Uplink

THE FORUM FOR COMPUTER-ASSISTED REPORTING

March 1993
Volume 4, Number 3

Getting data into the right hands: the readers'

By Chris Feola
Waterbury Republican-American
Waterbury, Conn.

If your database ran queries in a forest, but no one was there to access it, would your readers hear about it?

The Zen cliche about the tree falling in the forest is an examination of existence. My hackneyed refrain is a question of practicality.

In the end, a newspaper is no better than the editions it publishes. A paper can have racks of the latest equipment, access to every piece of data known to man, the deepest thinkers thinking the deepest thoughts — and all of that matters not if the paper that rolls off the presses each night is a piece of junk.

Here's the same question, then, in a more serious form: How do we get our data out of our databases and into the hands of our reporters?

We all know how to do it on the big projects. We all know how to run XDB on 42 bazillion records.

But what about the stories that aren't brainsurgery? What if a reporter doesn't need to do an asymmetrical outer join on a series of campaign contribution tables? What if they just want to know if Mr. Big gave the mayor money — Yes or No?

And what if they need to know on deadline? What if they need to know at the council meeting, so they can ask the mayor *tonight* why he's voting on Mr. Big's project?

We faced these questions at the *Waterbury Republican-American* about a year ago when we realized we were running incomplete stories because no one outside the computer-assisted reporting team had a clue when it came to the resources we had available.

We decided to attack the problem on two fronts: We started equipping all our reporters with i386 notebooks to give them their own computer access; and we began building information appliances to allow them to get the data without having to learn Paradox Application Language or anything similar.

The notebook setup is fairly straightforward. Each reporter is given a notebook — we are currently using AST i386SX 20 megahertz machines with 100 megabyte hard drives, six megabytes of RAM and 2400 baud FAX modems.

All that comes in a six-pound package that runs on batteries and fits in a briefcase. We also hand out monochrome VGA monitors and full-size keyboards that plug into the notebooks. In effect, each reporter has a notebook for the field and a desktop for the office.

Application appliances are programs designed to give users point-and-shoot access to data — the idea is that they are no more difficult to use than a microwave.

Here's an example. We have a Paradox database listing all those who contributed to our mayor's election war chest. Rather than having all our reporters learn Paradox — we teach everyone who is interested, but we don't force anyone — we gave them access to the data through a menu choice in their word processor.

Here's how we did it. First, we stripped our Paradox tables down to raw ASCII, then imported them into WordPerfect for Windows 5.1. Then we built a WordPerfect macro and hung it on a word processor menu.

The macro pulls up the file and posts a dialog box that asks "What is the name of the person?" The reporter types the name in and clicks "OK." The macro then runs a search based on the name. If there is a hit, the information is pasted into the story the reporter is working on. If the search comes up empty, the macro tells the reporter "That person is not on the list" and sends them back to their story.

In either case, the reporter has an answer as fast as they can ask the question.

We have used this technique on more than just this list. Reporters can also search municipal employee lists for their towns with just a click of a mouse button. And we're doing the same thing with census data for their towns.

This is probably not the most elegant piece of programming you've ever seen. But here's the bottom line: Our reporters can get their hands on the data they need in seconds. And so they use it.

Orange County Register settles the score on soft-money contributions

By Kristin Baird
MICAR

The stats read like the salary list of a major league baseball team: ARCO — $804,842; Michael Kojima — $613,850; Sony Corp. — $332,650; Walt Disney Co. — $196,207.

But this game was politics, played by corporations and private individuals from California who donated large sums of soft money during the 1991-92 elections. The final score: $13.4 million in soft money contributions from California alone.

Ronald Campbell of the *Orange County Register* analyzed campaign-contribution computer tapes from the National Library on Money & Politics to tally up the scorecard on California's soft-money contributors.

Soft money donations go directly to the national political party instead of a particular candidate. This provides an exception to the federal law that limits campaign contributions to $1,000 a person and prohibits businesses and labor unions from contributing.

Using FoxPro on a 386, Campbell quickly generated a long list of wealthy Californians who took advantage of the soft money loophole. But who were these people and corporations? That proved to be the hardest and most interesting part of Campbell's research.

Many contributors were household names, such as Sony, Chevron, Sunkist and Peter Norton. But what about Ben D. Kelts, or G.S. Ishiyama & Co.?

To unmask mystery donors, Campbell turned to a new "toy" in the *Register* library — a national phone directory on compact disc. Campbell struck gold.

"I went to the machine, typed in "Kelts, Ben," and there was one of them," he said. "Of all the dumb luck, to find a guy who gives $100,000 listed in the phone book."

Campbell also got help from his fellow reporters in identifying other contributors.

After identifying most of the donors, Campbell then broke the contributions down by industry. The help he needed was right there on computer tape. The National Library on Money & Politics had designated fields on the tape with interest codes.

