Uplink

July • August 2006
Volume 17, Number 4

Published bimonthly by the National Institute for Computer-Assisted Reporting

www.nicar.org

LATIN AMERICAN CAR

A bumpy ride

By Alejandra Fernández Morera
Missouri School of Journalism

Cars travel on highways, but in Latin America, CAR travels on unpaved roads. While computer-assisted reporting has gained popularity in Mexico and Brazil, it is still *terra incognita* in Central America, and few can point out more than one journalist who excels in the genre's many techniques in South American countries such as Chile, Venezuela and Peru.

"Few practice and master [CAR]," said Pedro Armendares, executive director of the Mexican Investigative Reporting Center, or CPI, a partner organization of Investigative Reporters and Editors, Inc., based at the Missouri School of Journalism. "There are countries where it isn't practiced at all."

In the countries where CAR is practiced, the success has been clear. In one case, a Mexican reporter discovered that taxpayers' money was used to pay for $400 towels for the president's residence. The reporter used data from a govern-

continued on page 21

SPOTLIGHT: ON THE JOB

Program for disabled misses the mark

By Tom Detzel, *The Oregonian*

Our stories on nonprofits that employ severely disabled workers sparked an FBI investigation and calls for reform of a little-known federal contracting program that spent $2.25 billion last year, much of it on military goods and services.

For these stories, simple databases produced profound results.

The Javits-Wagner-O'Day program sets aside government business for

continued on page 16

SPOTLIGHT:

For more about the workplace see:

- Navigating OSHA data for stories, p. 4
- Data and resources for reporting, p. 7
- Recent stories from IRE, p. 7
- The July-August *IRE Journal*

AIR SAFETY

Poor conditions plague medical helicopters

By Alan Levin, *USA Today*

There's no substitute for precision on a big project.

USA Today set out more than a year ago to learn more about why air ambulance flights were crashing at an alarming rate. A quick review of the clips revealed several fatal crashes that had killed patients. The numbers of crashes were climbing.

Our guts told us we had a good story. But you can't attribute major findings to gut instinct.

That started several months of searching for hard data that eventually led us to create our own database of crashes from scratch. Along the way, we occasionally cursed the sometimes archaic ways of aviation regulation and recordkeeping, which can make data retrieval very awkward. But, in the end, we had a roadmap laying out the worst crashes and the major safety shortfalls. Our data was more complete than anything the government had compiled.

continued on page 18

Bits & Bytes

National Conference

More than 800 journalists, educators and students attended the 30th Annual IRE Conference, June 15-18 in Fort Worth, Texas.

The conference offered more than 130 panels and workshops on covering beats, writing, editing, producing and hands-on classes in data analysis.

About 300 people showed up for Thursday's optional CAR day, which included panels on issues as far-ranging as introducing newsroom intranets and diagnosing medical and care data.

The conference also featured two rooms exclusively geared towards CAR: a hands-on room teaching the basics of Microsoft Excel spreadsheets and Access database manager, and a demo room where journalists showed off some of the newest and most advanced technology with presentations on CrimeStat, satellite imagery, Web scraping, social network analysis and more.

Federal Contracts Update

The IRE and NICAR Database Library recently updated its Federal Contracts database, making the latest additions available to journalists faster.

The database contains data about contracts between about 70 federal agencies and thousands of contractors. Those contractors include large corporations, local businesses, universities, nonprofits, and local, state and foreign governments. The database lists – among other things – the agency and office, the amount of money, date, contractor, purpose

continued on page 4

INSIDE NICAR

Tune in to CAR tips

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

Check out this incredible resource for helping you learn more about computer-assisted reporting from the top practitioners in the business.

For the first time, IRE and NICAR offer direct access to digital audio recordings from panels and speeches at the 2006 IRE Conference in Fort Worth, Texas. In the past, journalists who wanted recordings had to order tapes or CDs from a vendor.

IRE members who attended the 2006 Conference can simply visit the conference Web page at *www.ire.org/training/dallasfortworth06* and follow the links to find .mp3 (MPEG-3) files. Listen to them online or download for playback on a computer or portable digital audio player.

Conference attendees also may order a CD containing audio from all of the panels, paying only for shipping. IRE members who did not attend the conference can purchase the CD for $20 plus shipping. Order online from the conference Web page or contact the Resource Center at 573-882-3364.

So, if you want to learn more about mapping data, social network analysis, using CAR to report on campaign finance – you name it – you can revisit Fort Worth and get an earful of what you missed.

Contact David Herzog by email at dherzog@nicar.org.

About our contributors ...

Mike Casey is a projects reporter for *The Kansas City Star*, where he has investigated nursing homes and problems facing the Pension Benefit Guaranty Corp. At the *Dayton* (Ohio) *Daily News*, he won an IRE Award for reporting on workplace issues.

Tom Detzel, investigations editor at *The* (Portland) *Oregonian* since 2004, has coordinated projects focused on defective housing construction, Portland's police disability system and a deadly lapse in oversight by Oregon's medical establishment. He joined the paper in 1995 and served as political editor and Washington correspondent.

Gerry Lanosga, investigative producer at WTHR-Indianapolis since 1997, speaks and teaches on the subjects of computer-assisted reporting and public access to information. His journalism honors include the IRE Freedom of Information Award. Lanosga spent nearly a decade as a reporter and columnist for the *Indianapolis News* and *The Indianapolis Star*.

Alan Levin covers aviation safety and regulatory issues for *USA Today*. He previously served as database editor and an investigative reporter at *The Hartford Courant* and did database reporting at *The Boston Herald*. Though his job requires that he spend less time on data than before, he tries to maintain his skills doing analysis for fantasy baseball.

Alejandra Fernández Morera is a journalist from Costa Rica. A Fulbright fellow, she received a master's degree from the Missouri School of Journalism in May. During her graduate studies, she specialized in computer-assisted reporting and is currently working at the Center for Public Integrity in Washington, D.C. Since 1997, she has held reporting, producing and editing positions in top television and radio stations in her home country and abroad.

Neil Reisner is an associate professor of journalism at Florida International University in Miami and was training director for IRE and NICAR from January 1996-January 1997. He has worked at the *Daily Business Review* in Miami, *The Miami Herald* and *The Record* in Hackensack, N.J.

MaryJo Sylwester is the database editor for the *The St. Paul* (Minn.) *Pioneer Press* where she works on daily and long-term stories, maintains the newsroom intranet and offers CAR training. She also worked as database editor at *USA Today* and The Center for Public Integrity and is a former IRE and NICAR Database Library administrator.

FIRST VENTURE

Mapping finds holes in siren coverage

By Gerry Lagnosa, *WTHR-Indianapolis*

Taking my first crack at ArcView 9.1 reminded me of the first time my dad let me use a chainsaw – a little bit scary (probably more for him than me) – but a hell of a lot of fun.

I had tinkered around with ESRI's geographic information system (GIS) program for a couple of small maps after attending IRE and NICAR's Mapping Boot Camp last summer, but in March we decided to get serious with mapping by taking a look at tornado sirens in central Indiana.

It's certainly ground that has been tread before, but not in any in-depth way in our viewing area. With tornado season approaching, we figured it would be a high-interest story for our audience that would also give us a good initiation with ArcMap.

From start to finish, the project took just about every bit of three months for reporter Bob Segall, photojournalist Bill Ditton and me. First we made calls to get siren locations in Marion and the eight surrounding counties in the Indianapolis metropolitan area.

Most of the counties either lacked GIS files on the sirens or had files that didn't fit our needs. For example, we already had obtained the data from Marion County's GIS office, but it was quite outdated. So we decided to begin from the beginning, building our own database of sirens from which to create maps.

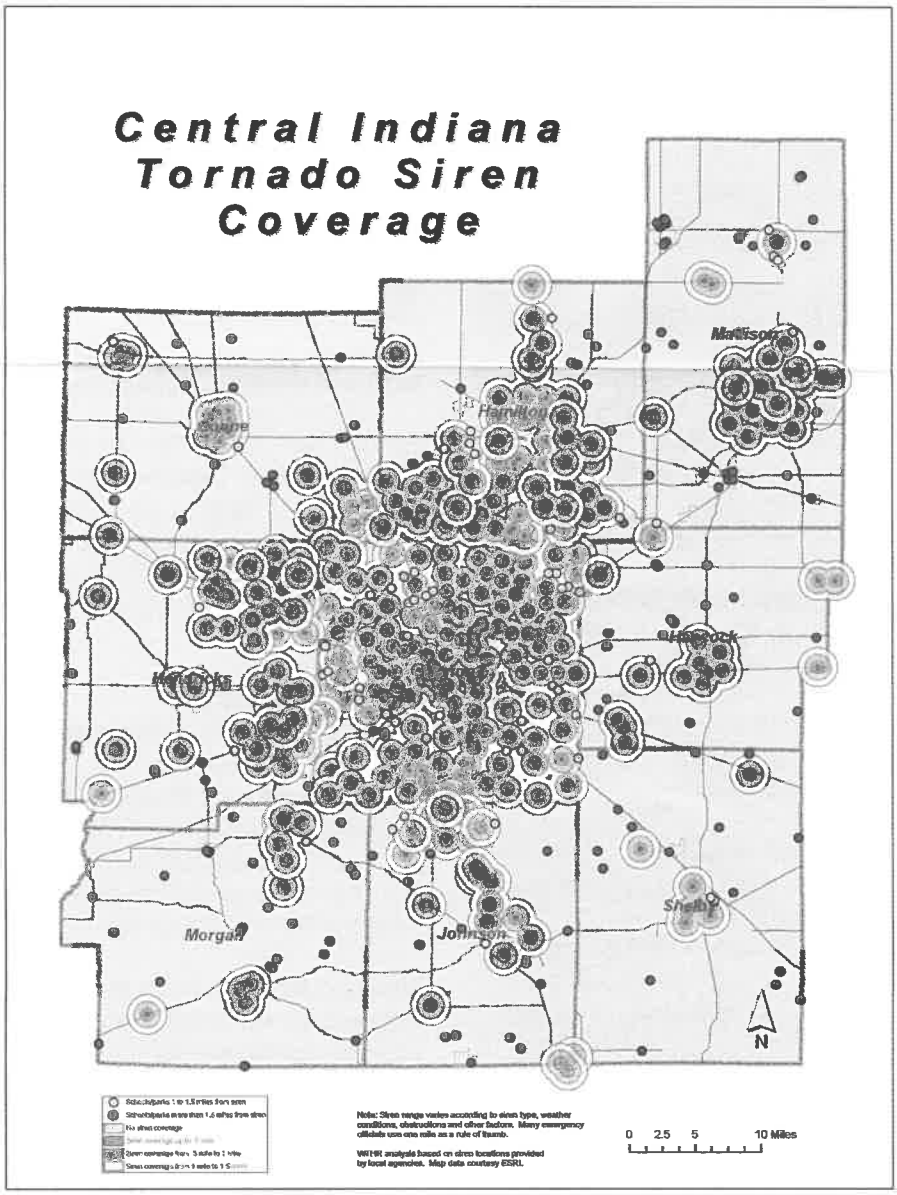
Data gathering

Getting the lists of locations was an adventure in itself. Some counties were extremely well-organized; others not so much. In Hancock County, just east of Indianapolis, it took several phone calls and interviews just to get to the person who reportedly knew where all the sirens were. When we found her at the sheriff's department, she jotted out a list of locations from memory. And, when we went to verify the locations, we found some of them didn't exist. The problems in Hancock, characteristic of the smaller counties in our area, would end up being central to one story in our series.

