Uplink

THE FORUM FOR COMPUTER-ASSISTED REPORTING

April/May 1993
Volume 4, Number 4

Suspicious figures

Be wary of government estimates for providing public data

By David Armstrong
Boston Herald

Two recent requests for public records in electronic form by the *Boston Herald* highlight many of the pitfalls and irritating techniques used by bureaucrats to keep information from reporters and their computers.

In the first case, which began with a request more than two years ago, the *Herald* sought a copy of the Massachusetts motor vehicle registration database.

We planned to use the database more as a people finder and research tool for reporters than for a particular project. However, the database's potential for projects was obvious.

The response to the request for the registration file was unfortunately predictable.

The information we requested was available, but would cost $168,000. However, the Massachusetts Registry of Motor Vehicles said it would offer the Herald a reduced fee of $3,134 because the Herald was a media outlet and was not using the information for commercial purposes.

While $3,134 was better than $168,000, it was still unacceptable.

The state said it had recently granted the rival *Boston Globe* the same reduced fee and the *Globe* purchased the database. We wanted to see the rationale for the *Globe* fee and formally requested a copy of any documents reflecting price estimates for the *Globe*.

This was important because, under Massachusetts law, a state agency can only charge the "actual cost" of producing the information requested by a public record seeker.

Sure enough, the Registry of Motor Vehicles bilked the *Globe*. The agency charged the newspaper $600 a minute for CPU time. Yes, $600 a minute.

In early 1991, the *Herald* questioned the CPU cost charged to the *Globe*, which was the basis for the fee the Registry offered to charge the *Herald*.

We informed the Registry that if the CPU time was calculated for a full year, based on 50 percent usage, the total cost would be $155 million. The registry's computer, purchased in 1989, cost only $2.7 million.

The Registry, however, refused to cave in and said the *Herald* was getting a great deal at $3,134. The Herald disagreed and appealed to the state's public records office to rule on the dispute.

Finally, on Dec. 29, 1992, the public records office determined the Registry price was inflated and ordered the agency to provide the database to the *Herald* at a price reflecting the actual cost. On Jan. 11, 1993, the Registry wrote to the *Herald* and informed the newspaper it was revising the price for the driver registration database.

The final price? $77. We bought it.

In a second case, the *Herald* sought a computer tape copy of the Boston area mass transit authority's payroll.

Initially, the transit authority simply ignored our request. After some prodding, from the *Herald* and the state's public records office, the authority gave us an estimate of $570 to provide the requested information.

The estimate was bogus.

The authority said two employees — a programmer and payroll manager — would have to work six hours to fulfill the request. This wasn't a problem since we realized some programming was necessary.

There was a problem, however, with the hourly rate the authority intended to charge the *Herald* for the work of these two employees. When the hourly rate was calculated for a full year's salary, these two mid-level employees appeared to earn in excess of $90,000.

I called the public records office and said if this was the actual salary of these two employees, the *Herald* would purchase the tape immediately. Imagine what the high-level officials were earning.

However, if this wasn't their actual salary, I informed the record's officials the newspaper wanted a new estimate. Only days later a new estimate arrived. This estimate was $203, less than half of the previous total.

The authority said it had mistakenly included all of the employee fringe benefits and overhead costs in the hourly rate.

The moral of this story? Never trust an estimate from a government agency. Double check all the figures and question anything that appears to be suspect. I've found many agencies back down after they have been caught trying to jack up the price and keep public records private.

CD-ROM may be the future, but don't give up your tape drive yet

By Deirdre Fleming
Missouri School of Journalism
MICAR

When the cost of CD-ROMs comes down in the next five years, government agencies may begin to replace nine-track tapes for storing public information.

Thomas Temin, chief editor of *Government Computer News*, has predicted that by the year 2000 government agencies will replace nine-track tapes with CD-ROMs.

What the metamorphosis means for journalists is that CD-ROM drives will become a necessity to access federal information. And the change could take place swiftly as the cost of CD-ROMs comes down.

