

Reporters sorted through 10,000 paper records from 9,700 drug tests of 4,600 athletes and entered the information into Excel and Access. They matched dates of drug tests with corresponding sporting events (found through LexisNexis, fan club sites and phone calls) to find whether the athletes attended these events.

continued on page 21

MAPPING IT OUT

The latest uses of mapping in news reporting.

Housing prices on rise

By Matthew Waite
St. Petersburg Times

Imagine buying a house for $60,000. Over three years you fix a bathroom, paint a few walls and put in some shrubs. And then you decide to sell it. Maybe you've heard that other homes on the block were selling well, so what the heck. And, on a whim, you make more in profit than you spent on the house in the first place.

Luck? An isolated case? Hardly, a *St. Petersburg Times* analysis of home sales found. The Tampa Bay housing market is in the midst of a transformation that has affected many bigger cities: the price of a home is skyrocketing, changing neighborhoods house by house and, in our case, greatly eroding the area's reputation for cheap housing. In 1998, the median price of a house around Tampa Bay was the same as Wichita, Kan. Now, it's the same as Dallas, a larger city with more wealth.

To get at this change on a neighborhood level, we largely relied on geographic information systems (GIS), in this case, ESRI's ArcView 3.2 and ArcGIS 8.3. We obtained databases of home sales from the property appraiser's offices in our five-county circulation area, ending up with more than 260,000 home sale records. We also got parcel map files from each county's appraiser. In most cases, the appraisers – elected officials – were happy to provide the data or maps free of charge. Others had standard charges, and just who had them seemed hit or miss. One county gave us maps for free, but charged us for data. Another county gave us data for free, but charged us for maps. The most expensive was $85, from one of our rural counties.

With the sales data and parcel data, we then turned to the problem of analyzing the data at the neighborhood level. Some cities have GIS departments, and gladly turned over GIS boundary maps of neighborhood associations. But our area includes large, heavily populated unincorporated areas, with no defined neighborhood associations, other than developer-created subdivisions that were too small for meaningful analysis. So we created our own, using ArcGIS. We started with a Census Designated Place map – which provides reasonable boundaries for cities and areas – and went from there. With help from editors, reporters and maps, we drew more than 100 neighborhoods onto our main map. In the end, we had more than 300 neighborhoods, towns, subdivisions and unincorporated areas in our analysis and maps.

That allowed us to dig much deeper than we could have with available reports from the Realtors' associations. We could compare cities, towns and neighborhoods from one to another. We could look at how one hot neighborhood spills over into the ones around it. We could get very, very local for our readers. The GIS portion of this, while key to the whole project, wasn't particularly difficult. I used ArcGIS to perform spatial joins, which assigned our neighborhood names to parcels based upon where the parcel was located. Then, with the names joined to the parcels, I could import the parcel table into Microsoft Access and join the parcel table to the sales table. Then, with a neighborhood name given to a sale, I could group the sales by the neighborhood name and then calculate median sale prices by years using SPSS, a statistics program. I then cleaned up the median output SPSS created in Excel and calculated percent changes. By the end, I used almost every software tool in my bag.

Since the *Times* is heavily zoned, the idea from the start was to give each zoned edition a story specific to its area. In the end, we published about 30 stories over two days, 25 of them on one Sunday. For anyone wanting to attempt this kind of story, some advice:

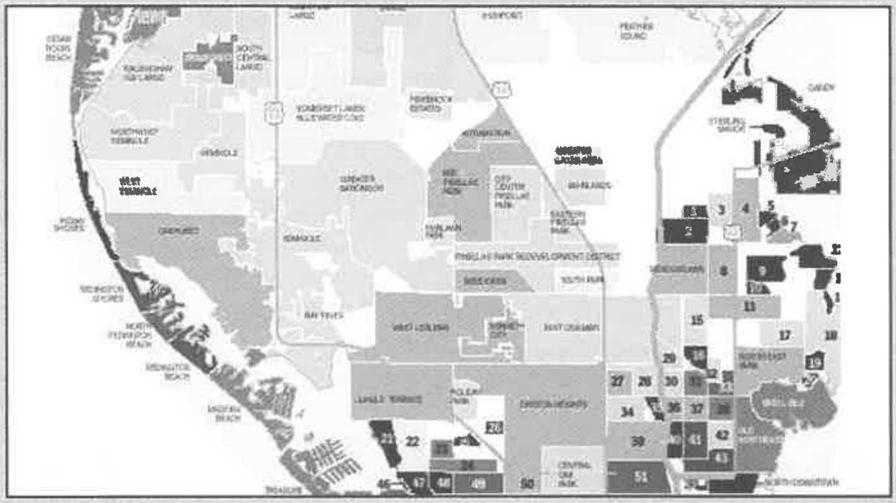
Brace yourself (part 1): In the beginning, this didn't sound like an enormous undertaking. It is. I promise you. The folder where this project resides on my hard drive comes in just under 8 gigabytes, which does not lend itself to easy management.

Brace yourself (part 2): Be prepared for what you're going to get from the property appraiser offices. Some are using Access and ArcView, making everything easy. Some use arcane data management software and unheard-of mapping applications. I was able to get maps and data from them, but it makes a great case for having both Arc 8 and Arc 3 on a machine. Arc 8 relies on a map layer having a projection assigned to it. Arc 3 does not. Data from one of our counties came from a computer-aided drawing system. The parcel file wasn't a collection of polygons – it was a giant collection of hundreds of thousands of lines that seemed to connect, but didn't. I could see what the parcel map was supposed to look like, but it was worthless to me as an analysis tool. The data, however, contained an X/Y coordinate that corresponded to the center of the parcel it was related to. To assign a neighborhood name to a parcel, I did a spatial join. The problem is, none of my layers – from the points to the parcels to the custom-made neighborhood file – had a projection attached to them. Arc 8 won't do a join without a projection, but Arc 3 will.

Check everything twice: Each appraiser's office stored data differently, and we could get only the barest details about a sale from all five counties. And we were after a very specific slice of the market: single family residential sales of houses (no vacant lots, no condos, no townhouses) that were qualified – an arm's length transaction between a buyer and a seller who didn't know each other and weren't giving the other guy a break that would unduly influence the price.

Some officials understood this perfectly, and knew exactly what we wanted. Others didn't get it, and gave us bad data. To check our data, we took more than 1,000 parcel ID numbers, some from each county, and looked them up, checking the sales history of each parcel for sales we had and sales we didn't have but should have. In one county, the appraiser's office gave us three dirty data sets before getting it right.

Write everything down: The last thing I did should have been the first. As one of the last things I did for the project, I wrote out, in excruciating detail, exactly what I did from the start. Much of it was cobbled together from handwritten scribbling on printouts and scratch paper. It was therapeutic almost to write it all out. In the future, I plan to start by writing out what I am going to do, then keeping an electronic journal on what I've done since then. It will come in very handy when editors begin asking questions about why this was done and that was done when it's been four months since you did them. I'm designing an intranet page whose whole purpose will be to track major data projects like this, from data acquisition to analysis to deadlines for stories.



The reaction to this story has been great – my editor called it the "feel-good investigative project of the year." People who own homes are awed by what they might be able to get for their home. I've received a couple of e-mail messages from house hunters thanking me for showing them that they weren't alone. People were buying the Sunday paper at our front counter days after the story ran. The folks who sell them told me they didn't have many left on Wednesday after it ran.

And, our editors have said, the story resonated with people more than anything we've done in recent memory. My thought is it resonated because we wrote about something people care deeply about: home. And that made all the work well worth it.

The complete package, including interactive features, can be found at: *www.sptimes.com/2004/webspecials04/homeprices.*

Contact Matthew Waite by e-mail at waite@sptimes.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: REAL ESTATE

Finding faults in new homes

By Katy Miller, *Orlando Sentinel*

One of the most expensive purchases a consumer can make is his or her home. For buyers of new homes, one benefit is not having to worry about repairs. However, an examination of house construction quality in the booming central Florida building market by the *Orlando Sentinel* and WESH-Orlando, found thousands of problems with the new homes.