With the locations in hand, I built a database in Microsoft Access and then pulled the data into ArcView for geocoding to create a point map of the sirens. I was able to match only about half of the 350 sirens. The rest we plotted the hard way.

Bill, Bob and I fanned out with handheld GPS units and gathered coordinates (in UTM, or Universal Transverse Mercator, projection) at each of the remaining siren locations. This proved

continued on page 13



Bits & Bytes

continued from page 2

of the contract and where the work is performed.

Journalists can use the data to examine the local economic impact of a federal project. Combine it with campaign finance information to look into the relationships between the government awarding the contract and the contractor who earns the business.

Previously, the data was updated annually, but for the first time, the Database Library is making new data available as soon as the agency's contractor releases updates.

For more information call Database Library Director Jeff Porter at 573-882-1982 or see *www.ire.org/datalibrary/databases/viewdatabase.php?dbaseindex=17*.

Educator's Center

With a new school year almost here, now is a great time for journalism instructors to look into the IRE Educator's Center. Top educators provide syllabi for computer-assisted reporting, online research and investigative reporting curricula. A materials page includes several reporting and writing exercises as well as links to information that could be turned into exercises.

The page also includes an interactive math test, where students can practice the kind of basic math challenges journalists face – from calculating percent change to averages, medians and means. See *www.ire.org/education*.

SPOTLIGHT: ON THE JOB

Making sense of OSHA data

By Mike Casey, *The Kansas City Star*

How much is a worker's life worth? What are the most deadly workplaces in your town? Are there factories in your town that are exposing workers to toxic chemicals?

The Occupational Safety and Health Administration database has the answers. OSHA is the federal agency responsible for worker safety, and its database details millions of inspections of factories, steel mills and construction sites since 1972.

Don't let the size of the database or its complexity intimidate you. For a recent project about worker safety around Kansas City, *The Star* purchased the data from the IRE and NICAR Database Library. See *www.ire.org/datalibrary/databases* for details about this database and ordering instructions.

To tailor the analysis for our circulation area, I extracted data for a dozen area counties. It's easy to do using codes assigned to each county in each state. I used Microsoft Access database manager for my analysis of nearly 30,000 records.

The big picture

One way to approach the database is to think of it as a news story. The main table is like the lead of a story that gives you a little about the most important aspects of each inspection. The accompanying tables are like the rest of story; they provide more detail and context.

The main table has more than 50 fields that give the name of the company, its address, the reason for the inspection, the number of OSHA violations, penalties the company paid and a number of other items.

Other tables describe the inspection in more detail. If the inspection concerns a fatality, the database tells you the name of the accident victims as well as the accident's causes.

Another table explains the violations in more detail and references the specific regulations that the company violated. Yet another table provides details about hazardous chemicals found in the workplace.

What connects all of these tables is the unique inspection number, which OSHA calls the activity number. It's just a matter of using your database manager to connect the tables using that field.

Job deaths

I was interested in looking at fatalities. In OSHA's codes, a fatality is classified as an "A" in the inspection-type field. I selected all "A" inspection types from the main table and joined that to the others using the inspection number.

Here are some of the things I learned:

- Low fines for workplace deaths or injuries are common even when OSHA cites employers for a serious violation. In 80 such fatal and injury accidents, half of the fines Kansas City area employers paid were $3,000 or less.
- A table detailing the citations showed that the agency sometimes downgraded its most serious willful violations to unclassified and then substantially cut its penalties in these fatal accidents. To find the change, look in the section that details the types of violations. "W" stands for willful and "U" stands for unclassified.
- The table that details fatalities gave the names of workers and I noticed that more workers, with Hispanic names were dying on the job. After

checking death records, I learned that indeed more Hispanics – particularly immigrants – were killed at local worksites.

The data also showed that falls and electrocutions often caused workplace fatalities in the Kansas City area.

Data dirt

Like all databases, OSHA's has strengths and weaknesses. I've found the database information nearly always matched the information in the written reports. Of course you would want to get written reports to verify information about specific incidents in your story.

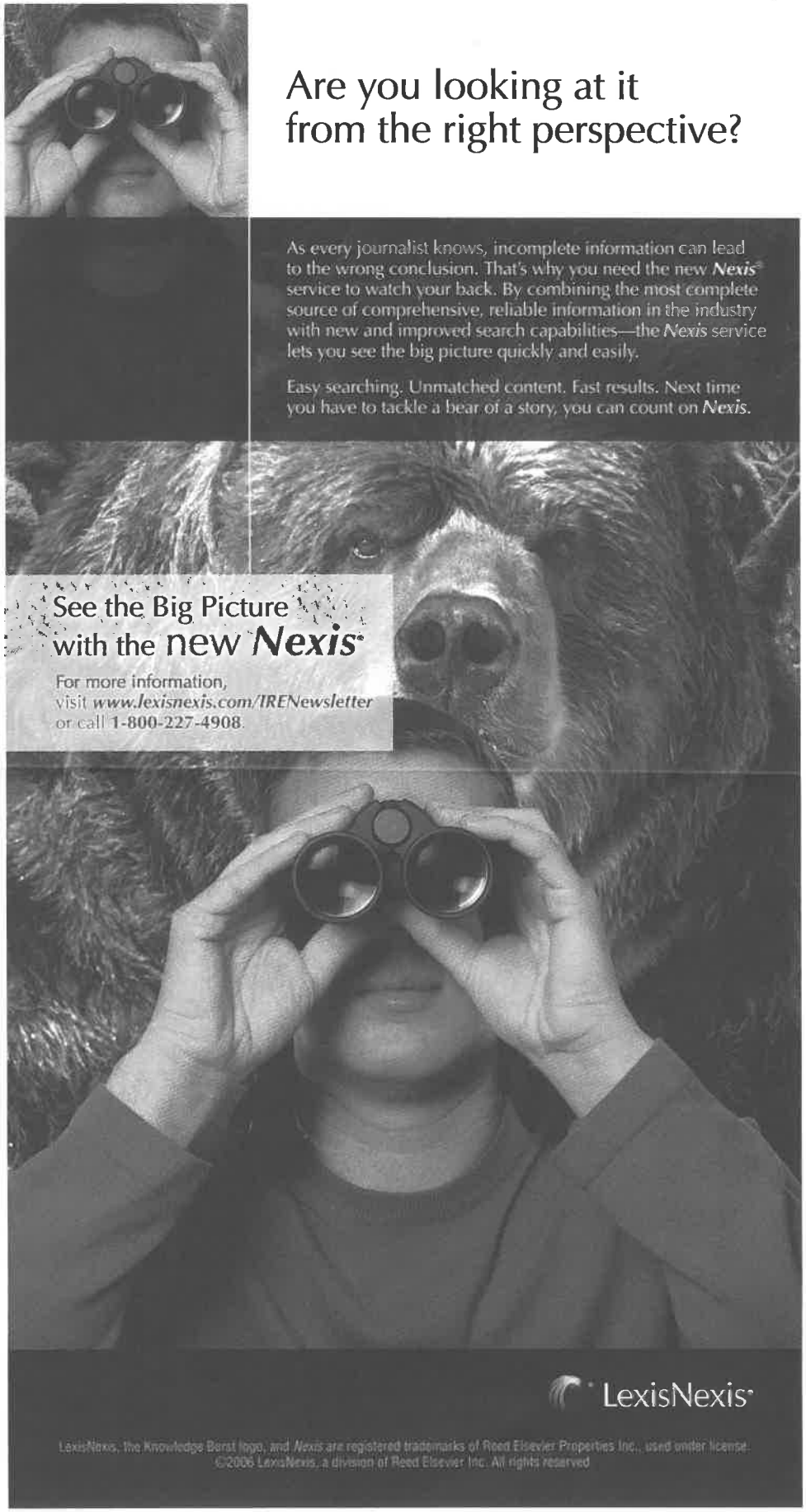
However, some of the OSHA fields contain dirty data. For example, Ford Motor Co. can be called Ford Motor, Ford or Ford Motor Co – no period. That creates a challenge in finding the companies with the highest number of violations.

Here's how to work with that problem: start by selecting and grouping the company names. That will give you an initial idea of the variations. Once you get the top five or 10, check their addresses to see if the name of the facility is connected to the same address. A company could have several locations, and you want to account for that.

Using the address, city and ZIP code fields can help you come up totals for each worksite. For example, Factory A has the most violations and its address is 100 Main St. Using wild cards, you also can expand the search of names and addresses to get all of the records at the address or variations the company name. Do the same thing with the company's name using ZIP codes and city codes as a way to check. Yes, this is time-consuming work, but it helps to make your reporting more accurate.

For construction inspections, the worksite is listed as the location, so a construction company may have a number of different sites listed. That makes coming up with a top violator among construction companies very difficult, if

continued on page 20



Call for Entries

Philip Meyer Journalism Award

Sponsored by

National Institute for Computer-Assisted Reporting

a joint program of IRE and the Missouri School of Journalism

and

John S. and James L.
Knight Foundation

the Knight Chair in Journalism

at the Walter Cronkite School of Journalism and Mass Communication, Arizona State University

The contest recognizes the best journalism done using social science research methods.

All entries are placed in IRE's Resource Center story library so IRE members can learn from others' triumphs and troubles.

POSTMARK DEADLINE: October 31, 2006

EXTRA! EXTRA!

YOUR GUIDE TO THE LATEST INVESTIGATIVE WORK

Check out the stories below for examples of journalists using computer-assisted reporting to report about the workplace. These stories and others can be found at Extra! Extra! on the IRE Web site (*www.ire.org/extraextra*).

Keith Rogers of the *Las Vegas Review-Journal* used Department of Labor statistics to review compensation applications made by sick employees who once worked at six radioactive and toxic materials sites. Rogers found that the Nevada Test Site had a compensation approval rate of 6 percent, the lowest of the sites. Yet only at the Nevada site did employees work in areas that had been detonated by nuclear devices. Nearly 900 cases were denied because former employees couldn't prove beyond a 49 percent likelihood that radioactive materials contributed to their illnesses.

Randy Ludlow of *The Columbus Dispatch* reports on respiratory illnesses contracted by more than a quarter of the employees working in a TRW brake plant in Mt. Vernon, Ohio. He also reviews the lack of Occupational Safety and Health Administration standards to regulate practices that caused these respiratory problems. As he describes in the July-August issue of *The IRE Journal*, Ludlow found illness trends by using a database of TRW employee complaints from the Ohio Bureau of Workers' Compensation. He incorporated that data with OSHA's inspection data and anecdotes gathered from dozens of interviews.

In *Time* magazine, Donald L. Bartlett and James B. Steele review how employees lost their retirement benefits when 401(k) plans tanked and pension plans were pulled out from under them. Company-sponsored pension plans are under-funded by $450 billion in the private sector and are quick to vanish when companies hit financial troubles. While 2004 reports claimed the average 401(k) is worth $61,000, the authors say that the median value was only $18,000.

Resources for workplace reporting

IRE and NICAR offer resources for journalists who are interested in using computer-assisted reporting to cover workplace issues.

The Occupational Safety and Health Administration inspections database, available from the IRE and NICAR database library, contains more than 3 million records of OSHA inspections in all states and U.S. territories from 1972 to February 2006. The OSHA database contains a variety of information on workplace conditions and accidents, including hazardous substance accidents. Journalists can use the database to identify companies that have dangerous working conditions for employees and to examine the safety record of companies involved in breaking news.