Since 1986, Jerry McFaul, a computer scientist with the U.S. Geological Survey (USGS), has helped a number of government agencies convert to CD-ROM information storage and retrieval with the group he formed, SIGCAT (Federal Special Interest Group for CD-ROM Application and Technology.

The lifespan of CD-ROMs has been questioned and SIGCAT, along with CD-ROM manufacturers, have begun testing the longevity of the media.

Mike Martin, a science data systems technologist at the California Institute of Technology's Jet Propulsion Laboratory, reported in the March issue of *CD-ROM Professional* that the idea of CD-ROMs lasting for hundreds of years, error-free, may yet only be a dream.

But Martin concluded in his report that CD-ROM is probably the most durable and long-lived media today.

Still, Robert Deller, a consultant for the government on computer equipment and the director of market research for Selbre Associates Inc. in Bethesda, Md., said that in five years, nine-track tapes will be replaced by erasable laser storage discs, which have been around for the last four years commercially, but are still too expensive for widespread use in government agencies.

Deller added that recently the move in government agencies has been to closed cartridges used in automatic tape libraries. He said many government agencies have turned to automatic tape libraries because the computer-run tape libraries are more efficient than open reel, nine-track tapes.

Ed St. Jean, director of input process systems for the Internal Revenue Service, said that two years ago the IRS launched a pilot project using automatic tape libraries that use cartridge tapes to store data.

St. Jean said that today the IRS uses mostly nine-track tapes because tapes are inexpensive. But, he said after using the automatic tape libraries, the agency expects to implement the new technology by the year 2008 in order to process information more efficiently.

St. Jean said the greatest benefits of the automatic tape library have been the savings in human resources and the decreased risk of damaging the tapes. Because nine-track tapes can be creased or bent, St. Jean said, cartridges are a more reliable method of processing information without altering the data.

"In our environment the task of loading the large number of tapes is too labor intensive," St. Jean said. "We have to mount 2,000 tapes e[illegible] week to load into the master file. If we had on line, we wouldn't need an operator to mount the tapes."

Other government agencies that use the libraries include the Department of Treasury, the Department of Agriculture and the Department of Health and Human Services.

Deller argues that as automatic tape libraries give way to laser discs in the next five years, CD-ROMs will continue to be used by government agencies. A problem that hastens the move to CD-ROMs is that nine-track tapes are sensitive to dust particles and can be creased and damaged.

While nine-track tapes deteriorate eventually, there is doubt that nine-track tapes will ever be extinct.

Temin said that when CD-ROMs or cartridges are commonly used to store public information, the information previously stored on nine-track tapes will remain on the tapes for years before the data is transferred to another means of storage.

"Just like punch cards still exist out there somewhere, nine-track tapes will be around for a long time, probably 20 years," Temin said.

Deller said nine-track tapes will still be used in governmental agencies to a small degree to move data because they are inexpensive, portable and they can hold a large amount of data.

St. Jean also agreed that nine-track tapes will not be replaced entirely, although they will used less by government agencies. He said t[illegible] IRS will continue to use nine-track tapes to share information with agencies that still only use nine-track tapes.

Bringing computer-assisted reporting to every reporter's repertoire

By Tom Foster
The Post-Standard
Syracuse, N.Y.

For every successful computer-assisted reporting effort chronicled in *Uplink*, there probably is an unfinished story languishing in newsrooms around the country.

Reporters using computers for the first time jump in with enthusiasm. Then reality sets in. They often find themselves struggling to tame the mountain of data they obtained on 9-track tape or (Ugh!) key-punched themselves.

These once-energetic souls now stare blankly at their screens. They are frustrated, puzzled and intent on avoiding eye-contact with editors. Two weeks later, there's still no story in the paper.

Computer-assisted reporting often becomes the domain of a few reporters and editors who work on projects. That's unfortunate.

Ideally, newspapers should be able to use computers to assist in investigations AND improve stories reporters write on a daily basis.

Chris Feola outlined the *Waterbury* (Conn.) *Republican-American*'s approach to democratizing the process in the March issue of Uplink. Feola's bottom line: Our reporters can get their hands on the data they need in seconds. And so they use it.