"That is very useful," Campbell said. "If you want to know how much money came from the entertainment industry, you have the computer go through, take the interest code and sum the amount."

Party loyalties became evident through the industry breakdown. Hollywood and labor unions went Democrat; energy moguls and agriculture went Republican.

But something was missing — the developers. Campbell noticed only a small number of them appeared among the big campaign contributors. An explanation was needed; Campbell decided to compare contributions to those made in 1988, a peak year in the Orange County economy.

Campbell looked once again to the Nati[illegible] Library of Money & Politics. Because the F[illegible] Election Commission did not require parties to report soft money until 1991, the library warned Campbell that the 1988 tapes would not be complete. Campbell was willing to take the risk.

"It is constructive to compare the partial picture of the late '80s with the complete picture of the '90s, because the contrast was quite glaring," he said.

Glaring, indeed. Real-estate developers who had dominated the county's soft-money heavy-hitters in 1988 were replaced in 1991-92 by a restaurant group and a health-care conglomerate.

The incomplete data from 1988 did cause a few problems. Several corporations listed as contributors in 1991-92 were known to be members of the Team 100, a designation for those that donated $100,000 or more to the Bush/Quayle campaign in 1988. When the 1988 computer tape failed to show several Team 100 members and the Republican party declined to release information on donations above $100,000, Campbell could only guess that the corporations may have donated larger amounts in 1988 as well.

Campbell is still following up on the story. Since the article's publication on Feb. 1, Campbell has discovered the identities of two previously unknown donors and is working on several others. But one question still weighs heavily on his [illegible]d.

"What the blazes are they getting for the money?" Campbell said. "The side with the most money wins. But what they do because of that money, I don't know."

The data entry ordeal

What to do when you don't have unfailingly accurate slave labor

By Deirdre Shesgreen
MICAR

Is the task of putting a few hundred thousand records on computer stopping you from cracking that great campaign finance story?

Data entry is not most investigative journalists' idea of a good time — but often it comes with the territory.

Hiring a data entry service is a solution, but they come with problems of their own.

The *Kansas City Star* learned the hard way while investigating the abuse of Missouri's "Second Injury" worker's compensation fund.

"We tried a so-called professional typist to do this, and it was disaster," said Rich Hood, the *Star*'s political correspondent. "Among other things, she left out the contribution amounts. She went like a bat out of hell, but it was unusable."

So the *Star* spent hundreds of hours entering the data itself. "We checked and rechecked the accuracy of the thing so we could stand behind the story," Hood said.

If the government wants to create a database, they can go to the nearest prison and enlist inmates to do the tedious task.

"Two ways that government inputs info into computers is through civil servants and prison inmates," said John Roure, who runs Optic Solution, a company that builds databases for law-enforcement agencies, unions, PACs and other associations.

Inmates at the Texas Department of Criminal Justice have done data entry for workers compensation commission reports, records from the Department of Transportation and other agencies, said Phillip Anderson, industrial supervisor of the Huntsville prison.

But the government mainly uses prisoner labor for non-sensitive data entry, Roure said.

Unfortunately, newspapers don't have access to the same human or financial resources. So enthusiastic and dedicated reporters end up spending months in front of a computer screen, coming away with a great story but glazed eyes.

One source of cheap labor newspapers do have access to is off-shore data entry companies. Although using a plant in Mexico that pays its employees $4 a day may raise some ethical eyebrows, it is cheaper than the U.S. firms.

Martin Horan, who runs Horan Data Services Inc., in Cincinnati, just lost a bid to an off-shore data entry company.

"We really can't compete with their prices," Horan said.

A local firm wanting to keep its business in the United States asked Horan to give them his best price. Unfortunately, Horan said, "the price they got from the off-shore place was half the best price I could offer."

But the offshore data services' accuracy is a problem. Horan said that the local company is frustrated with the quality of the off-shore service.

"They don't have the luxury of handling questions about the data on a timely basis, so they usually don't handle them at all."

When the *St. Petersburg Times* investigated abuse of a Florida law that allowed people charged with a crime to erase all records of their arrest, they had to go through hundreds of thousands of pages of information, said *Times* reporter David Barstow.

Where information is easily understandable, they could have sent the data to a data entry service, said Barstow. But with some types of documents, the chance of error, or of missing a crucial clue, was too great.

"You can't send it to someone who doesn't know what you're thinking."

— David Barstow, St. Petersburg Times

And as you sort through the piles of paper, other things can come up. "You can't send it to some one who doesn't know what you're thinking," he said.

"We did consider trying to get journalism students to do it because it is pure drudgery — mindless work — but the only way to be absolutely confident about the data is to enter it yourself," Barstow said.