Our series, "Building Homes: Building Problems," uncovered a serious decline in craftsmanship in central Florida's new homes. Among the major problems we found were a myriad of heating, ventilation and air conditioning issues, wall and foundation cracks and improperly installed windows. In addition, we found everything from major construction failures to simple poor workmanship.

Don Tracy, a senior projects reporter, asked me to help him build a Microsoft Access database. We had worked together on an Access database in the past, and I knew he had a working knowledge of the program. I didn't realize the extent of project we would be undertaking.

Tracy was doing a quantitative study of new home quality in central Florida. As part of the project, Ron Resch, a local building inspector, was selected to train engineering students at the University of Central Florida's home constructability lab to inspect homes built in 2001. My responsibility was to build the database for the inspections information and analyze the results.

The newspaper acquired a list of homes built in 2001 from local property appraisers. Clerks working for the newspaper and television station called homeowners and asked them whether they would be willing to participate in our study. Those names and phone numbers were submitted to the inspectors to set up appointments.

With the assistance of Resch, Tracy created a paper copy of an inspection report for the students. It had more than 800 categories of information. That should have been my first tip that Access was the wrong tool for this project because Access tables are limited in the number of fields. But at the time it seemed like the best choice, and it was a database the UCF students were familiar with.

I had two challenges in designing the database. First, I needed to make a form that somewhat resembled the inspection report used by the students for data entry. To achieve this I created multiple tables and linked them together using subforms.

I created tables that corresponded to the distinct sections of the inspection report. For example, there was a bathroom subform and a kitchen subform. Each table contained check boxes that indicated whether the student inspectors spotted a certain problem. There was a total of 17 subforms linked to the master form.

While the UCF students inspected the houses, Tracy obtained residential inspection data from the six surrounding counties for the past three fiscal years. We also acquired inspections conducted by the city of Orlando. The goal was to determine the pass-fail rate of the official inspections by county.

The cost for the data ranged from free to under $100, varying by county. The city of Orlando provided a Microsoft Access database, but it was read-only. We exported the tables into Excel and created a new database. Polk County provided text files. However, the data was in print format and looked like screen shots of a report. I used Monarch to parse the data into fields and export it as a delimited text file.

Osceola County at first refused to provide anything other than a printout; the press officer claimed that the database was not available in electronic format. After we spoke to someone in the technical department, we learned this was untrue and obtained the data in Excel. The remaining counties provided Excel spreadsheets that we easily imported into Access.

Unfortunately, the data sets had a couple of problems. Although we requested the past three years of data, the counties did not provide the exact same time periods. However, I felt there were enough records to show an inspection pass-fail rate.

The second problem with the data was that every county had provided some duplicate records. I used a select distinct query in Access to create a table with the unique values.

```
SELECT DISTINCT
Inspections.SITEADDRS,
Inspections.JOBDESCR,
Inspections.CONTRTNAME,
Inspections.CONTRTCOMP,
Inspections.INSPECTOR,
Inspections.INSPCODE,
Inspections.CODEDESCR,
Inspections.inspecdate,
Inspections.RESULTCODE INTO
Deduped

FROM Inspections
```

After I compiled the results, I attempted to verify our findings with the building inspection departments of each county. Only one of the counties had calculated pass-fail rates, and the results matched ours to the decimal point.

In Access, I ran queries to determine the number of inspections daily and a failure rate by inspector. I then generated a list of inspections by day for each inspector. This helped find inspectors who appeared to pass a majority of the houses.

We found an inspector who logged more than 90 inspections in one day. Using MapInfo geographic information system, I mapped the daily route for an inspector in each county, illustrating how little time was spent on each inspection because of travel time.

After UCF started entering data I needed to transfer it between UCF and the *Sentinel* to monitor the inspection results. Our technology department set up an FTP site and UCF uploaded the newest version of the database, with accompanying photos and videos, every Monday.

With the assistance of Resch, Tracy classified the problems found into three categories of severity: worth noting, concerns and priority problems. At this point I realized the restrictions of working in Access. I needed to group the problems into these categories using data from multiple tables.

The first hurdle was to convert the checkboxes on the forms into a data value. Even after converting the data to a value, I was unable to run queries on the data. I solved this problem by converting the values into new tables, and used the new tables when running a query. To avoid re-creating my efforts, the UCF students downloaded a copy of the database I had created with the queries and added data to that database.

When the inspections were almost completed, the editor for the project decided to put the database on OrlandoSentinel.com, our newspaper's Web site. I knew that this would take an enormous amount of programming, and using Structured Query Language (SQL) was the best choice.

In an effort to be fair to the builders, every problem found was not published in the series. Before converting the data from Access in SQL, I asked Tracy and WESH producer Travis Sherwin which problems should be included in the Web version of the database. They refined the original list of problems down to about 200. (See the database online at *http://extra.orlandosentinel.com/buildingproblems*)

After having a pared down list of problems, I sat down with a programmer from the *Orlando Sentinel*'s technology department assigned to build the Web database. Together we assigned each problem a location in the house (i.e., bathroom), a severity ranking, a system (i.e., HVAC), and a system element (i.e., ductwork) for classification purposes.

After the inspections were completed, we exported the tables from Access into the Microsoft SQL Server database. There were more than 120 builders, more than 400 homes, and more than 200 potential problems in each home.

On the inspection subforms, I created a comments box for problems not listed on the inspection sheet. To include this data in our database, I sorted through more than 800 separate comments to be classified. In many cases I had to create a new field for problems that had not been anticipated when the form was created. For example, there were 37 homes with mold on the HVAC unit.

I was put in charge of all the data analysis and statistics for the series. I signed off on every number and kept a spreadsheet listing every stat I checked and the date so I could easily refer back to what I had done when. This was vital because the numbers were flying everywhere, and there were several instances in which preliminary numbers made their way into the final versions of stories.

One of the main difficulties I encountered was the reporter and editor changing the severity of a problem several times. Unlike the data gathered from the inspection reports, the severity rating was the most subjective part of the series. For example, a problem originally classified as a "priority problem" was changed to a "concern" the afternoon before a segment was scheduled to air on WESH. At 3 p.m. I was scrambling to help Sherwin with the updated numbers for the segment running at 6 p.m.

I also worked closely with the graphics department, checking off on every chart or graphic. This also paid off. A chart for the pass-failure rate of inspections was mistakenly labeled New Home Inspections, when it was residential inspections of homes of all ages, a category that includes renovations of older homes.

Thanks to this diligence, there were not any corrections or errors in the data I was responsible for. There were a couple of corrections because data was entered incorrectly into the database by students, despite our efforts to compare written reports to the electronic database.

The response to the series has been very positive. The Web page devoted to the database generated an incredible number of hits and the message boards are full of homeowners reporting similar problems. Tracy had to set up a separate voice mail account to handle all the phone calls in response to the series. A state senator has said he will push for tougher standards on new home construction, and the local Home Builders Association has started to expand a homeowner advocacy program.

Contact Katy Miller by e-mail at
kmiller@orlandosentinel.com.

readme.txt

Interested in seeing how others are using computer-assisted reporting to cover news about real estate? Check out this and other recent stories that appeared on IRE's Extra! Extra! (*www.ire.org/extraextra*).

Robert Gebeloff and Maura McDermott of *The* (Newark, N.J.) *Star Ledger* examined tax assessment data for three northern New Jersey counties and found that residential property tax bills are off as much as $1,000 because of tax rolls that towns haven't fully updated in many years.

Credits

continued from page 1

Indiana's credit is a line-item reduction in each homeowner's tax bill. The state subsidizes the tax cut by writing checks to local governments that depend on property taxes.

The *Star* calculated the cost of the credit and deduction errors uncovered in Marion County at more than $6 million. No one in state government could say how widespread the problem was, but, after checking the newspaper's findings, budget and tax officials said they were concerned it could be occurring elsewhere in Indiana.

The news story ran March 28 and highlighted the 10 county taxpayers who had received the most credits. James E. Chalfant, who received unwarranted credits worth $3,512 on 19 of his residential properties, topped the list. He received related homeowner deductions worth another $1,000 or so.

The idea for the story emerged while I was cleaning a larger database of nearly 400,000 county property records. I wanted to evaluate the effects of a wrenching court-ordered overhaul of Indiana's system of valuing property for tax purposes. The Indiana Supreme Court had declared the old system unconstitutional and ruled that taxes on homes must be more closely tied to market prices.