We've had a similar experience at *The Post-Standard* in Syracuse. By concentrating on making data access as easy as possible, we've been able to make use of our databases for stories in nearly every day's paper.

For the most part, this involves helping reporters on deadline locate sources, check facts and add detail. A reporter scrambling to cover a late-night fatal fire, for example, can count on being able to quickly find a list of neighbors to call, identify the building's owner and its assessed value, and obtain a demographic profile of the neighborhood's housing values, rents and income levels from census data.

The most popular databases we have acquired are telephone directories on CD-ROM. Hardly an hour goes by without a reporter or editor using one of these databases. We use a NYNEX product which is updated monthly and a national directory called ProPhone, produced by ProCD Inc. The latest version of ProPhone allows searching by address, providing instant access to a national crisscross directory. It's an inexpensive gem.

Because the telephone book is such a familiar tool to reporters, the transition to the electronic directory is painless for most. From there, it's a small step to use other databases. Other heavily used in-house databases include voter registration records, city and county assessment rolls, government payrolls, campaign contribution records for federal, state and local candidates and 1990 census data.

The secret to this system's acceptance is its simplicity. No training is required to use any of these databases. The entire process is menu-driven. A shell program controls the DOS environment, with "databases" being one choice.

All of the databases, with the exception of those on CD-ROM, are accessed using programs written in Paradox's application language, or PAL. All of the directory changes, query structures, report parameters and other necessities usually required to tap database resources are invisible to users.

All a reporter has to do is keep making choices from menus, then type in the name, address or value he or she is searching for.

Writing these customized programs isn't as complicated as it might sound. All of the applications at *The Post-Standard* were produced by reporters and editors with no prior programming experience.

The databases enhance our newspaper every day. The surprise is that this approach may have helped us produce better investigative projects. If we had set out to devote all our resources to long-term computer-assisted stories, we would have drawn on the talents of only a few reporters and editors.

By having the entire staff familiar with our databases, we benefit from the story ideas they generate when they use the system. When a reporter gets an idea, he or she learns how to use interactive Paradox to do the analysis. Since the reporter already is familiar with the data, the learning curve is not as steep.

Consequently, we have a long and growing list of computer-assisted stories that we have been able to turn around fairly efficiently. Examples include:

- Detailing faults in prison health care by tracking the treatment of more than 1,100 inmates who had died in the state's care in the past decade;
- Analyzing salary differences between men and women in county government;
- Dissecting a public agency's telephone records to help show that its top executive was doing his job from California;
- Uncovering an obscure tax exemption the city was granting developers by designating new parking garages as bomb shelters.
- Writing a feature on people with leap year birthdays. Not every computer story has to be an investigative project. This staple of Feb. 29 editions everywhere was painless because the voter registration roll provided the reporter with a list of names of people born on leap day, complete with phone numbers.

All of those stories involved different reporters, and several other members of the staff are being drawn by increments into computer-assisted, long-term stories.

We've had our share of glitches and failures. But so far every reporter who has been involved with a computer-assisted story has survived. A few are even brazen enough to look their editors in the eye.

By having the entire staff familiar with our databases, we benefit from the story ideas they generate when they use the system. Consequently, we have a long and growing list of computer-assisted stories that we have been able to turn around fairly efficiently.

The National Law Journal uncovers injustice in environmental cleanup

By Kelly Devine
Missouri School of Journalism
MICAR

It's no surprise that wealth and race have long separated neighborhoods and communities nestled in the American landscape. It is surprising, however, that the U.S. government has allowed waste and pollution to fester in minority communities, becoming a deadly element of that injustice.

Children suffer birth defects. Mothers and fathers battle cancer. Brothers, sisters, aunts, uncles complain of pink eye and asthma, ringworm and respiratory problems. This is life in one of the many minority neighborhoods that are not receiving equal protection under environmental law. The Environmental Protection Agency has promoted a racial divide in the United States by acting faster to clean up toxic waste sites in white communities and punishing polluters in white neighborhoods more severely.