The Database Library also offers the Wage and Hour Enforcement database kept by the U.S. Department of Labor. It shows violations and penalties related to labor laws, including the Fair Labor Standards Act, Family and Medical Leave Act, and the Migrant and Seasonal Agricultural Worker Protection Act. It contains records regarding more than 700,000 cases that have been closed between the 1980s and 2004. Journalists can investigate the fairness of compensation paid by employers, identify industries with frequent offenders, and check adherence to child labor laws.

In addition, IRE resources can provide guidance in using the databases to for stories. Chapter 12 of The Investigative Reporter's Handbook has a section specifically about safety and health in workplace.

In the March-April 2002 issue of *Uplink*, Ron Nixon explains pitfalls of OSHA data and offers some measures to overcome them. In the same issue, Stephanie Armour lists workplace data resources, along with some specific story ideas.

Tipsheets contributed by journalists are also good resources to refer to in covering workplace issues.

Barbara Clements and Daniel Wickert of *The* (Tacoma, Wash.) *News Tribune* offer advice about analyzing OSHA data and suggestions for making the data more manageable. They also provide a list on web resources regarding an OSHA project. (Tipsheet No. 2458)

Justin Pritchard of The Associated Press explains where to get information about death in the workplace and how to handle that information. (Tipsheet No. 2062)

Journalists have written a number of stories about the workplace and available from the IRE Resource Center:

David Barstow of *The New York Times* revealed that OSHA declined to seek prosecution in 93 percent of the cases in which it determined an employer's "willful safety violations" caused a death. The report also finds that "OSHA's reluctance to seek prosecution persisted" even when the violations caused multiple deaths, or when the victims were teenagers. (Story No. 20850)

An investigation by *The* (Fredericksburg, Va.) *Free Lance-Star* disclosed that state government agencies are among the most common violators of worker safety standards, but they never get fined. The story identifies the area's top violators in the past 25 years, including the City of Fredericksburg and the County of Stafford. (Story No. 18656)

Contact the Database Library at 573-884-7711 to order workplace data. More information about the databases is available online at *www.ire.org/datalibrary/databases*.

Contact the Resource Center at rescntr@ire.org or 573-882-3364 to order copies of the stories or tipsheets.

CAR TOOL

Wrestling useful data from PDFs

By Neil Reisner, *Florida International University*

Let's start with the obvious: PDFs are devils masquerading in angelic garb.

They emerged in the early '90s as a purportedly benevolent effort by developer Adobe Systems Inc. to create "a universal way to communicate documents across a wide variety of machine configurations, operating systems, and communication networks."

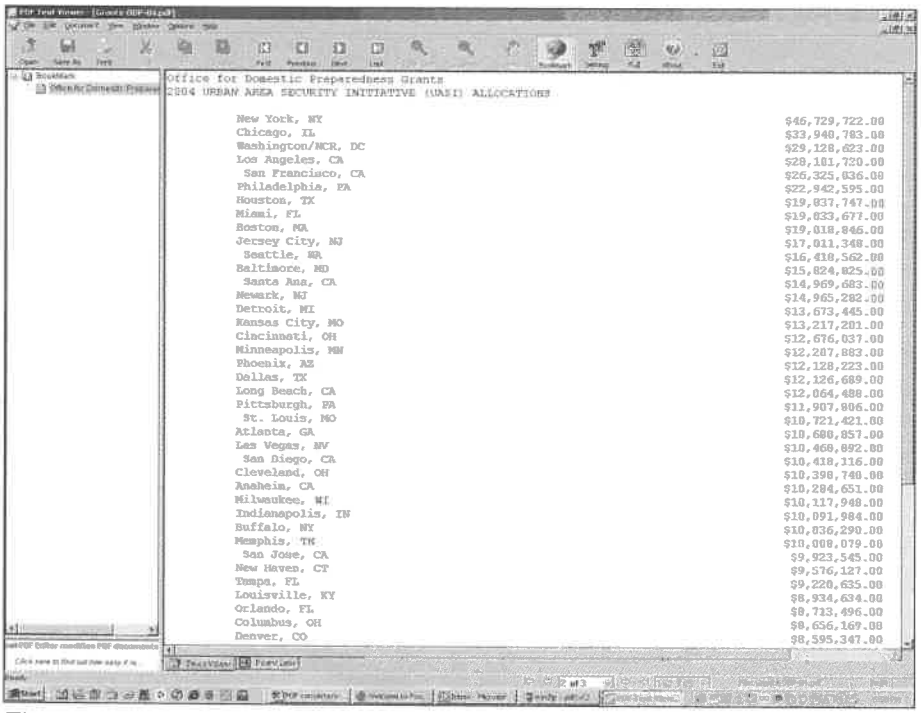
Translation: Let's create software that, with a mouse click, can change any file – including data tables – into an image that cannot be imported into a spreadsheet or database manager. In other words, let's make documents that look like data but cannot be sliced, diced or analyzed.

That's not what Adobe said, and it's probably not what the company intended. But PDFs have been driving geeks crazy ever since. Fortunately, there are a number of free or inexpensive programs that extract data from PDFs and turn it into text that imports easily into CAR software.

PDFs – the acronym also stands for "Portable Document Format" – weren't born evil.

The format, in fact, resolved an important dilemma because they can be created from the output of any software and can be viewed on any computer. A PDF rendition of an Excel spreadsheet created on a PC can display on a Mac, Linux or Palm device even if Excel isn't installed. All that's required is the free PDF viewer distributed by Adobe for all common platforms and some not so common – anyone working on an OS/2, IBM AIX or HP-UX box?

Moreover, because PDFs are hard to edit – not impossible as conventional wisdom holds – there was at least a reasonable assurance that a PDF version of a document was the same as the original document. PDF quickly became the format of choice for attorneys looking to exchange documents for review and for document managers seeking to save storage space by scanning paper documents and saving them as easily retrievable PDF images.



Office for Domestic Preparedness Grants
2004 URBAN AREA SECURITY INITIATIVE (UASI) ALLOCATIONS

New York, NY	$46,729,722.00
Chicago, IL	$33,940,783.00
Washington/NCR, DC	$29,128,623.00
Los Angeles, CA	$28,181,720.00
San Francisco, CA	$26,325,036.00
Philadelphia, PA	$22,942,595.00
Houston, TX	$19,837,747.00
Miami, FL	$19,833,677.00
Boston, MA	$19,018,846.00
Jersey City, NJ	$17,011,348.00
Seattle, WA	$16,418,562.00
Baltimore, MD	$15,824,825.00
Santa Ana, CA	$14,969,683.00
Newark, NJ	$14,965,282.00
Detroit, MI	$13,673,445.00
Kansas City, MO	$13,217,201.00
Cincinnati, OH	$12,676,037.00
Minneapolis, MN	$12,287,883.00
Phoenix, AZ	$12,128,223.00
Dallas, TX	$12,126,689.00
Long Beach, CA	$12,064,488.00
Pittsburgh, PA	$11,907,806.00
St. Louis, MO	$10,721,421.00
Atlanta, GA	$10,680,857.00
Las Vegas, NV	$10,468,892.00
San Diego, CA	$10,418,116.00
Cleveland, OH	$10,398,748.00
Anaheim, CA	$10,284,651.00
Milwaukee, WI	$10,117,948.00
Indianapolis, IN	$10,091,984.00
Buffalo, NY	$10,036,290.00
Memphis, TN	$10,008,079.00
San Jose, CA	$9,923,545.00
New Haven, CT	$9,576,127.00
Tampa, FL	$9,220,635.00
Louisville, KY	$8,934,634.00
Orlando, FL	$8,713,496.00
Columbus, OH	$8,656,169.00
Denver, CO	$8,595,347.00

Figure 1

Creating PDFs used to require purchase of the $300 Adobe Acrobat program. Today all it takes is any of the widely available PDF print drivers distributed for free. Which means, of course, it's even easier for records custodians to make our lives harder.

So, how do we deal with them?

As always, the key is to know the public records law for your state. The question is whether the law views PDFs as responsive to a request for data.

Here in Florida, for example, the law reads, "An agency must provide a copy of the record in the medium requested if the agency maintains the record in that medium…"

Journalists interpret this to mean that, if the data is in Excel, then we can get it in Excel. But some officials (and their lawyers) apparently think "medium" to refer to CDs, DVDs or floppy disks and thus believe that a PDF copy complies with the law. Others claim that their computers can produce nothing but PDFs, obviously incorrect based on the technology itself.

No one in Florida has mounted a challenge to these positions, and the practice seems to be spreading. The law may be clearer, though, in other states.

If a PDF is all you can get, it's time to turn to one of several free or low-cost programs that will turn PDF documents into text; some even use optical character recognition to convert scanned PDFs – paper documents that have been scanned and turned into PDF images. Note, however, that none of these

programs will crack password- or security-protected PDFs.

XPDF/PDFTOTEXT (Free)

For the last few years, PDFTOTEXT has been the way to go. Part of the open source XPDF PDF tools suite, it's available at no cost at *www.foolabs.com/xpdf*. XPDF is a simple program with no fancy Windows interface. It's command-line driven – for you young'uns, that means typing commands from the C:/ prompt.

(Note: IRE and NICAR use PDFTOTEXT in CAR training exercises. A NICAR tutorial with a conversion example can be downloaded at *www.ire.org/training*.)

XPDF does a good job of pulling text from PDFs, but a few other new programs make the job a little easier and add some bells and whistles that XPDF lacks. None are free, but they won't break the bank, either. Evaluation copies of each program can be downloaded from their Web sites.

Foxit Software (*www.foxitsoftware.com*) specializes in PDF utilities and offers a number of products useful in CAR. The company's free PDF Reader is an Acrobat substitute that does not require installation, opens immediately and is less than 1Mb in size.

Foxit PDF Text Viewer ($39)

(Figure 1)

Among the simplest PDF utilities, PDF Text Viewer offers a simple Windows interface and the ability to toggle between PDF and text views. It's a self-contained executable that requires no installation. Text can be saved with a simple "File/Save As" command or it can be cut and pasted from the display. Output is in a fixed-width format easily handled by Excel's import wizard.

Foxit PDF to Text Converter ($59)

(Figure 2)

PDF to Text is a hyped-up version of PDF Text Viewer that lacks the PDF Text toggle function of the viewer, but it makes up for that flaw in other ways. Its Windows Explorer-like interface displays a list of folders, PDFs in each folder and the properties of a selected document.