The General Assembly, fearing hordes of angry homeowners at the ballot box, had passed a tax-relief package in mid-2002 that raised Indiana's sales tax to 6 percent from 5 percent to cover tax revenue losses that resulted from an increase in the homestead credit and a related deduction.

But in early 2003 legislators found a 17-year-old mistake in how the homestead credit was calculated. They chose to correct the error, which had favored homeowners. The action diminished the value of the credit and generated hundreds of millions of dollars for a state budget that was running in the red. Months later, in the summer of 2003, thousands of Indianapolis homeowners learned their property tax bills would be double or more as a result of the court-ordered reassessment.

I had originally sought the tax data so I could recalculate Marion County's tax bills and figure out how much of the relief taxpayers had been promised had been taken back. (The *Star* had previously reported they had paid an extra $30.2 million.)

The county's computer vendor burned a copy of the data onto CD-ROM in delimited text format for me at no cost, at the request of a township assessor. The data had some inconsistent owner names and I spent a lot of time cleaning the data to create uniform names. In addition, some townships had combined a number of parcels into one, making it impossible to compare the tax bills over the years. So I created a field called "combo" that indicated whether the parcel was created by combining other parcels.

In the course of cleaning the tax data in Microsoft Excel, I noticed during sorts on taxpayer names that many had more than the single homestead credit allowed by law.

We already knew about owners getting more than one credit. Earlier in the year, two reporters for the *Star*, Vic Ryckaert and Matthew Tully, had disclosed that Marion County's prosecutor had more than one homestead credit in 2003.

My data analysis showed the extent of the problem in Marion County and put a price tag on it. Inside the newsroom, Neill Borowski, the *Star's* assistant managing editor for news, championed the story. I largely credit him for the Sunday A1 play the story received.

Going into the story, I had assumed the multiple homestead credits were a result of homeowners applying for more than their fair share of tax breaks. (You can read the story and search a database of property records at *www.indystar.com/news/politics*).

But I began to question that assumption after taxpayers with the most homestead credits began calling me back and vehemently denying they had applied for them. I reached six of the top 10 before the stories ran. I reached two others by sending them letters via Federal Express at their tax-bill addresses. One contacted me after the story ran and I never did hear from one. They all told the same story, and upon checking paper land records, I found they were telling the truth. They had not applied for the breaks; they had inherited them.

The errors had turned out to be a result of poor record keeping by the county auditor, the local official responsible for administering this state-funded program. Problem No. 1: State officials were relying on local officials to make sure tax relief went to the right people, but local officials had no reason to care about the data's accuracy, because money to pay the credit was coming from the state's bank account.

County officials told me that it would be difficult, if not impossible, to pull data off the county's old mainframe computer and put it into a format that I could use. Problem No. 2: State and local officials usually have no idea how computers work.

The data was made available at no cost after I persisted and found a township official who asked the contractor to provide it as an ASCII text file on CD-ROM.

The *Star* used a snapshot of property tax data taken Aug. 18, 2003, to determine which taxpayers had received more than one homestead tax credit. The database was narrowed to include just the 209,767 land parcels with these credits. I knew I had the correct parcels because the sum spent on homestead credits in my database matched, to the penny, the amount of the tax relief check the state had to write Marion County last year.

Using Microsoft Access, I analyzed the data to determine which taxpayers with the same first and last names received more than one credit by grouping and counting them. Because the county's data does not include unique identifiers such as Social Security numbers, I eliminated people with common first and last names from the analysis. Admittedly, that required making some judgment calls.

I assumed the most valuable homestead credit for each taxpayer with more than one credit was valid. I also assumed credits on properties near those with the largest credit each taxpayer had received were legitimate, because in Indiana the homestead credit applies to a primary residence and up to an acre of adjacent land. Any additional credits the remaining property owners received were totaled to determine the cost to taxpayers of paying for undeserved credits.

This analysis found 9,315 taxpayers with 20,780 credits; the 11,465 extra credits cost taxpayers an additional $1.11 million.

In addition, a homestead credit comes with a related property tax deduction of up to $35,000 per home.

To determine the effect of giving county taxpayers extra homestead deductions, I assumed that the largest deduction for each property owner with more than one deduction was the valid one.

For each invalid homestead deduction, property value was removed from Marion County's tax rolls, effectively increasing tax rates. I did not attempt to re-compute county tax rates for the county's 60 taxing districts, which would have left room for error and added weeks to the project.

Instead, after talking to property tax experts, I used a proxy measure of the likely effect of property owners getting more than the one homestead deduction they were entitled to by law.

To calculate this effect, the property value these undeserved deductions had eliminated from tax rolls was totaled ($171 million), divided by 100, and multiplied by $2.9421, the weighted average tax rate per $100 for homes in Marion County last year. This proxy measure indicated the extra deductions had shifted the burden of paying $5.03 million of property taxes, mostly onto businesses and homeowners with just one homestead credit.

I queried Access for a list, grouped by name, of each taxpayer with more than one deduction. In addition, I asked the program to create fields totaling the values of these deductions for each taxpayer and listing the highest-valued deduction for each taxpayer. Then I exported the data to Excel, where I subtracted the highest deduction from the deduction total to get the invalid amount of deductions. I added the 6,146 undeserved deductions to arrive at the $171 million total.

I found the unwarranted homeowner tax credits and deductions cost state and local taxpayers least $6.14 million.

Several months before the story ran, I provided a list of property owners with duplicate credits to the county auditor so she could review them. She told me her own study, performed after the prosecutor's homestead credits had drawn scrutiny, found 9,000 more duplicate credits than the *Star's* conservative estimate had.

In addition, the auditor's aides double-checked the list of the two dozen people with the most credits and removed more than 210 homestead credits from their properties. This list became the basis for the Top 10 list the paper published.

I also walked through my methods and results with state property tax analysts at the State Budget Agency, Department of Local Government Finance and nonpartisan Legislative Services Agency. They reviewed my methodology and did not dispute my findings. After the story ran, the budget agency assigned an analyst to check some other counties. During a cigarette break outside the Statehouse he told me that he'd found a public school district in suburban Hamilton County with six homestead credits.

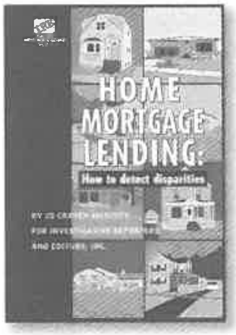
When I contacted property owners with the most homestead credits, one local real estate agent getting too many credits asked, "How much are you going to cost me?"

The answer: At least $3,000 a year after the auditor canceled his credits.

Not everyone was so upset. Chance L. Felling, No. 8 on the *Star's* list, wrote a check to pay the $1,691 owed for 11 invalid credits.

"Regardless of who made the mistake," Felling said, "if I owe money, then I should pay."

Contact Kevin Corcoran by e-mail at Kevin.Corcoran@indystar.com.



Voting

continued from page 1

a local resident. Investigators were trying to determine whether Teaneck's at-large voting system should be replaced with a ward system that would help promote the election of black candidates to the town council and school board. Teaneck was 28 percent black but none of the members of the town council and only one member of the school board were black. In fact, there hadn't been a black resident on the town council since 2000.

We worked the story on two tracks. Fallon, through his sources within the town, began retracing the steps of the investigators from the Department of Justice. He requested the same records from the town and interviewed the same town officials and residents. Meanwhile, I began to work on the statistical analyses the investigators were likely to conduct.

We contacted Edward Still, an Alabama attorney who specializes in voting rights cases. Still explained what the investigators would be looking at and how we could mimic their analysis. I also asked two university professors to act as expert advisers.

We learned that the investigators would be seeking to answer three key questions. First, was there racially polarized voting? Put simply, do whites vote predominately for white candidates and blacks vote predominately for black candidates? Second, do black candidates routinely run for office and lose? And, third, can a ward system be created that includes at least one ward in which minorities would be dominant?

So, while Fallon did the shoe-leather reporting, I dug into the numbers.