These are some of the startling discoveries *The National Law Journal* made when reporters Marianne Lavelle, Marcia Coyle and Claudia MaClachlan embarked on an eight-month project combining computer-assisted analysis of thousands of incidents with in-depth portraits of families who live each day in these minority communities that have become hazardous pollution dumping grounds. The article "Unequal Protection: The Racial Divide in Environmental Law" not only unveiled the bias of environmental law in the U.S. government, but also earned the NLJ an IRE Award for its accomplishments.

It was not easy. The road to obtaining documentation on diskette from the government was paved with denials, rejections, and non-cooperation.

"We knew that there was so much data that we would need to do the project by computer," Lavelle said.

The reporters first difficulty was obtaining computer data on Superfund locations. Superfund is a complex federal program for identifying and cleaning up the worst of the nation's hazardous waste sites. The original diskette the NLJ received from the EPA for $200 was inadequate and incomplete. The data did not include when cleanup occurred and what happened after the identification of a Superfund site.

At first, the EPA denied that such data even existed. But eventually officials conceded that the information could be obtained through each EPA regional office. In all, 10 FOIA requests were filed with the 10 regional offices.

"They responded very slowly, and each field office had a different response, from outright refusal to wanting to charge us $500 for the information," Lavelle said.

Fortunately, one regional EPA officer was very cooperative and agreed to produce a national Superfund diskette for the NLJ reporter at no charge.

"It wasn't through FOIA requests but through reporting skills and a reasonable EPA officer that we were able to obtain the data we needed," Lavelle explained. "We used the FOIA process to open another door."

Reporting skills helped to open still another door when the NLJ reporters attempted to obtain data from the Census Bureau. Because the bureau sells its data to individual contractors, retrieving the data can be very expensive, Lavelle explained. A contractor wanted to charge them $10,000 for the information.

"Luckily, through some reporting we were able to locate a non-profit organization, the Population Reference Bureau, whose purpose is to make census data public," Lavelle said. "They merged our EPA data with their census data and gave it to us on a computer disk we could use on our personal computers for $2,000.

Again, the NLJ reporters were denied civil case data from the EPA. However, through a low[illegible]el judge, the NLJ reporters located and retriev[illegible]ta on every civil case filed since 1972 on two com[illegible]puter disks through the Department of Commerce and the National Technical Information Service.

After obtaining the various data, the reporters built two databases titled "Enforgraph" and "Supergraph," linking demographic data with EPA's enforcement results and with EPA's Superfund progress, respectively. They collected and entered data on diskette using their Magnavox 386 SX-16 personal computer and the word-processing programs XYWrite and WordPerfect.

The reporters also corrected nearly 200 ZIP codes from EPA's records by calling corporate headquarters, post offices, city halls and EPA regional offices, as well as interviewing agency engineers to fill in the various gaps in the EPA's Superfund data. They continued to manipulate and massage the data on their personal computers. Then they conducted nearly 500 interviews for their report, including many minority families living in these highly polluted communities.

Overall, the project took eight months. Five months were spent gathering and analyzing the computer data and researching environmental injustice, two months were spent reporting in communities and one month was reserved for writing and editing the article. "My advice to reporters wanting to get involved in computer-assisted reporting is to jump into a project," Lavelle, who had never worked on a computer-assisted project before. "In my opinion, you need to use computers to learn how to do it. Only by using your hands can you really learn."

"It wasn't through FOIA requests but through reporting skills and a reasonable EPA officer that we were able to obtain the data we needed," reporter Marianne Lavelle explained.

Tech tips

Turning print-image files into databases

By Brant Houston
The Hartford Courant

In the land of data garbage one of the most painful files to get from a government agency is a print file.

This file looks just like the information you get on a print-out and contains header information that looks like "page one, judicial files of america, section 203AB, the office of Eddie Haskell, 10/20/92" and so forth for each "page" of information.

If you can't convince the agency to give you a nice, flat file then it's laundry time.

You can clean these files, if they are small, by throwing them into a word processing program like XYWrite III Plus and manually defining the offending headers and deleting them.

This won't work very effectively for a 10 megabyte file, however, and that is where, whether you are a programmer or not, you should head for a program.

At *The Courant*, I discovered that a good person in our publishing systems had a handy XYWrite III Plus program that with a little modification would go in and destroy headers until it reached the bottom of the file.