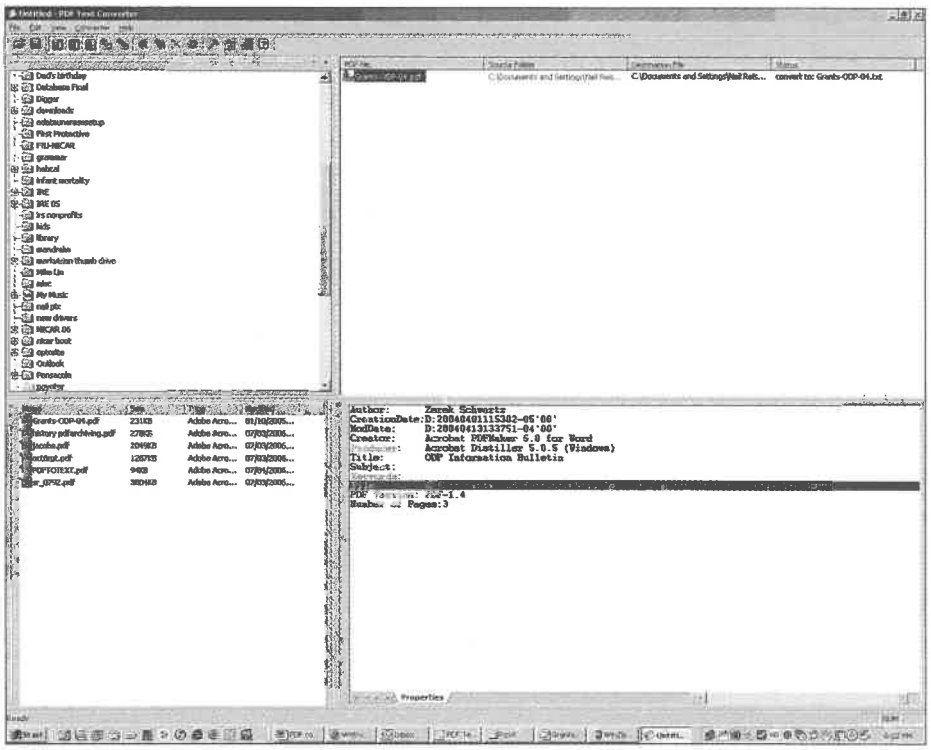


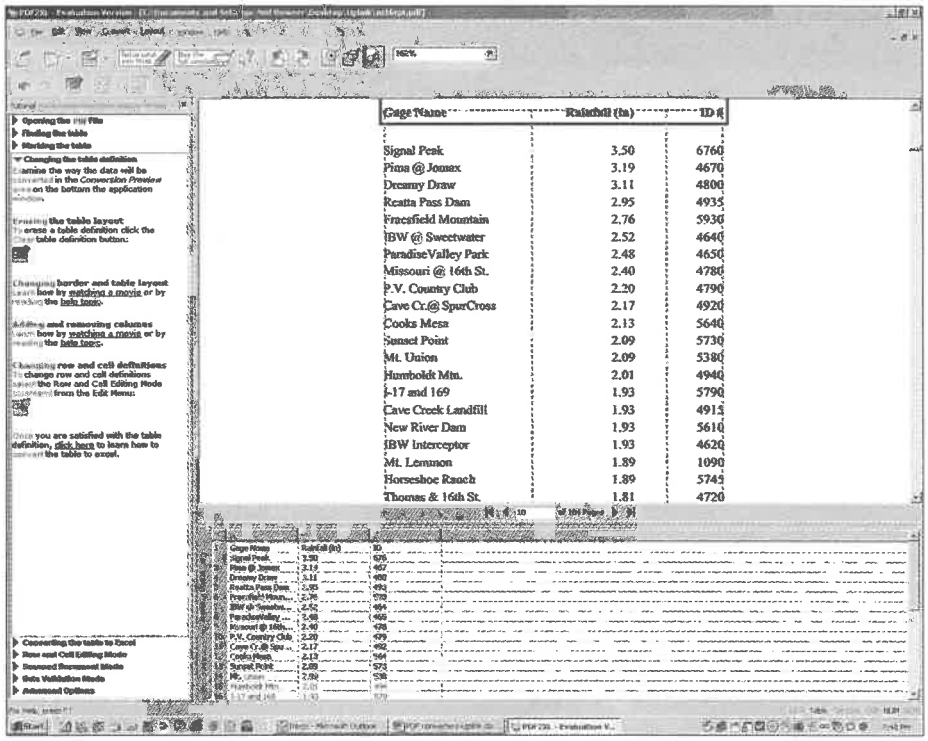
Figure 2

Clicking "Convert" outputs a text version of the PDF. Especially useful is the program's batch conversion tool, which allows users to process multiple PDFs simultaneously. It, too, is a simple executable requiring no installation.

Cogniview PDF2XL ($95)

(Figure 3)

As its name implies, PDF2XL by Cogniview (*www.cogniview.com*) converts PDF tables into text and imports them directly into Excel. It's more sophisticated than Foxit's ap-



Gage Name	Rainfall (in)	ID #
Signal Peak	3.50	6760
Pima @ Jonnex	3.19	4670
Dreamy Draw	3.11	4800
Reatta Pass Dam	2.95	4935
Fraesfield Mountain	2.76	5930
IBW @ Sweetwater	2.52	4640
ParadiseValley Park	2.48	4650
Missouri @ 16th St.	2.40	4780
P.V. Country Club	2.20	4790
Cave Cr.@ SpurCross	2.17	4920
Cooks Mesa	2.13	5640
Sunset Point	2.09	5730
Mt. Union	2.09	5380
Humboldt Mtn.	2.01	4940
I-17 and 169	1.93	5790
Cave Creek Landfill	1.93	4915
New River Dam	1.93	5610
IBW Interceptor	1.93	4620
Mt. Lemmon	1.89	1090
Horseshoe Ranch	1.89	5745
Thomas & 16th St.	1.81	4720

Figure 3

PDFs

continued from page 9

plications, allowing greater control over the conversion process. It also attempts, with some success, to deal with scanned PDFs as well as conventional documents.

PDFs opened in the program are displayed in one pane of a three-pane window. Users select the desired data with the cursor; the program then makes its best guess on column widths. The resulting extracted data is shown in a spreadsheet-like preview window below the PDF pane. Column borders can easily be adjusted by dragging until the preview is correct.

PDF2XL also tries, with some success, to convert scanned PDF images into text using built-in optical character recognition. On detecting a scanned PDF, it allows the user to select a table as usual. It then applies OCR to the document, displaying the result in the preview window. The preview is editable so the OCR output can be proofread against the original document. While the process is on the clunky side, it is far better than keyboarding an entire document or analyzing it by hand.

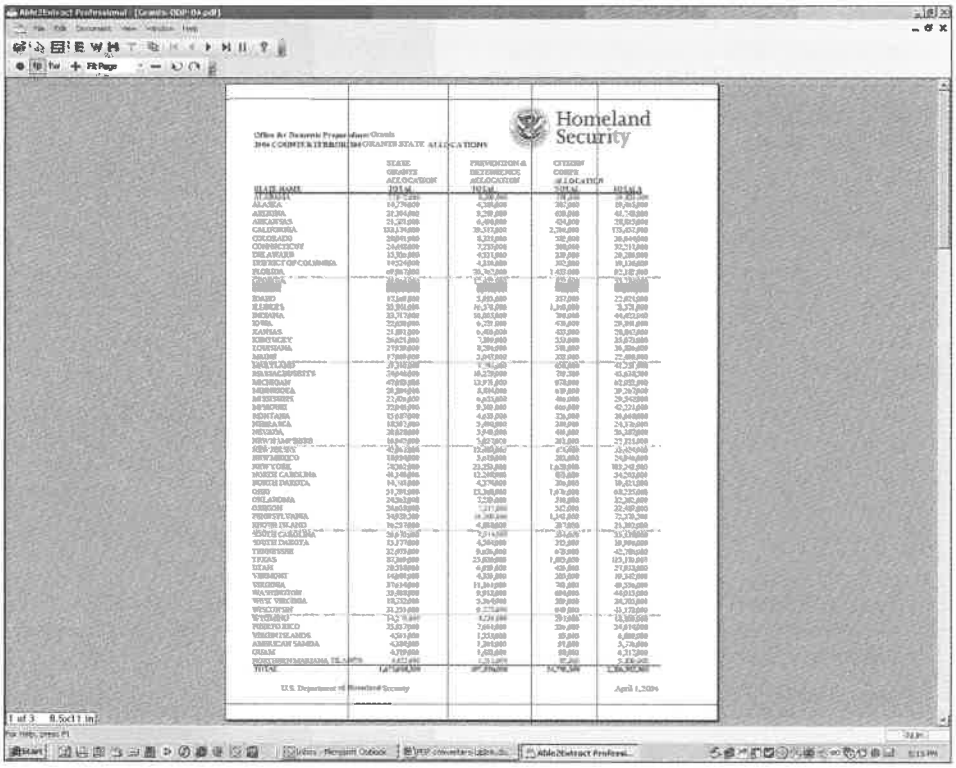


Figure 4

Investintech Able2Extract Pro ($119.95)

(Figure 4)

Like PDF2XL, Able2Extract by Investintech (*www.investintech.com*) extracts text from PDF files and imports them directly into Excel, as well as converting documents to Microsoft Word, text and HTML. It also uses OCR to handle scanned PDF images, though not very successfully.

The program features a slick, intuitive interface and lets the user select exactly what data to convert. It takes its best guess on the data layout but also allows adjustments. It does not, however, have the preview function offered by PDF2XL; you see the result only after the conversion is complete.

The program's OCR function leaves a lot to be desired both in terms of accuracy and because it lacks PDF2XL's data verification/editing functionality.

There are other programs that will convert PDFs into importable text. But based on conversations on the NICAR-L listserv, some adroit searching and some quick tests – in other words, the kind of we-need-a-program-to-do-something-now research journalists often conduct – the programs reviewed here are both affordable and promising.

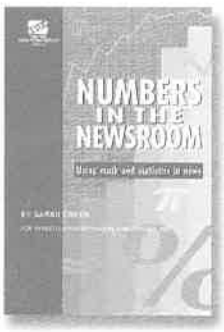
Until now, the venerable XPDF has been the only silver bullet that could tame devil PDFs, and it merits our respect. These programs add more bullets to our gun belts and maybe a wooden stake and some wolf's bane potion as well.

Contact Neil Reisner by e-mail at nr@nreisner.com.

REQUIRED READING For Your Newsroom

Numbers In The Newsroom
Using Math and Statistics in News, by Sarah Cohen

Pulitzer Prize-winning journalist Sarah Cohen guides reporters through fractions, rates, percents and per capitas. Making inflation adjustments. Understanding averages. Doing the budget story. Questioning surveys and polls.



ORDER TODAY!
Call 573-882-3364
with your VISA, MasterCard or American Express
-OR-
Visit our Web site at **www.ire.org/store**
for online ordering or order form downloads

BREAKTHROUGHS

Please make your annual contribution to **IRE**!

"There is nothing more important for a journalist than to seek the truth, speak for those who have no voice, and try to hold power accountable.

Investigative journalists have always felt this was more than a career – it's a trust, and a duty. In the age of blogs and the instant assertions of the internet, it is even more critical to have journalists who check and doublecheck and care.

Investigative Reporters and Editors has always been a place where journalists have learned these skills. IRE has and will always honor the mission and lead the way.

Please join me in supporting IRE during this important time in our history."

– Diane Sawyer
ABC News

FUNDING **IRE**'s FUTURE

2006 Annual IRE Endowment Appeal

Help us ensure that IRE's independent voice, state-of-the art training, up-to-date resource library, and vast array of networking possibilities for journalists continue into the future. Join the hundreds of IRE members who have donated and pledged to IRE's endowment fund.

And, for a short time longer, the John S. and James L. Knight Foundation will give $1 for every $2 you donate or pledge under a $1 million partial matching program. See your philanthropic dollars immediately grow!

IRE seeks donations from all audiences that value quality investigative journalism and recognize its importance in keeping governments, businesses and individuals accountable.

IRE depends on contributions to maintain and improve its services and to create long-term financial stability. At present, more than one-third of IRE's $1.4 million annual budget is funded through donations.

By achieving a $5-million endowment, IRE will ensure its uninterrupted support of investigative journalists and editors and will be able to focus its efforts on emerging programs. In short, a strong endowment will allow IRE to continue to foster excellence in investigative journalism, a mission essential to a free and democratic society.