To answer the first question I needed to conduct a regression analysis. Sounds frightening, doesn't it? It is. At least that was my initial thought. But just before I ran away screaming, I pulled out notes from a weeklong IRE and NICAR statistics boot camp that I had attended in Chapel Hill, N.C., where I learned statistical analysis. I said to myself, "You can do this."

So what is regression analysis? It is simply an attempt to determine the predictive value of the relationship that exists between two variables. In this case, our variables were the demographics of the voting age population and voting results for each of the 21 voting districts within Teaneck from several recent elections.

My first step was to collect the data. Determining the racial makeup for each of the voting districts became a long and tedious process. First, I grabbed block level race data from the 2000 Census and then mapped that information out using ESRI's ArcView 3.2a. That sounds simple enough, but here's where the real work started. Teaneck did not maintain an electronic version of the voting districts map.

Fortunately, the borders of the voting districts and those of the Census blocks were, in all but one case, identical. So I placed a street map of the town over my map of the Census blocks and followed the streets to determine which blocks were within which voting district. I then created mapping shapefiles for each of the 21 voting districts and also exported the underlying data to a Microsoft Excel spreadsheet.

Now that I had the demographic data and the voting results data I was ready to conduct a regression analysis. (NOTE: It would have been better to use the racial breakdown of registered voters rather than the racial breakdowns in the Census voting age population data but that information was unavailable.)

I conducted the initial regression analysis in Microsoft Excel after my expert advisers, Bernard Grofman at the University of California-Irvine and David Epstein at Columbia University, recommended the program. I built a spreadsheet containing both the racial breakdown and voting results for each of the 21 voting districts. Then, with a few clicks, Excel produced a scatter plot graph and several statistical measures of the data to indicate what - if any - relationship existed between the percent of the population that was black and the performance of black candidates within each district.

Excel's regression routine does not report statistics that tell whether the any relationship is meaningful. I replicated the analysis in SPSS, a statistical program that has the tools, and also ran the results by Grofman and Epstein.

I analyzed five recent Teaneck elections in which a black candidate ran and found at least some level of racially polarized voting in four of those. I then asked Epstein to double-check my analysis, and he came to the same conclusions. However, Epstein was quick to point out that the presence of racially polarized voting is not uncommon.

Complicating matters further was the success of Gordon Johnson, a black state assemblyman, who had great success within Teaneck during primary and general elections.

So, while it appeared that the town did suffer from racially polarized voting that didn't necessarily mean the at-large voting system needed to be replaced. But it did indicate that the Justice Department investigation was not a waste of time, and that the town had some difficulties.

The second and third questions were much easier to answer. We found that since 1990, seven of the 22 black candidates (32 percent) for the township council and school board won while 56 of 90 white candidates (62 percent) ran successfully. So black candidates did have a history of running and losing. By using ArcView and Excel I was able to determine that the population in the northeast section of the town was 65 percent black so the creation of a ward where minority group members would dominate would not be difficult.

At the end of the analysis we were able to paint a statistical portrait of Teaneck for our readers that helped explain what had happened in recent elections and also gave them insight into what was probably happening inside the federal investigation.

One important piece of advice I might offer is that you don't need a federal investigation to do a story like this. Many elected bodies in towns and cities all across the country look nothing like the communities they represent. Does that mean those towns or cities are violating federal law? No, but it is worth the time and effort of reporters to find out why and, in the process, perhaps begin a constructive discussion within those communities about how best to ensure elected officials represent all segments of a community.

Contact Benjamin Lesser by e-mail at lesser@northjersey.com.

readme.txt

Interested in seeing how others are using computer-assisted reporting to cover news about the elections? Check out these recent stories that appeared on IRE's Extra! Extra! (*www.ire.org/extraextra*).

Chris Davis and Matthew Doig of the *Sarasota Herald Tribune* examined Florida voter data and found that a quirk in the state's drive to clear convicted felons from the voter rolls would spare Hispanics, who tend to support Republicans, in greater numbers than other races.

Mark Greenblatt of WBBH-Fort Myers, Fla., built and analyzed a database of the voting records of 34 local politicians and found that a handful voted less than half the time since 1990.

CAR TOOL

Vigilant program tracks Web changes

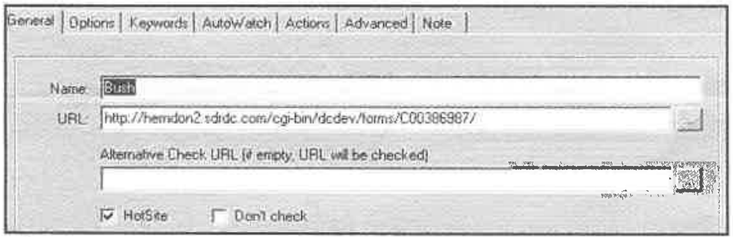
By Sarah Cohen, *The Washington Post*

As any journalist who has to deal with presidential campaign finance raw data can tell you, the 20th of each month this year poses a competitive nightmare. Candidates must report their monthly contributions and expenditures to the Federal Election Commission by midnight of that day. But some report a day or so early, and others miss the deadline by a few minutes.

This means that journalists might have a day to analyze and report the results or may still be waiting, at deadline, for the data. Knowing when the data is available is crucial and getting the filings as soon as they are posted on the Web becomes a competitive obsession.

Enter Website Watcher. The utility is available for purchase and download at *www.aignes.com*. A single-use commercial license costs $99. There are several similar tools on the market. I chose Website Watcher because it was the first one I found that had a few features I needed.

Basic use of the program is simple: You type a Web address for the program to check and give it a name. (See figure 1)



General | Options | Keywords | AutoWatch | Actions | Advanced | Note

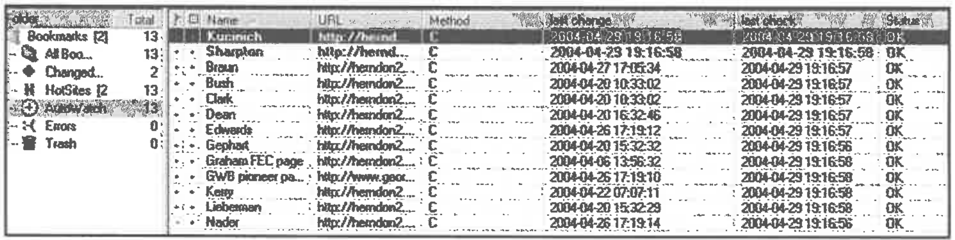
Name: Bush

URL: http://herndon2.sdrdc.com/cgi-bin/dcdev/forms/C00386987/

Alternative Check URL (if empty, URL will be checked)

☑ HotSite ☐ Don't check

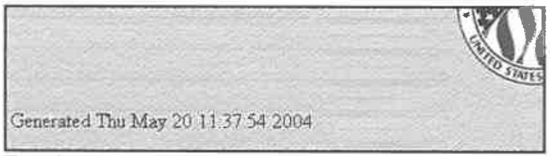
Figure 1



Folder	Total
Bookmarks [2]	13
All Boo...	13
Changed...	2
HotSites [2	13
AutoWatch	13
Errors	0
Trash	0