Barstow said the *St. Petersburg Times* did consider giving campaign contribution data to an off-shore data entry service.

"We explored the possibility of farming out to Jamaica, but it was still enormously expensive and it was not a possibility in terms of hardcore reporting. Too many things can go wrong."

In the United States, data entry is an expensive endeavor. Getting 100,000 typed records with 16 fields put into a database could cost $27,000 — $54,000 if you want the information double-checked, said Gayle Myers of International Data Corp. in St. Louis. And it could take on average four weeks, depending on other factors.

"I built my own database with index cards. Then I decided it was time to move into the twentieth century."

— David Morris, AP

International Data estimates costs on a per-record basis. The rate is higher for smaller jobs because they still charge a set-up fee. For 1,500 typed records with 16 fields, the bill would be between $250 and $425.

"They (data entry services) don't handle small jobs. They're used to doing things for IBM. If they do a few thousand records they charge exorbitant rates," Barstow said.

But for the *Cleveland Plain Dealer*, the $10,000 tag on the database was worth keeping their reporters reporting. The *Plain Dealer* paid 18 cents per record.

"There's a trade off with concerns about accuracy and judgment calls. But if you do data entry you're not doing reporting," said Dave Davis, investigative reporter for the *Plain Dealer*. Davis said that if it doesn't take reporters away from reporting for very long, they would prefer to do data entry in house.

"But it's not worth it if it's something that's going to take lot of time," Davis said. Time pressures added incentive to hire outside help.

So the *Plain Dealer* went to a company they knew and had worked with before.

"We created the record layout and hired people to go through the records and input the data. We were looking over their shoulder the whole time."

"I told them not to guess and gave them my home number. They called a lot."

Another possible solution is illustrated by a cooperative effort among state legislative reporters and a professor at Pennsylvania State University.

An organization of full-time Pennsylvania state government reporters funded a project to computerize all the campaign contributions to the candidates for governor and state legislature in 1989 and 1990.

"The reason I got interested in the project was because I spent about a year tracking contributions from lobbyists — just doctors, lawyers and insurers. But there were 253 legislators," said David Morris, then-president of the Pennsylvania Legislative Correspondence Association and a state reporter for The Associated Press.

"I built my own database with index cards. Then I decided it was time to move into the 20th century," said Morris, who is now head of the AP bureau in Sacramento.

He teamed up with James Eisenstein, a Penn State professor of political science who has studied campaign finance since 1976. With the help of work-study students, Eisenstein computerized almost 35,000 contributions in all, plus another 2,000 records of summary information. "We computerized all contributions except individual contributions between $50 to $250," said Eisenstein.

The project cost about $20,000.

"What we are trying to do should be done by the state," Eisenstein said.

Eisenstein has tried to keep the project going by forming a coalition of all people who would be interested in the information.

The Pennsylvania Newspaper Publishers' Association was the largest financier, giving about $5,500.

In return, three versions of the information were made available to all members of PNPA.

The first — the blue book — was an index of all contributions by name of candidate, type of contribution and alphabetical listing of contributor in each category.

The second — the red book — gave all PAC contributions sorted by the name of the PAC and then the name of the recipient.

The third — a database application — allowed users to make specific inquiries by candidate name, by contributor name and SIC code. The searches could be specified by chamber party, election outcome, and date.

"So, for instance, you could find out all contributions that went to Republican senate losers in the primary between March 2 and April 2 in 1990. It is a very powerful program," said Eisenstein.

By including the SIC codes, "it does one thing that no other data base does," according to Eisenstein. It automatically classifies PACs by industry.

The date functions lets the reporter see if there is a relationship between the timing of a contribution and the legislative agenda.

But the project is petering out. The coalition only came up with $8,250 to computerize 1992 campaign contributions, not enough to get past the primary election reports.

But Eisenstein said he hasn't given up on the idea yet.

"About 20 organizations got at least one version of the reports," Eisenstein said. "We had a good, but not massive, response."

Using the database "became pretty routine," said Morris. Now that he's moved to Sacramento, Morris said he misses the luxury of using the database.

"It's a pain in the butt because you have to go through all this paper by hand. You have to go to the secretary of state, find the paper and take notes or make copies."

Morris is trying to get some news organizations interested in starting up a California database, but "it's a long, cumbersome process," he said.

Adventures in sampling:

Using brute force to compensate for your ignorance about statistics

By George Landau
St. Louis Post-Dispatch

Using your computer's ability to generate random numbers and to repeat tasks thousands of times in a few minutes, you can use common sense — rather than a statistical formula — to figure out how likely it is that the results of your investigation are a chance occurrence.

The technique is called sampling. It requires only that you write a program that tells your computer to dip repeatedly into the data while keeping track of the results.