All endowment gifts make a difference because, taken together, they form a strong foundation that continues to grow for years to come.

Make your dollars grow: Don't forget to inquire...

Don't forget to inquire whether or not your employer will match your donation. For a list of media companies that match gifts to IRE, see: www.ire.org/endowment/matchingcompanies

$500 pledge (over 5 years)
+ $500 match from employer = $1,000
+ $500 Knight Foundations match
= $1,500 endowment increase

To make a contribution, please use the form below, visit www.ire.org/endowment or phone IRE Development Officer Jennifer Erickson at 573-884-2222. All contributions are tax deductible to the fullest extent allowed by law.

YES! I would like to support IRE's Endowment Fund

Name _____

Address _____

City _____ State _____ Zip _____

Company _____

Company Address _____

Wk Phone _____ Hm Phone _____

❒ I will pledge $_____ over _____ years.
❒ My check is enclosed and **made payable to IRE.**
Please write "Endowment" in the memo line of your check.
❒ Please charge my credit card with the amount indicated
❒ VISA ❒ MasterCard ❒ American Express

Account Number _____ Exp. Date _____

Signature _____

I would like my gift to benefit IRE in this way:
❒ Endowment - general operations
❒ Endowment - specific program, services or resource area
Name area _____

Amount of gift $_____
❒ My company will match my contribution
❒ **CHECK BOX IF YOU DON'T WANT YOUR NAME DISCLOSED AS A DONOR.**

Mail or fax this form to:
Investigative Reporters and Editors, Inc.
138 Neff Annex, Columbia, MO 65211
PH 573-882-2042 • www.ire.org
FAX 573-882-5431

2007
CAR
CONFERENCE
IN CLEVELAND
DON'T
MISS A
BEAT!
March 8-11
Presented by Investigative Reporters and Editors and
the National Institute for Computer-Assisted Reporting
2007
CAR Conference
March 8-11, 2007
Cleveland, Ohio
Renaissance Cleveland Hotel
2007 Conference
IRE
Phoenix, Arizona
2007
IRE Conference
June 7-10, 2007
Phoenix, Arizona
Arizona Biltmore Resort & Spa
Find out more at:
www.ire.org/training/conferences.php

Sirens

continued from page 3

to be more painful than geocoding. In fact, it was the most time-consuming aspect of the project. We logged hundreds of miles searching for sirens, and the locations we were given weren't always as accurate as we'd hoped.

Once we had the data, I made the first of many novice mistakes on this project by hand-keying the data captured in the GPS units into the database. (Learning how to upload the data from the GPS is something on my to-do list.) In any event, after the database was complete, it was a snap to add the data to the map, as X,Y points. After adding geographic features such as townships, populated places, and census tract data, I created buffers around each siren. We had our maps – the first comprehensive picture of tornado siren coverage in central Indiana.

I should note that we didn't get to this point without some epic battles with ArcMap. I struggled with everything from projections to buffers. At one point, the program simply crashed repeatedly, sometimes without warning and other times with cryptic error messages – "The instruction at 0x77fcc043 referenced memory at 0x00000038. The memory could not be written. Click on OK to terminate." After scouring technical papers, talking to technical support and re-installing the software, the problem persisted on a couple of county files. I could eliminate the problem by re-creating the files, but I still don't know what caused the error.

Our initial stories looked at the overall picture of tornado sirens in the area, from how well they work to how much area they cover. Using census block data, we were able to report roughly how many people live in dead zones where there are no tornado sirens.

On the WTHR Web site, we posted our own maps of the siren coverage for each county and hired a consultant who created an interactive Google Maps application that allowed users to look up sirens near their homes.

(See the "Cause for Alarm" stories and maps online at *www.wthr.com/global/Category.asp?c=79536*. In the March-April 2006 *Uplink*, Sarah Okeson of the Journal Star in Peoria, Ill., describes using GIS to report on siren coverage for print journalism. Former NICAR data analyst Brian Hamman won a Student Society for News Design award for his news graphic on tornado sirens in the Columbia Missourian.)

Lessons

I learned a few lessons and would do things differently if starting over.

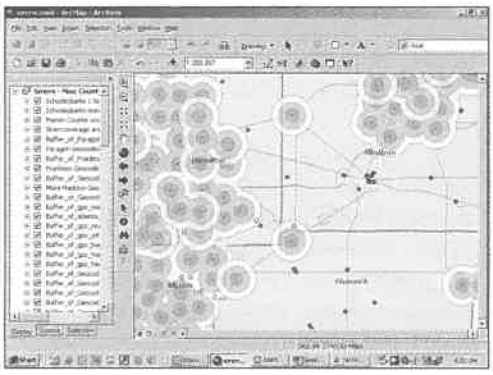
Initially, we had decided to use a generous 1.5-mile siren coverage radius, even though only the best sirens under ideal conditions operate at that range. The reason: most counties were unable to give us specific decibel-level data on individual sirens so we wanted to err on the side of caution. In retrospect, I wished I had handled those buffers differently, creating multiple rings rather than one radius.

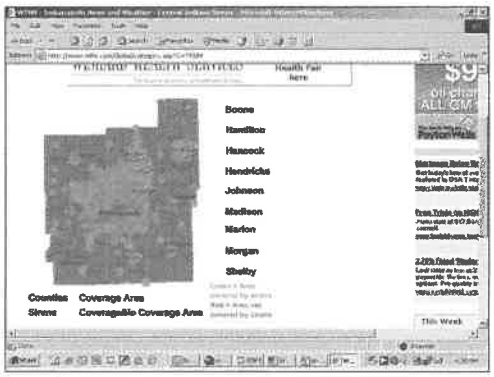
In fact, for a second take on our stories, we did that, showing graduated half-mile distances from the sirens and explaining the variance in siren range in a legend. For this second go-around, we were able to report how many schools and parks are outside the practical range of the sirens. Getting the schools and parks on the map was much simpler than the sirens because we were able to download shapefiles from Indiana's GIS clearinghouse Web site, *http://igs.indiana.edu/arcims/statewide/index.html*.

Another thing I'd change relates to the display. Early in the project, I decided to add each county's siren data as a separate layer because I wanted to have an area map and county maps. I had 16 layers of geocoded and GPS-recorded points. When I added buffers to each of the layers, some overlapped and caused unwanted color variations. If I had to do it again, I would create one master data table that includes a county field.

While ArcMap is exhilarating to use in terms of its analytical power, I also discovered that it's just plain fun to create these visuals that can add so much to our storytelling. I found myself thinking much more like a graphic artist than I ever thought I would, though I readily admit that my initial maps look pretty rudimentary. But, humble as they were, the maps helped us get our point across in ways that we couldn't do before, and they helped us find out what the point of story actually was. (And I didn't even saw off my foot.)

Contact Gerry Lagnosa by e-mail at glagnosa@wthr.com.





The NICAR Database Library's Storm Events database is another weather-tracking resource that contains mappable data. See more at www.ire.org/datalibrary/databases/viewdatabase.php?dbaseindex=49.

SCHOOL DATA

Finding diamonds in the rough

By MaryJo Sylwester, *St. Paul (Minn.) Pioneer Press*

The results from Minnesota's standardized test – used to measure progress for No Child Left Behind – yields a predictable list when sorted best to worst. At the top are schools from the wealthy neighborhoods and suburbs. At the bottom are schools from poor neighborhoods.

This didn't seem like a fair way to assess how schools are performing; yet that is often how test results are portrayed. We decided to dig a little deeper.

In June, the *Pioneer Press* published a series of stories called "Schools that Work," stemming from a statistical analysis of test scores that revealed which elementary schools were doing better than expected, considering the socioeconomic and other demographic factors of their students.

We found 13 schools serving predominantly low-income students that had consistently done better than expected. Then we visited those schools and came up with five things these schools were doing that experts said were the keys to success. The five things were: focusing on reading above all else; using small-group instruction whenever possible; having an effective and cohesive staff; maintaining a structured, disciplined environment; and providing services beyond the classroom such as homework help, dentists and donated clothing.

The bulk of the work was traditional street reporting, but we wouldn't have known where to start without the data analysis to point us in the right direction.

Test score data is readily available in Minnesota from the state department of education, in delimited-text format. Since all states are required to collect test score information as part of No Child Left Behind, chances are good that your state will have data as well. What might differ from state to state is the accessibility and the level of detail provided. Some states allow you to download the data directly from a Web site; others require a public records request. Some provide student-level data, others (like Minnesota) offer only school-level data. You might also have trouble getting the scale (or raw) score. Some states only provide the percentage of students meeting proficiency standards.

Numerous academic studies have shown that a child's family income is the most significant predictor of how he or she will perform on a standardized test. Typically, 50 to 60 percent of the difference in scores between one student and another can be explained by income.

As a result, it's no surprise that schools serving predominantly low-income populations will have much lower average scores. The statistical procedure we used – regression – estimates the impact of income on a school's test scores, then uses that measure to "predict" what each school's average should be.

There are several types of regression, and of course, it can be used to measure many other things besides test scores. The type we used is called multiple regression because we had more than one independent variable.

Our dependent variable was the test score (the scale score). The independent variables were the percentages of students on free or reduced-price lunch (a proxy for income), percentage with limited English proficiency and percentage in special education. As expected, the income variable proved the most significant. English proficiency played a small role, while special education was not significant at all.

Our test results were averages for each school in math and reading for third and fifth graders. And we had three years' worth of data. So we had up to 12 scores for each school (if they had both third and fifth graders). We only used scores where at least 30 students were tested, and schools that didn't meet this mark across all three years were excluded.

Then we used SPSS statistical software to run separate regressions on each set of scores (3rd grade math for 2004-05 would be one set, for example).

I won't get into the specifics about how to run a regression here (you can learn more by reading tipsheets in the IRE Resource Center or attending the IRE and NICAR Advanced CAR Statistics Workshop), but you should know that one of your options in SPSS is to instruct the program to add fields to your data table showing the predicted score and something called a residual.

The residual shows the difference between the predicted score and the actual score. You can have the residual expressed either standardized (the number of standard deviations) or unstandardized (the raw difference between the two numbers). The general practice is that any residuals of one standard deviation or higher can be considered above the predicted results. A negative residual of one standard deviation or more is considered below the prediction.

We decided to go one step further and make sure any schools we were labeling as "above predicted" earned it for being consistently better than expected, across multiple grades, subjects and years. So we averaged the residuals and used those to determine whether a school would be labeled as above, within or below predicted scores.

The statistical analysis took several weeks, partly because we worked on this in between other assignments and partly because the analysis required a lot of trial and error, and double-checking our work with two professors. One was an education professor in Minneapolis who not only taught quantitative analysis

classes but also did her own analyses for publication. The other was a statistics professor from Georgetown University whom I had worked with previously. He is especially good at explaining things in layman's terms. Having two people review my work and help me interpret the results proved comforting.

At this point in the analysis, the big question was, "What do we do now?" We had a list of a couple hundred schools identified as above predicted. Many were the ones you'd expect: the rich suburb or the elite magnet school. But there were some surprises in there, too.

After talking with some school testing experts, we decided to focus on schools with 60 percent or more low-income students. One expert told us that if a school with high-income students did a merely adequate or mediocre job teaching, the students likely would still have above-average performance. But if a school with low-income students had just mediocre teaching, those kids would likely fail.