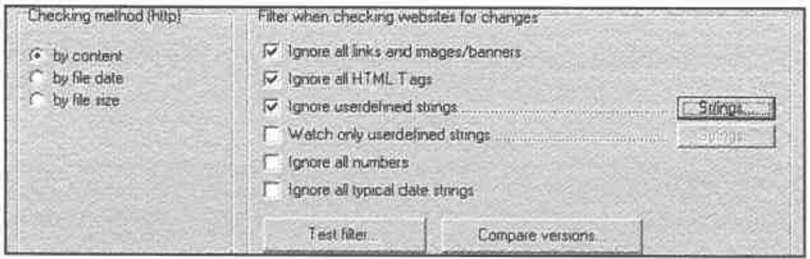
Name	URL	Method	last change	last check	Status
Kucinich	http://hernd...	C	2004-04-29 19:16:58	2004-04-29 19:16:58	OK
Sharpton	http://hernd...	C	2004-04-29 19:16:58	2004-04-29 19:16:58	OK
Braun	http://herndon2...	C	2004-04-27 17:05:34	2004-04-29 19:16:57	OK
Bush	http://herndon2...	C	2004-04-20 10:33:02	2004-04-29 19:16:57	OK
Clark	http://herndon2...	C	2004-04-20 10:33:02	2004-04-29 19:16:57	OK
Dean	http://herndon2...	C	2004-04-20 16:32:46	2004-04-29 19:16:57	OK
Edwards	http://herndon2...	C	2004-04-26 17:19:12	2004-04-29 19:16:57	OK
Gephart	http://herndon2...	C	2004-04-20 15:32:32	2004-04-29 19:16:56	OK
Graham FEC page	http://herndon2...	C	2004-04-06 13:56:32	2004-04-29 19:16:58	OK
GWB pioneer pa...	http://www.geor...	C	2004-04-26 17:19:10	2004-04-29 19:16:58	OK
Kerry	http://herndon2...	C	2004-04-22 07:07:11	2004-04-29 19:16:58	OK
Lieberman	http://herndon2...	C	2004-04-20 15:32:29	2004-04-29 19:16:58	OK
Nader	http://herndon2...	C	2004-04-26 17:19:14	2004-04-29 19:16:56	OK

Figure 2



Generated Thu May 20 11:37:54 2004

Figure 3



Checking method (http)

○ by content
○ by file date
○ by file size

Filter when checking websites for changes

☑ Ignore all links and images/banners
☑ Ignore all HTML Tags
☑ Ignore userdefined strings
☐ Watch only userdefined strings
☐ Ignore all numbers
☐ Ignore all typical date strings

Figure 4

You can keep adding pages. For those you want to check routinely, add them to a HotSite. For those you just want to keep track of once in a while, you can just keep it manually. (See figure 2)

After you enter the sites, set them to "Autowatch," choose how often they should be checked and you can walk away. You can set Website Watcher to alert you of changes by playing a sound, opening the page in a browser, e-mailing a copy of the page or even running a program.

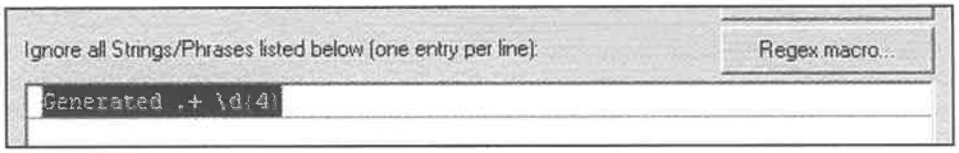
There are a few features that make this program especially good for the kind of work we journalists do.

The most important feature is the ability to ignore changes that don't matter. For example, the FEC electronic filing system re-generates the page each time it's checked. This means that technically, the page always changes. (See figure 3)

That little phrase, "Generated Thu May 20 11:37:54 2004", in the page would create a new version each time the page was checked. But Website Watcher can ignore all changes in dates, ignore a phrase or pattern of characters that you define, or to do the opposite: Only count it as a change when a certain pattern appears. Here is how I've told the program to ignore this little "Generated" phrase. (See figure 4)

I've told Website Watcher to ignore links and images, since I know they won't include a new filing. I've also told it to ignore internal coding. But I didn't want it to ignore "all typical date strings", because that might be included in a new filing. Instead, I gave it a pattern to look for by checking "Ignore userdefined strings", then typing in this pattern. (See figure 5)

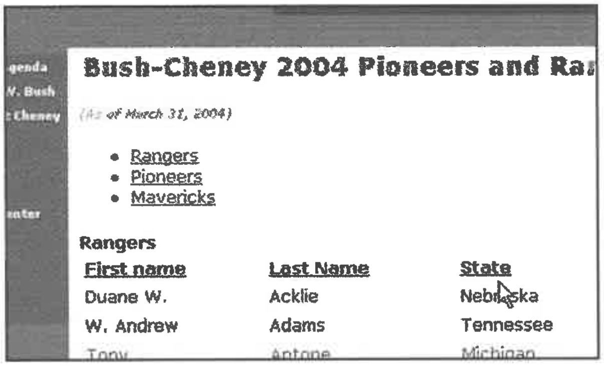
This expression means look for the word "Generated", then anything, then four digits in a row. If this is the only thing that changed in the page, then Website Watcher won't consider the page changed.


Ignore all Strings/Phrases listed below (one entry per line):

```
Generated .+ \d{4}
```

Figure 5

Another feature I like is that the program archives the most recent copy of the page. When there is a change, Website Watcher alerts you and then saves a new copy with the changes highlighted in yellow. In this case the phrase "(As of March 31, 2004)" has been highlighted and formatted as italic in my viewer. (See figure 6)



Bush-Cheney 2004 Pioneers and Ra

(As of March 31, 2004)

- Rangers
- Pioneers
- Mavericks

Rangers

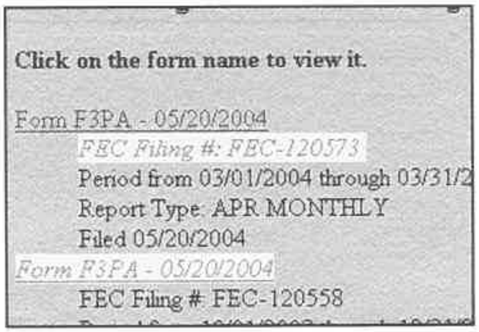
First name	Last Name	State
Duane W.	Acklie	Nebraska
W. Andrew	Adams	Tennessee
Tony	Antone	Michigan

Figure 6

For the Bush-Cheney fund-raisers, called Pioneers, it highlights new last names. It can also show you what has been deleted from a page at the end.

There are many more advanced features I haven't explored. Among them are options to e-mail copies of the changed page and run programs, such as a Perl script, when a change is identified.

But I have to go. Website Watcher just beeped to let me know that the Bush campaign filed an amendment to its April report. (See figure 7)


Click on the form name to view it.

Form F3PA - 05/20/2004

FEC Filing #: FEC-120573

Period from 03/01/2004 through 03/31/2

Report Type: APR MONTHLY

Filed 05/20/2004

Form F3PA - 05/20/2004

FEC Filing #: FEC-120558

Figure 7

Contact Sarah Cohen by e-mail at cohensh@washpost.com.

Planes

continued from page 4

to an enforcement investigation or required follow-up action.

Valuable information about PTRS can be gleaned from watchdog reports issued by agencies such as the General Accounting Office. For example, in 1998 the GAO reported that PTRS records understate the incidence of problems and violations. The *Observer* obtained an unpublished memo that described in detail the GAO's methods and confirmed ours.

Key questions about PTRS were answered by interviewing inspectors and others in the FAA who use the system and its data.

- **EIS.** The Enforcement Information System data is the FAA's record of its actions against people and organizations in aviation. Current data used to be available online. Its distribution was limited and some key Web pages removed following the 9/11 terrorist attacks. An extract of the data became available again in mid-2003. The IRE and NICAR Database Library makes the EIS available to journalists. For more information see *www.nicar.org/data/faae.*

- **TranStats.** This online service by the Bureau of Transportation Statistics provides data on airline spending, including maintenance spending, departures and passengers. It's available at *www.transtats.bts.gov.*

- **Other data.** Several other databases were used in background. These included FAA Accidents and Incidents Database, System Difficulty Reports and lists of contract repair stations. These and other databases are listed on the Aviation Data Systems (AFS-620) Web page at *http://afs600.faa.gov/AFS620.htm.*

You can read the stories on the Web at *www.charlotte.com/mld/charlotte/news/special_packages/planes*, and an archive of the newspaper pages in Adobe Portable Document File format is at *http://161.188.204.190/charlotte/air.*

Contact Ted Mellnik by e-mail at tmellnik@charlotteobserver.com.

readme.txt

For more details about this investigation, see the July-August *IRE Journal.*

The IRE Beat Book "Covering Aviation Safety: An Investigator's Guide" provides guidance for journalists interested in probing airplane safety. For more information and ordering instructions, see *www.ire.org/store/books/aviationbook.htm.*