I decided to use this technique when I was working with the data from a corrupt workers compensation fund in Missouri. Using a database of about 12,000 claims, we had identified 15 lawyers who were working both sides of the fund — alternately filing claims on behalf of injured workers and then, with other cases, working *for* the state defending the fund against excessive claims.

Our analysis showed that this group of 15 lawyers somehow managed to win settlements that were about three-and-a-half times higher than everybody else's. One question we were asking ourselves (and that we knew our critics might raise) was this:

What are the odds that this disparity between the 15 lawyers and the rest (who numbered about 1,000) is due to chance? In other words, if you randomly picked groups of 15 lawyers from the database, how often would their combined average settlement be 3.5 times higher than everybody else's?

As you can probably gather at this point, the answer lay in the question itself. We created a table that listed each attorney whose name appeared in the database, along with the total number of cases he filed against the fund and the grand total of his settlements.

Then we wrote a program (in FoxPro, although you could do the same in Paradox) that instructed the computer to do the following:

- Do three things 15 times:
  1. Pick a number between 1 and the total number of records.
  2. Goto that record (i.e., if the random number is 553, goto the 553rd record).
  3. Mark the record, either by tagging it for deletion or setting a value in a marker field.
- Count the total value of settlements and the total number of claims for these 15 lawyers (summing the marked records only).
- Calculate their average settlement (total dollars/total cases) and place this number in a field called "avg_15" in another database table called "results."
- Count the total value of settlements and the total number of claims for everyone *but* those 15 lawyers (summing the unmarked records only).
- Again, calculate their average settlement (total dollars/total cases) and put this number in the "results" table, into a field called "avg_rest" in the same record you just wrote to.
- Unmark all records in the table of lawyers.
- Do it all again. And again. And again.

The technique is called *sampling*. It requires only that you write a program that tells your computer to dip repeatedly into the data while keeping track of the results.

We used a **FOR/NEXT** loop that counted from 1 to 10,000, doing the above routine 10,000 times. It took maybe 20 minutes to run. If the database had been much larger or the computer slower, we simply would have run it overnight. We probably didn't need to repeat the sampling loop so many times, but hey — we wanted to be sure.

With the inhumane chore of data sampling done, the answer to the question — how often would this happen by chance — was quickly in hand. We just counted the number of records in our "results" table where the 15 lawyers' average settlement was at least 3.5 times larger than everybody else's. There were about 60 such records among those 10,000. So the likelihood that our story was the result of chance was 6 out of 1,000, or 0.6 percent.

Knowing that pollsters and scientists are happy when the uncertainty falls below 5 percent, we were downright giddy.

Bits, bytes and nibbles

Congratulations to reporters at the *The Cleveland Plain Dealer* and *The National Law Journal* whose computer-assisted stories won 1992 IRE awards for investigative reporting.

"Unequal Protection: the Racial Divide in Environmental Law," by Marianne Lavelle, Marcia Coyle and Claudia MaClachlan of *The National Law Journal* won in the category for newspapers with circulation less than 75,000.

The National Law Journal reporters used computerized records to prove that a community's color influences the way the federal government cleans up toxic waste sites and punishes polluters.

They used pre-collected data and also built two of their own databases for the story.

"Lethal Doses — Radiation That Kills," by Dave Davis and Ted Wendling of the *Plain Dealer* won in the category for newspapers with circulation over 75,000.

Davis and Wendling used several Nuclear Regulatory Commission data files to show how the agency has failed to protect the public from radiation overexposures and accidents in the nation's hospitals.

The February issue of Uplink featured a story by Davis and Wendling on how they did the series.

• • • • • •

Public Information Research offers NameBase — a database of citations for over 67,000 names of groups and individuals.

NameBase, compiled from over 400 investigative books and thousands of pages from periodicals, is said to specialize in sources that are ignored by expensive on-line search services. Citations for each name include author, title, date, and page number. Almost all titles contain a screen of annotations.

Areas emphasized include the international intelligence community, U.S. foreign policy, political elites from the right and the left, assassination theory, Latin America and big business.

Leading letter and phonetic searches can be used to locate difficult or transliterated names.

Some sources which have been added within the last year include: 2,800 names from the 1991 Membership Directory for the Association of Former Intelligence Officers; 395 names from the 1992 Federal Staff Directory; 996 names from the state Department Biographic Register.

NameBase fits on diskette and costs $79. Updates are available for $39. For information, contact Public Information Research, P.O. Box 5199, Arlington VA 22205, (703) 241-5437.

• • • • •

■ Two computer-assisted stories win IRE awards

■ NameBase offers 67,000 hard-to-find names

THE MISSOURI INSTITUTE FOR
COMPUTER-ASSISTED REPORTING

120 Neff Hall
University of Missouri
Columbia, Mo. 65211
(314) 882-0684