We also decided to focus only on the overachieving schools and essentially create a roadmap for what schools should be doing, rather than harping on things the other schools were doing wrong. Many of the schools we visited were shocked to learn we were writing about them in a positive light, but none were surprised to hear they were performing above the expectations.

We tapped several other data sources such as student and teacher demographics, teacher experience, and student mobility rates. We generated crosstabs in Access for the schools that fell into each group. For example, this allowed us to see that the above-predicted schools had lower mobility rates and higher averages for teacher experience than schools in the other two groups.

Instead of publishing the predicted scores and residuals, the stories simply identified schools as above, within or below predicted. One reason is that we thought the numbers might add more confusion than it was worth, and that they implied a greater level of precision than our analysis truly had. Experts told us student-level data would yield much more rigorous findings, but Minnesota does not make that data public.

Several people have asked me how I convinced my editors to let me get a regression analysis into the paper. The simple answer is that I didn't. I convinced them to put a good story in the paper. The analysis helped us get there. You'll notice that the stories (except the methodology) are almost devoid of evidence of our analysis. We picked out some key points and relegated most of the analysis results to the file cabinet.

All of the test score and other education data were posted in a searchable database online. The stories and database can be found under "Special Projects" at *www.TwinCities.com*.

Contact MaryJo Sylwester by e-mail at msylwester@pioneerpress.com.

Stateline.org
WHERE POLITICS AND POLICY NEWS CLICK

How can Stateline.org help you?

- Each weekday, *Stateline.org* features an original story on trends in the 50 states -- great story ideas for you.
- Free daily and weekly email newsletters covering top state news.
- Research state trends with our searchable archives.
- *Stateline.org* is a FREE service -- no registration or subscription required.

Disabled

continued from page 1

charities that mainly employ workers who are blind or severely disabled. But our reporting found that nonprofits were increasingly using workers with mild or questionable disabilities.

The biggest contractor, the National Center for the Employment of the Disabled in El Paso, Texas, repeatedly failed to document that it was using enough disabled labor. And it had paid the management firm run by its president a stunning $8.7 million over two years.

Although much of our information came from sources inside the charities themselves, reporters Jeff Kosseff, Bryan Denson and Les Zaitz used the federal Freedom of Information Act and Internal Revenue Service charity disclosure forms (Form 990) to get a grasp of overall trends in the program and target specific charities for investigation.

Kosseff used FOIA to obtain five years of annual reports from the 627 charities. The reports, kept by a tiny federal agency (Committee for Purchase from People Who are Blind or Severely Disabled) that oversees the program, contained data on annual sales, wages paid to the disabled, and the hours worked on federal contracts.

When he received the data in a Microsoft Excel spreadsheet, it was immediately clear that the bulk of the contracts went to the 50 largest charities. We decided to focus our reporting on this group, which the data showed had tremendous revenue growth since 9/11.

War bucks

Denson wanted to know the impact of the wars in Afghanistan and Iraq. He looked up the Dun & Bradstreet D-U-N-S nine-digit identification numbers of each of the 50 contractors, punched them into a spreadsheet and sent them to the Defense Department, asking for a list of their military contracts.

The Pentagon can find contracting records quickly if you provide a contractor's D-U-N-S. You can find these numbers online on Nexis, Hoover's and other sites.

The Pentagon turned the request around in two days. We then had a list of every defense contract for each charity over five years, detailing contract amounts and the type of goods or services provided. The bulk of the growth had indeed come from a burst in military spending.

Defense business was only part of the picture. What about other sources of income – contracts from local governments and private firms – for the charities? And how could we measure whether workers or charities managers were benefiting?

To get aggregate numbers, Denson turned to the charities' tax forms. Gail Hulden, the newsroom's head researcher, used Guidestar Premium (*www.guidestar.org/services/guidestar_premium.jsp*) to acquire 990s from the top 50 charities for 2000 and 2003. We built a data entry form in Microsoft Access, and Denson and two copy aides began the painstaking process of building and proofing a database.

The information they punched into the database included the 9-digit employee identification number (EIN); the tax return year; the name of the nonprofit; the total amount of government contributions; total revenue; total expenses; net assets; officer compensation; total salaries and wages; other wages; related income; and four-year income. This data allowed us to measure growth in all categories.

Money trail

In a access form, they focused on payments to charity officials, entering total compensation for up to five officers, five employees and five contractors paid $50,000 or more. In all, they entered and verified the names, compensation totals and deferred compensation totals for 1,165 officers, employees and subcontractors.

Denson also flagged any insider or related-party transactions noted on the forms for follow-up reporting. In the case of the National Center for the Employment of the Disabled, they were numerous enough to lead us to add a reporter to the project and look hard at the charity's president.

We later added 2004 data when it became available. It allowed us to report that average compensation for top executives at the 50 biggest charities grew three times faster than hourly wages for their disabled workers. At least a dozen execs pulled in $350,000 a year or more in pay and benefits, while many workers were paid less than minimum wage.

The management firm run by Bob Jones, who ran the El Paso nonprofit, had received $14 million in charity payments over five years and loaned millions more to other businesses in which Jones had an interest. Records we requested under FOIA, plus documents from internal charity sources, showed that the charity had repeatedly been unable to document its workers' disabilities. Many were counted as disabled because they spoke little or no English.

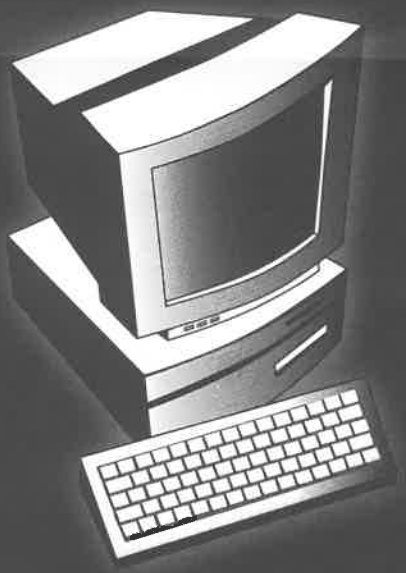
Jones resigned the day our story about him appeared under the headline: "Texas charity a launchpad for entrepreneur's empire." He is now under investigation by the FBI, the IRS and the Texas attorney general's office.

Our first-day story about weak oversight of the contracting program, "Charity leaders prosper as 'disabled' is redefined," prompted calls on Capitol Hill for an overhaul of the Javits-Wagner-O'Day program. It also sparked an intense and ongoing debate within the community of nonprofits that serve the disabled.

The stories and accompanying slide shows can be found online at *www.oregonlive.com/special/disability*.

Contact Tom Detzel at
tomdetzel@news.oregonian.com.

Jumpstart Your Career!



COME TO THE MISSOURI SCHOOL OF JOURNALISM AND HONE YOUR SKILLS AT INVESTIGATIVE REPORTERS AND EDITORS AND THE NATIONAL INSTITUTE FOR COMPUTER-ASSISTED REPORTING (NICAR) WHILE GETTING A GRADUATE DEGREE IN JOURNALISM.

THE MORE THAN A DECADE-LONG COLLABORATION BETWEEN THIS TOP JOURNALISM SCHOOL AND IRE HAS RESULTED IN DOZENS OF JOURNALISTS MAKING THE LEAP TO THE SOME OF THE BEST NEWSROOMS IN THE COUNTRY.

IF YOU QUALIFY FOR ADMISSION TO THIS SCHOOL AND A GRADUATE RESEARCH JOB AT IRE, YOUR TUITION WILL BE WAIVED AND YOU WILL RECEIVE A MONTHLY STIPEND WHILE GETTING HANDS-ON EXPERIENCE NICAR'S RENOWNED DATABASE LIBRARY AND OBTAINING A MASTER'S DEGREE FROM THE SCHOOL.

GRADUATE STUDENTS WORKING IN THE DATABASE LIBRARY HAVE GONE ON TO CAR JOBS AT NEWSROOMS ACROSS THE COUNTRY INCLUDING:

- THE ATLANTA JOURNAL-CONSTITUTION
- THE CENTER FOR PUBLIC INTEGRITY
- THE DALLAS MORNING NEWS
- DATELINE NBC
- DETROIT FREE PRESS
- THE KANSAS CITY STAR
- MIAMI HERALD
- THE NEWS AND OBSERVER IN RALEIGH, N.C.
- THE NEW YORK TIMES
- THE SEATTLE TIMES
- ST. LOUIS POST-DISPATCH
- THE WASHINGTON POST

FOR MORE INFORMATION ON IRE AND NICAR, SEE WWW.IRE.ORG.

FOR MORE DETAILS ON GRADUATE STUDIES AT
THE MISSOURI SCHOOL OF JOURNALISM,
VISIT HTTP://JOURNALISM.MISSOURI.EDU/GRADUATE

OR, E-MAIL BRANT@IRE.ORG OR CALL 573-882-2042.

Helicopter

continued from page 1

Since the Vietnam War, air ambulances have gradually become commonplace. Every major metro area has a steady stream of helicopters landing on the roofs of major trauma centers, bringing the most severely injured and gravely ill patients to the best facilities.

In recent years, air ambulances have helped rural areas cope with major changes in the health care system. As more and more small rural hospitals shut down, helicopters have become increasingly important. For millions of Americans, the only quick way to an emergency room is by air.

During the 1980s, the number of air ambulance flights surged. But in the race to provide more service, safety standards suffered. In the years 1985 through 1988, there were 15 fatal air ambulance helicopter crashes, compared to only seven in the previous four-year period.

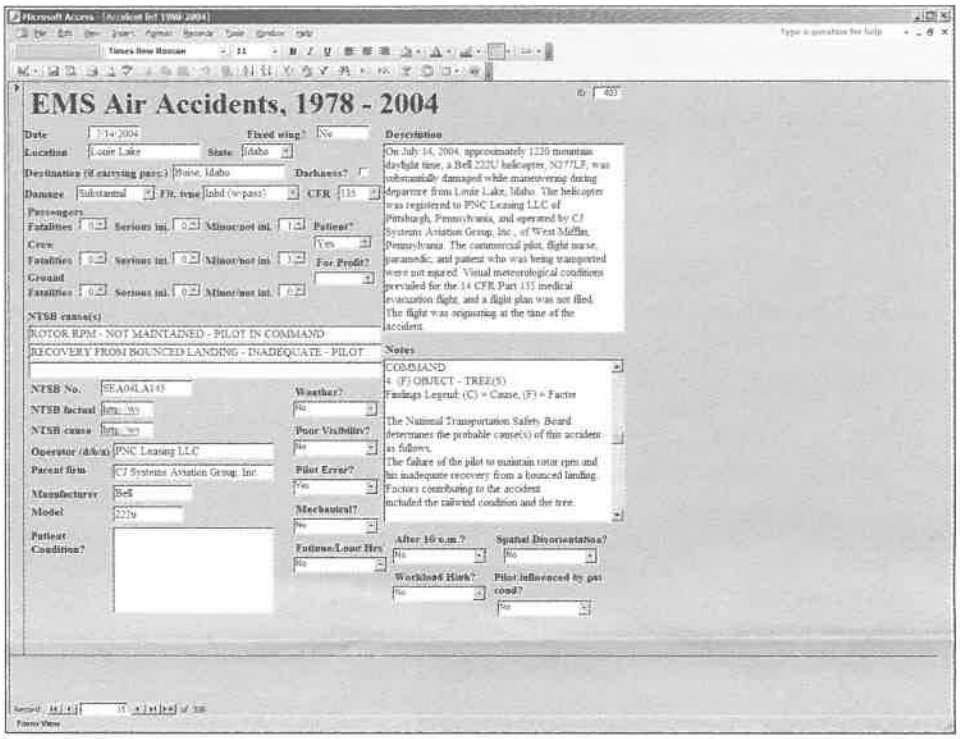
The National Transportation Safety Board attributed the spike in crashes and deaths to carelessness that had crept into overstressed operations and to inadequate federal oversight. The crashes fell rapidly under the glare of attention from investigators and federal aviation regulators.