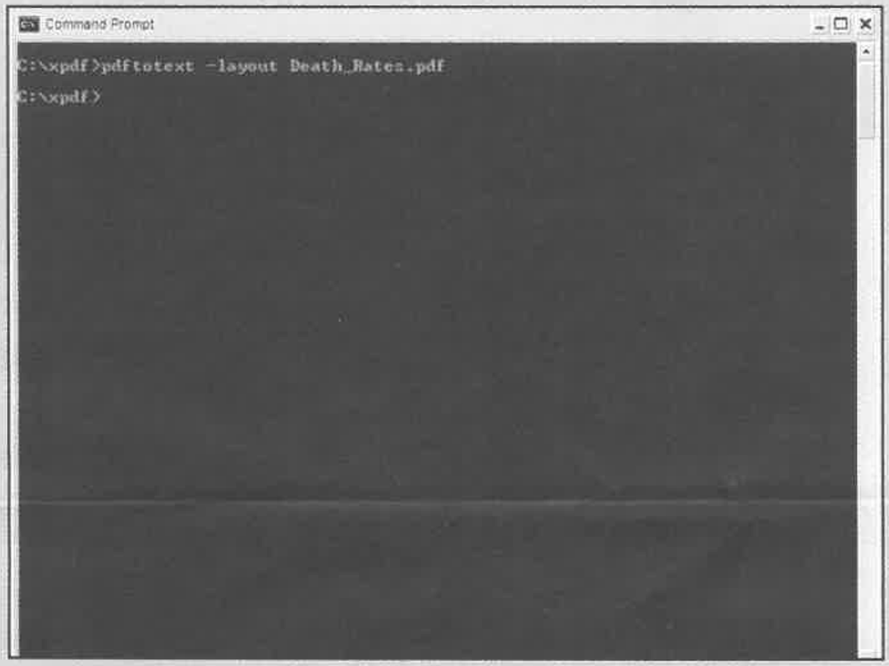
Tech tip...

PDF to text using XPDF

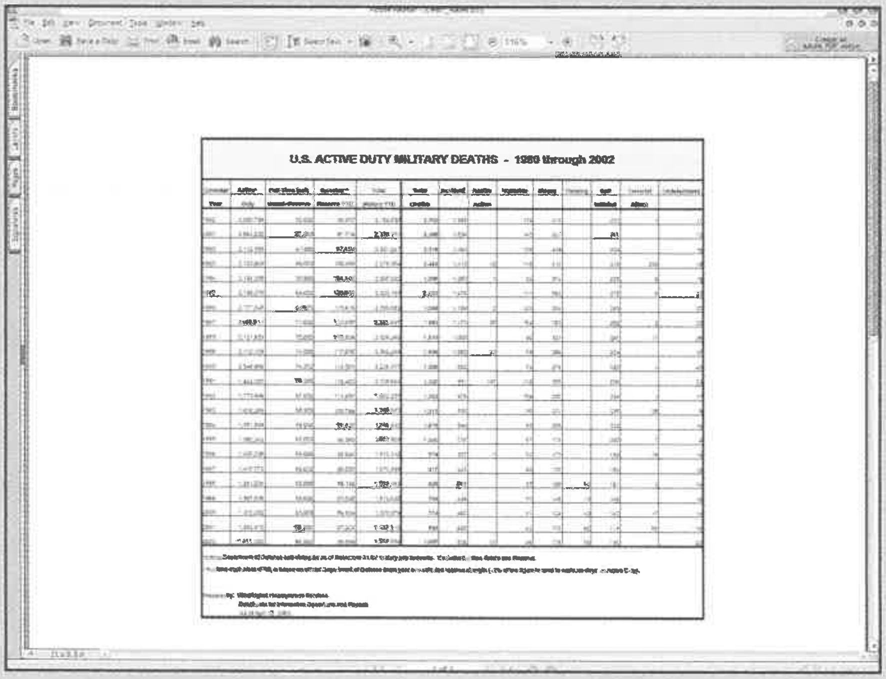
By Derek Willis, *Center for Public Integrity*

Here's a question that should have a familiar ring for CAR veterans: How do I get text out of a PDF file?

If your experience has been anything like mine, the answer is likely to be "painfully, if at all."



```
C:\xpdf>pdftotext -layout Death_Rates.pdf
C:\xpdf>
```



U.S. ACTIVE DUTY MILITARY DEATHS - 1980 through 2002

The existence of Adobe Portable Document Files has been a boon for reporters because governments and agencies everywhere have been able to make documents broadly available over the Internet. It's hard not to love the idea of a file that looks the same no matter what kind of computer you have.

But if you've ever tried getting tables out of a PDF document – and we all have – the results usually aren't worth the effort. Until now. (See "Slicing and dicing those pesky PDFs" in the July-August 2003 *Uplink*.)

If you're ready to leave behind the annoyance of copying and pasting from PDF files, a free command line utility called Xpdf will save you time and aggravation. It will, in most circumstances, enable you to go from PDF to Microsoft Excel in a matter of seconds, rather than minutes or hours. Did I mention that it's free?

Although Xpdf comes with many versions of the Unix or Linux operating systems, Microsoft Windows users will need to download and install the small application in order to run it.

Xpdf can be downloaded from *www.foolabs.com/xpdf/download.html*, and it comes in packages for Windows and Linux/Unix. Look for the precompiled binaries for your operating system. A DOS version is available, but unless you run DOS as your primary operating system, the Windows version should be your choice. The Windows version does not have all the features of the Linux/Unix version, but it has all you'll need for converting PDFs to text tables.

First, save the download file (a zip file) to your computer and then unzip it into a directory (c:\xpdf works fine). Then, and this is the technical bit, when you have a PDF you want to get tabular text from, simply place the PDF in that directory, open a command prompt, then navigate to c:\xpdf and type:

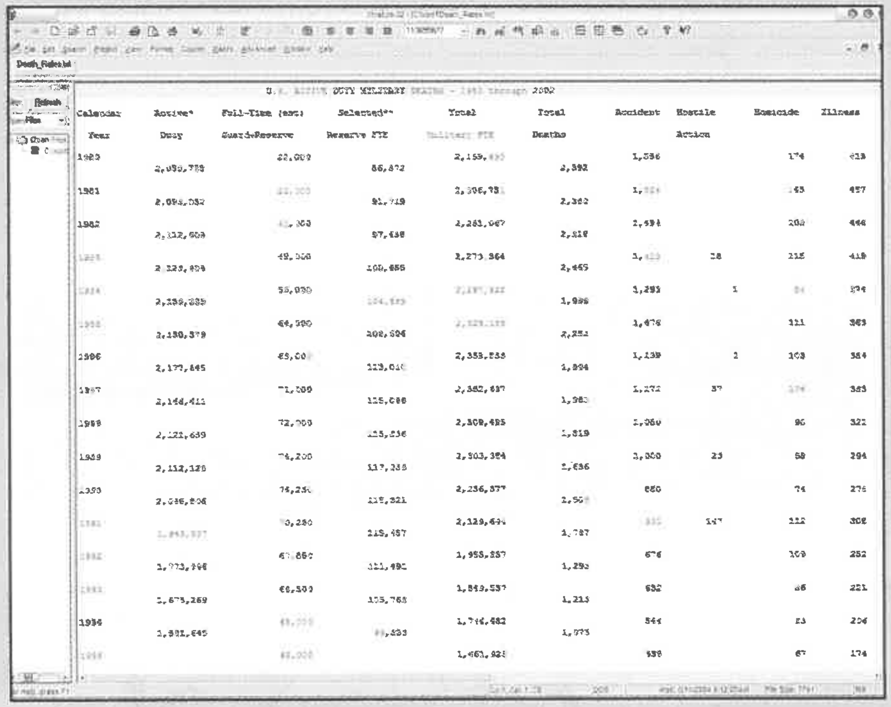
```
pdftotext -layout pdfname.pdf
```

Depending on the size of the PDF file, your output text file (with the same name as the original) will be in the same directory in a matter of seconds.

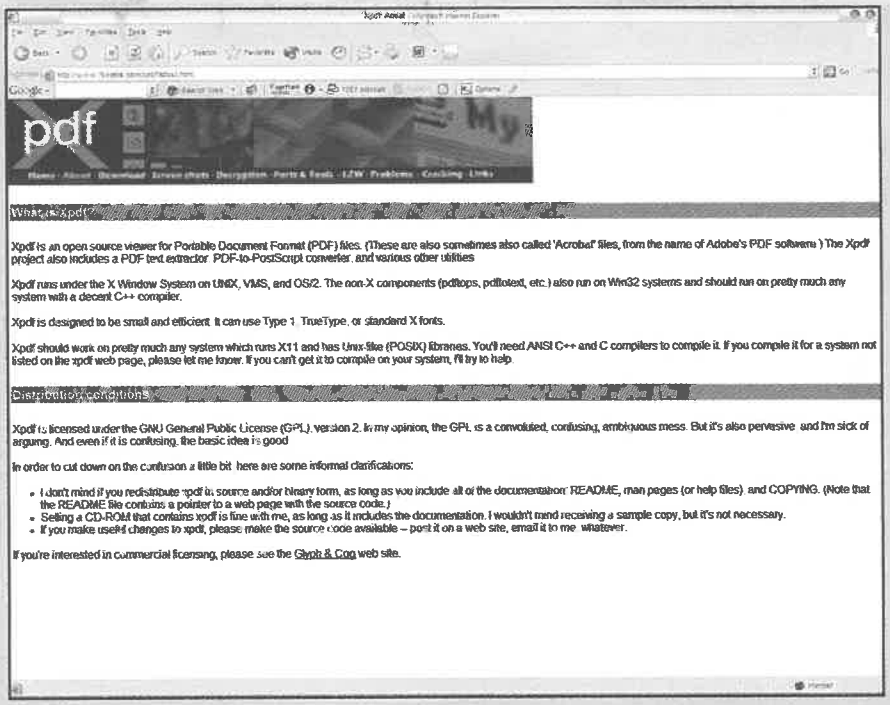
Let's go through the command line syntax. First, the command "pdftotext" is required for this process, and "pdf2text" won't work. The "-layout" tag tells Xpdf that you want to preserve the layout that the PDF file uses, which keeps the text in those nice, clean tables. And you need to have the full name of the file (I recommend a single-word name, even though Windows supports filenames with spaces). That's it. You would be forgiven for thinking that it can't be that easy, but it really is.

The resulting text file will be the entire text of the PDF, meaning that you may have to wade through pages of text in order to get to your tables. The preservation of the PDF's layout means that if a page contained two tables side-by-side, that's the way they will look in the text file, too.

Xpdf doesn't work in all instances; specifically, it won't convert PDFs that have been locked by their creators. Don't bother asking the author of Xpdf, either, as he has posted a message on his Web site indicating that he will not add that ability.



```
                        U.S. ACTIVE DUTY MILITARY DEATHS - 1980 through 2002
Calendar   Active*     Full-Time (est)   Selected**     Total           Total      Accident  Hostile   Homicide  Illness
  Year      Duty       Guard+Reserve     Reserve FTE    Military FTE    Deaths               Action
1980                      22,000                        2,159,630                  1,556                174       419
           2,050,758                      86,872                        2,392
1981                      [illegible]                   2,306,731                  1,524                145       457
           2,093,032                      91,719                        2,380
1982                      [illegible]                   2,251,067                  2,494                208       446
           2,112,609                      97,458                        2,319
[illegible]               49,000                        2,273,364                  1,[illegible]  18    115       419
           2,123,494                     100,455                        2,465
[illegible]               55,000                        [illegible]                1,293      1      [illegible] 374
           2,138,335                     [illegible]                    1,999
[illegible]               64,000                        [illegible]                1,476                111       363
           2,150,379                     102,604                        2,252
1986                      65,000                        2,353,538                  1,139      2         103       364
           2,177,845                     123,016                        1,894
1987                      71,000                        2,382,637                  1,272      37       [illegible] 383
           2,168,411                     125,088                        1,983
1988                      72,000                        2,309,495                  1,060                96        322
           2,121,659                     115,236                        1,819
1989                      74,200                        2,303,384                  1,000      23        58        294
           2,112,128                     117,056                        1,636
1990                      74,250                        2,236,377                  880                  74        276
           2,046,806                     115,321                        1,507
[illegible]               70,280                        2,129,644                  [illegible] 147      122       308
           [illegible]                   115,457                        1,787
[illegible]               67,850                        1,955,357                  676                  109       252
           1,773,948                     111,492                        1,293
[illegible]               66,500                        1,849,537                  632                  86        221
           1,675,269                     105,768                        1,213
1994                      [illegible]                   1,746,482                  544                  83        206
           1,581,645                     [illegible]                    1,075
[illegible]               [illegible]                   1,661,928                  538                  67        174
```


Xpdf About
Google
pdf
Home · About · Download · Screen shots · Decryption · Ports & Tools · LZW · Problems · Cracking · Links

What is Xpdf?

Xpdf is an open source viewer for Portable Document Format (PDF) files. (These are also sometimes also called 'Acrobat' files, from the name of Adobe's PDF software.) The Xpdf project also includes a PDF text extractor, PDF-to-PostScript converter, and various other utilities

Xpdf runs under the X Window System on UNIX, VMS, and OS/2. The non-X components (pdftops, pdftotext, etc.) also run on Win32 systems and should run on pretty much any system with a decent C++ compiler.

Xpdf is designed to be small and efficient. It can use Type 1, TrueType, or standard X fonts.

Xpdf should work on pretty much any system which runs X11 and has Unix-like (POSIX) libraries. You'll need ANSI C++ and C compilers to compile it. If you compile it for a system not listed on the xpdf web page, please let me know. If you can't get it to compile on your system, I'll try to help.

Distribution conditions

Xpdf is licensed under the GNU General Public License (GPL), version 2. In my opinion, the GPL is a convoluted, confusing, ambiguous mess. But it's also pervasive, and I'm sick of arguing. And even if it is confusing, the basic idea is good

In order to cut down on the confusion a little bit, here are some informal clarifications:

- I don't mind if you redistribute xpdf in source and/or binary form, as long as you include all of the documentation: README, man pages (or help files), and COPYING. (Note that the README file contains a pointer to a web page with the source code.)
- Selling a CD-ROM that contains xpdf is fine with me, as long as it includes the documentation. I wouldn't mind receiving a sample copy, but it's not necessary.
- If you make useful changes to xpdf, please make the source code available – post it on a web site, email it to me, whatever.

If you're interested in commercial licensing, please see the Glyph & Cog web site.

For dealing with most government documents, Xpdf can be a huge time-saver and allow you to spend more time actually analyzing data rather than trying to free it from the confines of the PDF.

Contact Derek Willis by e-mail at derek@thescoop.org.

Fire

continued from page 1

Crowd evacuation simulation software is well accepted in the building design and fire safety industries. In most jurisdictions, such software can be used to demonstrate that a building complies with codes governing emergency evacuations. An owner whose building might otherwise fail to meet fire codes that prescribe, among other things, the number and size of exits, can still pass inspection if an evacuation simulation shows building occupants would reach safety in time.

The Station nightclub fire was the ideal situation for a newspaper to use a computer model. The newspaper, through interviews with more than 200 survivors, had excellent data regarding where people were when the fire broke out and how they got out of the building. Because of those interviews and because the first minute of the fire was videotaped from inside the building by a television camera operator, the newspaper had a good idea of what actually happened, providing an excellent test of the computer model against the real world.

A model would produce worthwhile results, both in terms of visualizing what happened and in exploring "what ifs", such as what if the exits had been configured differently.
The *Journal* researched several evacuation simulation software packages online before selecting two – Simulex and STEPS – for evaluation.

Though the software is normally expensive, the developers of each package, after learning what the newspaper wanted to do, offered free licenses and technical support. Both packages were similar in their features and calculations, though they had subsfantial differences.

STEPS produced excellent three-dimensional color animations of the evacuation, allowed the user to have somewhat greater control over the movements of building occupants and easily rendered the animation in .AVI video format.

Simulex was easier to set up, the movements of individual people in the model appeared more lifelike, and – most important for the *Journal* – the overhead, two-dimensional animation of the evacuation was better suited for conversion to a newspaper graphic, though it had to be done by hand by a graphic artist.