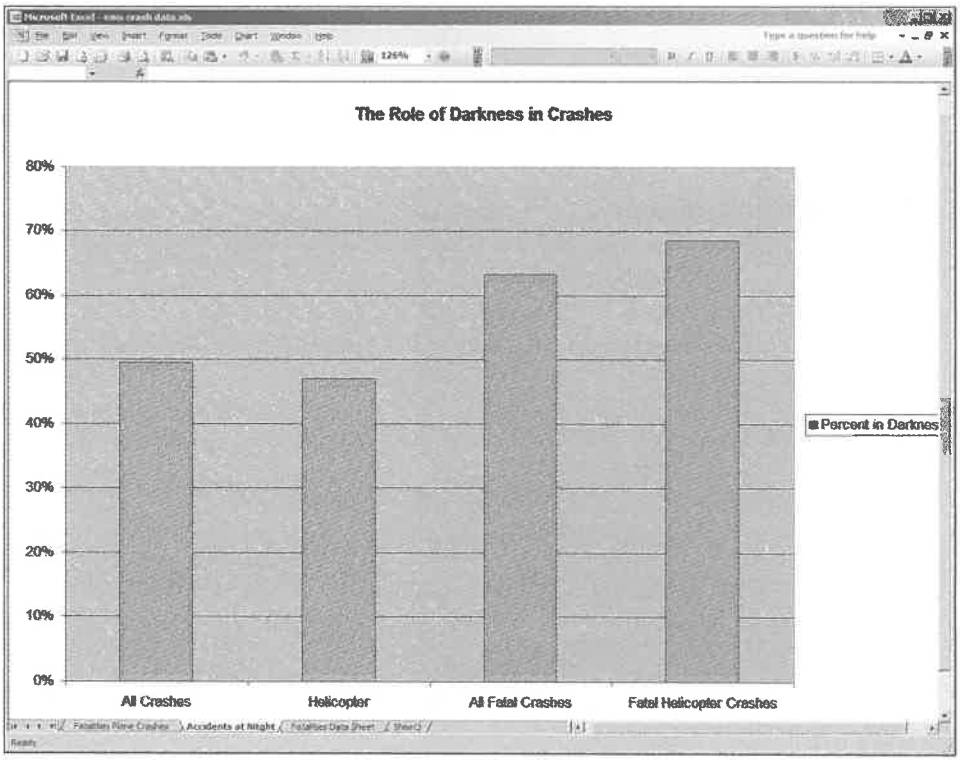
As we began our reporting last November, we suspected that many of the same dynamics had crept into the industry over the past few years, but the NTSB was just beginning its own review of the crashes and had not issued any findings. So we started our own data collection.

As a general rule, the NTSB keeps one of the cleanest databases in government. (As a former database editor at the *Hartford Courant*, I've seen dozens of them. As many of you know, they're often poorly constructed and dotted with typos.) Because the agency has no regulatory power and relies solely on the bully pulpit, it is also unusually generous with its public records.

But when it comes to niches such as the air ambulance world, the database can be difficult to work with.

In order to be included in the NTSB's database, an accident must meet a test that requires factors such as fatalities, serious injuries or significant damage to the aircraft. All air ambulance accidents that met the NTSB's statutory requirements were in its database, but there is nothing in the data to flag these flights. They are simply listed under a broad category of charter

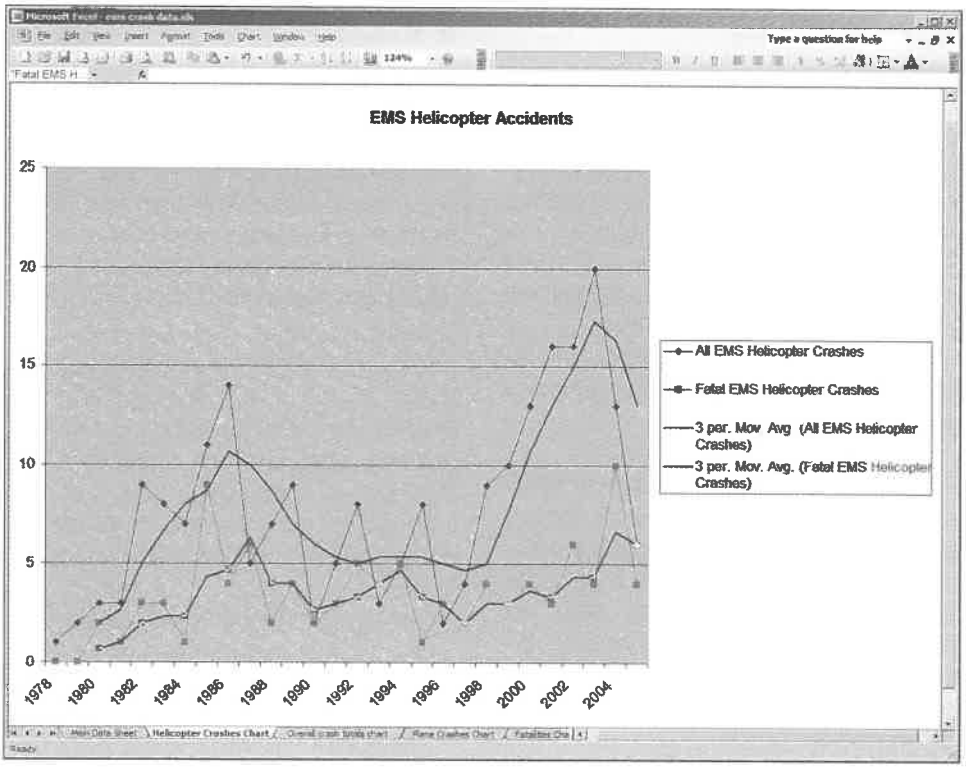




flights. (Though it hasn't been formally announced, we understand the NTSB is planning to modify its data to make it easier to identify crashes involving medical flights.)

We did our best to collect lists of the accidents from a variety of sources. NTSB provided us with a spreadsheet of accidents after 1995 and we combed reports and other sources.

Paul Overberg, a *USA Today* database editor, helped design the database and created a data entry form in Microsoft Access. We set our able intern, Marie Skelton, loose on the NTSB's database. The web version of the database (*www.ntsb.gov/ntsb/query.asp*) includes a fairly robust search function that allows broad text searches. She began plugging in words like "nurse," "ambulance," and "hospital" to look for any cases we might have missed.



To be as thorough as possible, we also entered into our database the accidents involving medical flights on planes, not just helicopters.

Working one or two days a week, Marie finished in about two months. I don't recall the exact number, but by using this crude-but-effective manual search, she found a handful of accidents that others had missed. At our conclusion, we had 279 accidents from 1978 to the early months of 2005.

It turned out that finding the crashes was the relatively easy part of the job. Next, I had to audit Marie's work (which was extremely clean) and to categorize all of the accidents.

The NTSB provides detailed information about the cause of each accident. But that information comes in the form of a fairly lengthy string of text, and over the years, the agency has used different formats.

Here's a typical cause finding: "The failure of the pilot to maintain rotor rpm and his inadequate recovery from a bounced landing. Factors contributing to the accident included the tailwind condition and the tree."

We needed to boil this down to key components, namely the pilot's error, in order to know what patterns were behind the crashes. How important were errors by pilots versus mechanical failures? Were crashes occurring in bad weather? At night?

In this particular case, I concluded that an error by the pilot contributed to the crash, but I did not include this case among those caused by weather. Our definition of a weather-related crash required more than a normal breeze.

The only way to identify whether these broad types of issues were leading to crashes was to read over every single NTSB cause finding. I then added a bunch of additional fields to our database for every conceivable type of accident, from pilot error and mechanical failure to factors such as the time of day.

I went through each record twice to ensure consistency and accuracy. Later on, I met with a statistician at the Federal Aviation Administration to compare our database with one they had compiled to make certain we had not missed any accidents or included cases that didn't belong.

I also went through the data and made sure we had accurate counts of the dead. In particular, we needed to know how many patients had died on flights.

It wasn't pretty, but I hand-entered crash totals and various pieces of information into a Microsoft Excel spreadsheet. That allowed us to help visualize some of the trends. For example, it became obvious very quickly that poor visibility from bad weather and flying at night was frequently a contributing factor to crashes.

continued on page 20

The NICAR Database Library offers several datasets related to aviation accidents and safety. For information, see www.ire.org/datalibrary/databases.

OSHA

continued from page 5

not impossible. Still, if the data shows a company with a high number of violations you can probably say the company was among the top violators.

Another area to look at is the industries with the most violations. You can do that by using the Standard Industrial Classification codes in the inspections table.

These SIC codes pertain to all types of businesses. For example, it breaks down construction by carpentry, electrical work, masonry work and other fields. Picking the industries with the most violations and then looking for companies in those industries with a number of violations will give your story an immediate focus.

The OSHA database goes back to the days of Richard Nixon's presidency so a number of the companies on the database are no longer in business or have changed names. I focused on a 10-year span from 1994 to 2004 to get around that problem.

I've done many stories on workers' safety and I've always gotten a warm response from victims' families. These people are eager to tell their stories, which help make for a dramatic project.

Contact Mike Casey by e-mail at mcasey@kcstar.com.

Helicopter

continued from page 19

We were far from the first news organization to notice that air ambulance flights were crashing at alarming rates. But the painstaking collection of data (and my editors' patience while I completed the work) allowed us to write about the problem with authority.

Even though the NTSB had not yet issued its own findings about the growing number of air ambulance accidents, we were able to reach strong findings on our own. Here are two of our main conclusions:

- "Industry safeguards are so lax that pilots have repeatedly caused accidents by knowingly flying into bad weather, failing to check weather conditions or otherwise violating federal or company regulations. In at least 17 cases since 1995, pilots crashed after flouting fundamental flight rules."

- "About two-thirds of fatal air ambulance crashes occur in poor visibility, the newspaper's analysis shows. Even so, pilots are not required to have special training about what to do when they encounter fog, snow or darkness."

(To see the story online, go to *www.usatoday.com/news/nation/2005-07-17-air-ambulance-crashes_x.htm*).

The information on these crashes was available for anyone to see, but no one had bothered to compile it in one place. The data also helped us identify the most egregious crashes. The narratives of these accidents were horrific and helped make the story readable.

The one thing we wished we could have done was to determine the rate at which air ambulance flights were crashing. The raw number of crashes was the highest in history, but the industry had grown considerably in the past decade. Unfortunately, neither government nor industry collected the number of flights or hours of flight.

Of course, the data was only one of the elements we relied on to report the story. Sources helped us identify patterns of problems and pointed us to documents that highlighted the shortfalls of regulators. We spent just as much time using traditional reporting techniques as we did doing computer analysis.

But without our computer analysis, the story would not have been half as good.

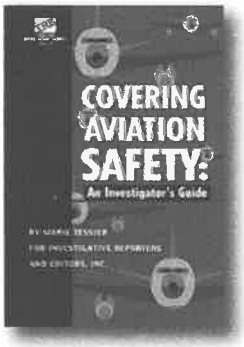
Contact Alan Levin by e-mail at alevin@usatoday.com.