Screen-grab shareware Screen Movie Studio was used to generate .AVI animations that were posted on the newspaper's Web site because Simulex could only create video in a proprietary format.

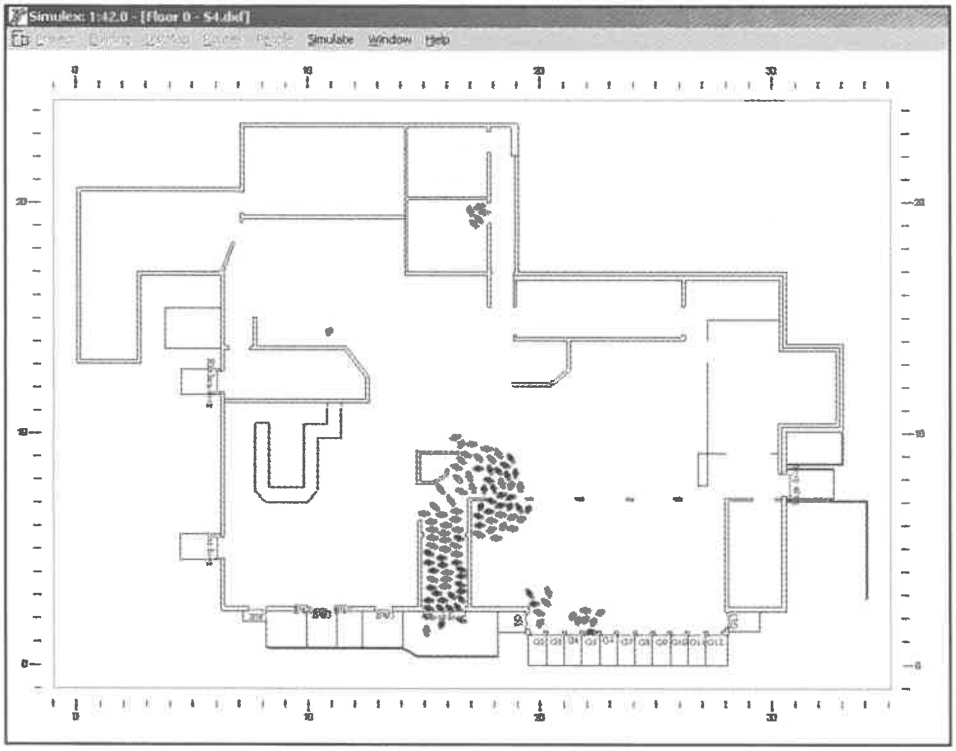
Simulex, as does STEPS, starts with a building floor plan, which is created in a Computer-Aided Design (CAD) program and imported into Simulex. The *Journal* had used public records requests to obtain excellent floor plans of the nightclub on paper that were recreated using CAD software by the *Journal* employee who oversees remodeling of the newspaper's building.

Next we defined the exits by entering their widths and graphically placing them on the floor plan. Then we entered information about the people inside the nightclub, to account for gender, body sizes, walking speeds and alcohol consumption.

We placed the people on the floor plan, along with the exit they would use and a defined delay factor. The delay factor was critical in The Station fire. Because rock band Great White started its show with fireworks and many in the crowd had been drinking, few were alarmed to see flames shooting up the stage walls. The television video showed it took more than 30 seconds for the crowd as a whole to start moving toward the exits.

After we placed the people, we ran the simulation, which records an animation of the evacuation, and provides statistics that are updated 10 times a second.

The *Journal* employed a few tricks in getting Simulex to mimic what happened the night of the fire. Because the program is designed to calculate how long it takes to evacuate a building, it does not account for the "crowd collapse" that led to people piling up in the front door.



Winners

continued from page 5

The article won second place in the Associated Press Sports Editors investigative reporting category. (See Story No. 20338)

- "Damaged Lives: Lead's Toxic Toll," *Detroit Free Press*, Emilia Askari, Tina Lam, Megan Christensen, Marsha Low, Hugh McDiarmid, Jr., Dan Shine, Shawn Windsor and Wendy Wendland-Bowyer

A team of reporters investigated the contamination from lead smelters and leaded gasoline. They found government agencies neglecting to clean up lead in neighborhoods and test children for lead poisoning.

They used Access to analyze a Michigan Department of Community Health database for children tested for lead poisoning and mapping software to show the results of their soil tests. (See Story No. 20369)

- "Shell Game," *The Times-Picayune*, Aaron Kuriloff and Jeffrey Meitrodt.

This series investigated $2 billion in court judgments against the state in favor of oyster farmers. The reporters found the awards to be either unjustified or excessive. Using tax and harvest records of oyster farms, court documents and sales data, they built tables in Excel to show results of a project initiated by the state to rebuild the coast. They found oyster beds were doing better – although oyster farmers claimed they were destroyed – than before the project.

The series won the Louisiana Press Association investigative award, and the Associated Press/Louisiana-Mississippi story of the year. (See Story No. 20608)

- "Silent Alarm," *The Kansas City Star,* Karen Dillon and Mike McGraw

This series provided an in-depth look at the government's handling of the Hepatitis C virus. They used Microsoft Visual FoxPro and Excel to work with some of the data. Spreadsheets containing information about people alive today with the virus contracted through blood transfusions showed the estimated mortality of these patients. Other spreadsheets showed the number of Hepatitis C cases since the 1980s that could have been prevented by the government with a test. They also created a database of medical journal articles about the virus. "Silent Alarm" won first place in the Kansas Press Association award for investigative reporting. (See Story No. 20921)

Medium newspapers

(100,000-250,000)

Medal

- "Buried Secrets, Brutal Truths," *The Toledo Blade*, Michael D. Sallah, Mitch Weiss and Joe Mahr

The investigation of atrocities committed by the Army platoon Tiger Force in the Vietnam War brought out details that were buried by the government decades ago. Going through old documents and interviewing former soldiers and Vietnamese villagers gave essential details of the case.

continued on page 22

Crowd collapse can occur for many reasons. In The Station fire, people at the back of the crowd were the first to feel the effects of heat and smoke, and rushed forward to get nearer the exit. Meanwhile, people at the front of the crowd slowed down as they reached the outer doors, a normal reaction once safety is reached in an evacuation.

This difference in speed led to people in front being knocked over, and, as the crowd continued to surge forward, more people tripped and fell on top of them. This pileup effectively closed the front door of the club.

But the Simulex software was not designed to model doors that close during an evacuation. The *Journal* tricked Simulex into doing that by programming several large extra people who tried to push their way into the door at the moment the crowd collapsed. Having people trying to go both ways through the same door effectively jammed it.

In another instance, we had to compensate for a window broken after the start of the fire. We did that by adding some other people into the model and setting their delayed reaction to the time that the window broke. These people were then placed right in the window, making that area impassable until the delay had elapsed.

In the case of both tricks, the extra people were electronically removed from the graphics published in the paper and the video published on the Web.

The model showed that, in addition to the pile of people at the front door, fire victims were trapped deeper inside the club, at the beginning of a hallway leading to the front door.

The model also showed that if the club had a rear exit if, the hallway at the front exit had not been built or if fewer people were inside the club, which was over capacity, virtually everyone would have had a chance to escape.

The stories developed from the computer models were part of The *Journal's* package that was recognized as a finalist for the Pulitzer Prize for public service and in the IRE Awards.

Crowd evacuation models such as Simulex and STEPS could be useful for a variety of computer-assisted reporting projects, such as evaluating existing or proposed buildings or even, with a few tweaks, examining outdoor venues.

Contact Paul Edward Parker by e-mail at pparker@projo.com.

Uplink

Uplink Info

A newsletter of the National Institute for Computer-Assisted Reporting

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
Christina Caron
Megan Clarke
Jaimi Dowdell
Holly Hacker
Andrea Lorenz
Tim Ragones

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

Director of Publications
Len Bruzzese
len@ire.org

Advertising Coordinator
Pia Christensen
pia@ire.org

Subscription Administrator
John Green
jgreen@ire.org

Subscriptions
IRE members $40, nonmembers $60

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

Postmaster: Please send address changes to IRE-NICAR.

Ira Chinoy
Philip Merrill College of Journalism
1117 Journalism BLDG
University of Maryland
College Park MD 20742