REQUIRED READING FOR YOUR NEWSROOM

Covering Aviation Safety

An Investigator's Guide, by Marie Tessier

Learn to develop a newsroom crash plan, report from the scene of a crash, start an aviation beat, interpret aviation records, negotiate Web-based data, and investigate planes and airlines on deadline.



ORDER TODAY!

Call 573-882-3364
with your VISA, MasterCard or American Express
-OR-

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

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

Latin

continued from page 1

ment Web site that the president himself was unaware of. In another prominent story, CAR tools helped to determine that millions of dollars in pension funds intended for the poor were actually given to people with property and stable government jobs in Costa Rica. CAR helped find that 27 representatives from Rio de Janeiro in Brazil's Chamber of Deputies (the country's lower legislative chamber) had doubled their wealth in five years in suspicious ways.

Spreading CAR skills south of the U.S. border has been a struggle for the past 10 years. Two of the latest initiatives by Latin American journalists had their epicenter in Paraguay and Costa Rica.

In April, a group of Paraguayan journalists launched a new organization, the Paraguayan Journalists Forum, FOPEP, supported by the Knight Center for Journalism in the Americas. Its inauguration was marked by a visit from José Roberto de Toledo, a leading Brazilian CAR practitioner, who conducted a workshop.

One day later, on April 28, Lise Olsen of the *Houston Chronicle* and Costa Rican investigative reporter Giannina Segnini teamed up to train 25 journalists in a CAR workshop at the University of Costa Rica's School of Communication Sciences in San José. The Knight Center and the new Costa Rican Institute for Press and Freedom of Speech, IPLEX, supported this seminar.

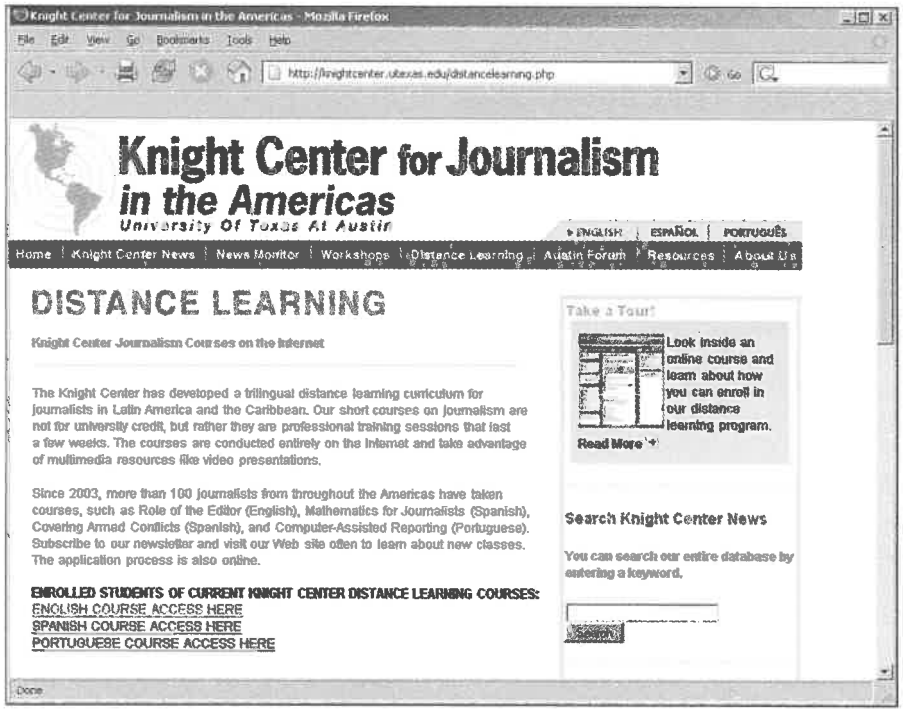
A survey of 89 Latin American CAR-trained journalists, conducted earlier this year as an academic project at the Missouri School of Journalism, sheds light on the export of CAR skills from U.S. journalists to their continental neighbors.

In Latin America, Microsoft Excel is the predominant software used for data analysis. Database managers, such as Microsoft Access or FoxPro, have yet to make headway, and mapping and statistical applications are virtually unknown.

The majority of CAR-trained journalists agreed or strongly agreed that CAR is feasible in their countries. But only a minority said that CAR is understood in their newsrooms, the survey found.

Learning to drive

CAR arrived in Mexico in 1996 as an IRE initiative to develop CPI, a similar training organization with Olsen its first director.



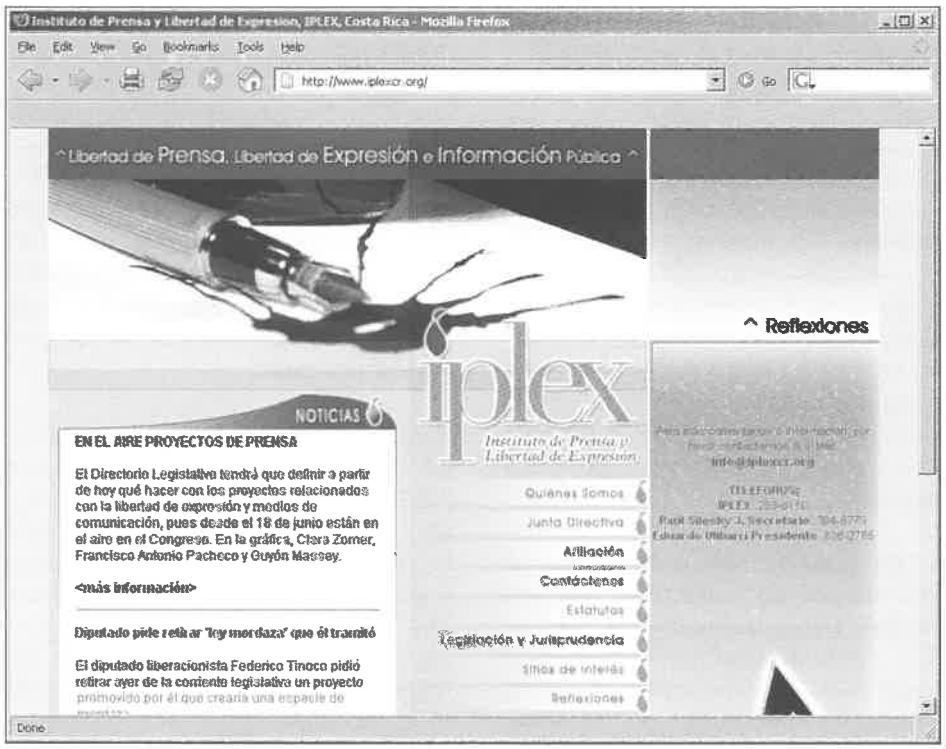
CPI introduced the skills of finding, breaking down and analyzing electronic information to Spanish and Portuguese-speaking newsrooms with varying degrees of success. After being a pioneer in Latin America, CPI is no longer the lone active CAR trainer. New projects have emerged in the region while CPI explores new ways to revitalize its mission after having exhausted the U.S. funding, a $544,000 McCormick Tribune Foundation grant.

In 2002, professor Rosental Alves founded the Knight Center for Journalism in the Americas – a professional training and outreach program for journalists in Latin America and the Caribbean – at the University of Texas at Austin. The introduction of online CAR training can be rightfully attributed to the Knight Center, a project currently funded through June 2008. The distance-learning project started in 2003 and is located at *http://knightcenter.utexas.edu/distancelearning.php*. Since then, two online CAR workshops have been taught in Portuguese, reaching 26 journalists from more than 12 Brazilian cities at a time. The online technology also provided seven sessions of Mathematics for Journalists, benefiting more than 100 reporters.

From a different trench, a former CPI trainer also helps to keep the legacy alive. Alma Delia Fuentes, an instructor since 1997, joined the National Training Program for Journalists in Mexico, a not-for-profit organization, and is a CAR teacher at the Instituto Tecnológico de Monterrey.

Fuentes says CAR "wears diapers" in Mexico, but it is steadily evolving: "I am just getting them [the students] to under-

The IRE Resource Center now offers a growing collection of Spanish-language stories, tipsheets, and beat resources for journalists at www.ire.org/esp.



stand that CAR is also for the daily story, that you do not have to save it for 'special occasions,' that to master CAR the best route is to start with small projects."

According to the Missouri School of Journalism survey, in Mexico, 22 percent of CAR-trained journalists reported using the skills every day and 34.1 percent said they use them once a week. In Brazil, 42.9 percent reported working daily on their own CAR-related projects.

Latin CAR leaders

Some can claim excellence in practicing CAR. Their names surface time and again as field leaders; Colombian Gerardo Reyes, Venezuelan Carlos Subero; Brazilians de Toledo, Marcelo Soares, Fernando Rodrigues, and Angelina Nunes; Argentineans Daniel Santoro and Sandra Crucianelli, Costa Rican Segnini and Mexicans Armendares and Fuentes.

They have written stories that have embarrassed presidents, reformed bureaucracies, revealed political patronage and exposed drug traffic details.

Brazil seems to be a harbor for many leaders. José Roberto de Toledo, 40, was among the first Brazilian journalists to be trained by CPI in 1996. His first CAR story was a detailed account of the flight of human capital from Brazil to the U.S. lured by attractive professional offers.

"Lise [Olsen] showed us an American immigration database," said de Toledo, whose story ran on the front page of *Folha de São Paulo*, the largest circulation newspaper in Latin America.

Like de Toledo, 30 percent of Brazilian and 61 percent of Mexican CAR-trained journalists have had a CAR story published on page one, according to the Missouri survey.

Brant Houston, IRE's executive director, highlights common ground for the vitality in Brazil and Mexico: CPI's impact in the region during the 1990s and U.S. funding.

CAR also got a big push in Mexico in June 2002, when the first federal open-records legislation was approved. Sixty percent of Mexican CAR-trained journalists surveyed reckoned the new law eased their access to digital public information.

Because of the law, "Mexico's unique case might not be representative of the state of CAR in the whole region," Armendares said.

In spite of adverse legal climates compared to Mexico, CAR has grown the strongest in Brazil, and even their Mexican colleagues consider Brazil's CAR development over the last decade a success story.

A combination of initiatives is responsible for this achievement. *Folha de São Paulo* has for 20 years sponsored a training program for both its staff and for journalists fresh out of school. By the end of 2002, a group of journalists founded the Brazilian Association of Investigative Reporting, a not-for-profit organization with economical and technical support from the Knight Center, meant to train journalists throughout Brazil.

Bumpy road

However unthinkable it might be to disassociate U.S. journalism and computers, many Latin American journalists from rural areas have limited access to technology and the education needed to use it.

"Frequently journalists did not study journalism, might not have finished high school and have two or three jobs to make ends meet," Mexican CAR trainer Fuentes said.

The lack of familiarity with arithmetic is also an obstacle. Gradually, CAR-trained journalists have made peace with mathematics. Only 20 percent of Spanish and 18 percent of Portuguese-speaking journalists said they did not enjoy it, according to the Missouri survey. Seventy-one percent of Brazilians and 60 percent of Mexicans reported feeling comfortable computing ratios and percentages.

There are also absurd situations. *O Estado de São Paulo*, a São Paulo newspaper, asked de Toledo to give a CAR workshop in their newsroom, but they had removed Excel from their computers.

"They asked me 'How do you practice CAR without Excel or a computer?' Well,