All you have to do is identify a mark or string of characters that establishes the beginning of the header — "page" for example — and the end of the header — "92" perhaps — and run it through its loops. Sometimes you might have to make another pass through the data, but the job gets done.

I know the same kind of program can be written more gracefully in other word-processing programs or database programs, but the message of this tip is not to waste time on manual deletion when a little programming can help.

Also, I have been told that a database program — Monarch — takes care of this problem during importing and look forward to reading about it while trying to find money to buy it.

(Editor's note: We knew Tom Boyer in Virginia has some experience with Monarch. Here's what he has to say.)

By Tom Boyer
The Virginian-Pilot and Ledger-Star

Somebody in our data-processing department recently tipped me off to a piece of software that can digest a report print-image file and turn it into a live database. It's called Monarch, and I think its a gem.

Monarch is like a database report writer in reverse. It takes fixed-format reports — even very complex ones — and slices them into data fields. You can export to a bunch of formats, including .wks spreadsheet, .dbf database, or delimited ASCII.

For reporters, it helps with those agencies who aren't set up to output fixed or delimited files to tape, but who send their key data out in long printed reports. If you have a good FOIA, you can pay them to produce a tape file for you, or you can FOIA their print file if it's what you need. It's usually a minor programming change for them to filch a text file — such as a payroll list — from the mainframe's print queue and output it to 9-track tape or floppy disk.

Monarch requires no programming, and you can figure it out in a morning. The new version handles report files up to 30,000 pages. It is very fast on a 486. (No, I'm not getting a commission.)

There are three catches: 1) It must be a fixed-format report with fields of information appearing at the same spot within a line. 2) If it's outputted to tape, you'll need a utility to move the entire file en masse to your PC hard drive. 3) When you get the file in your PC, you may have to clean out weird print characters, extra line feed commands, etc. (More on that below.)

Monarch works by recognizing the various lines of a report by their contents. Say you have a print-image of a report on criminal incidents (this is a report we get weekly). Each page contains headers and footer, and within the headers are three "detail lines" of information describing each crime.

```
CRIME    DATE    TIME     ADDRESS          PROPERTY TAKEN
OWNER    BUILDING TYPE  ENTRY  METHOD
SUSPECT          SUSPECT OCCUPATION
```

The information, in a fixed format, looks like this:

```
BURGLARY    042193   1330     421 APPLE LANE      STEREO
HOUSTON BRANT    CONDO          REAR DOOR         KICKED IN
SCHMID JON              GRADUATE STUDENT
```

Working with this file in Monarch, you first teach it to recognize, or "trap" the first detail line by the pattern of characters, spaces and numbers. Once it knows the first detail line, it can pick the two lines immediately below it. Then you "paint" each bit of information with the cursor and give it a name. Monarch then chops up the report into fields according to your instructions. It can even include header and footer information as fields. Monarch has some analysis features of its own, but I generally have it export to .dbf files and do the analysis in FoxPro.

Now for the hairy part. Sometimes the print-image files contain mainframe junk: weird IBM printer instructions, extra line feeds and such. If Monarch chokes on these — you'll know because the report won't look right on your screen — you need to remove the offending characters.

I won't go into detail, but you can use the hex dump feature of NineTrack Express to diagnose the characters that are giving you problems and write a routine to remove them or replace them with spaces. A word processor like XYWrite can clean up a print file with its search-and-replace capability, but its very slow with big files (editor's note: Brant Houston says the newest version of XYWrite is fast even with files up to 20 meg). I use the text editor that comes with FoxPro 2.0, which can search and replace the entire ASCII character set and is blazing fast with big files.

All this sounds complicated, but once you have it figured out, it need not take long. You can run a complete Monarch translation from a DOS batch file, and it can spit out thousands of records in a couple of minutes. In one case I've gotten the whole routine — from tape, through Monarch, to FoxPro's report writer, to an ASCII report that runs in the paper — down to 10 minutes.

Monarch is put out by Personics of Wilmington, Mass. Call them at (800) 445-3311 and they'll send you a demo disk free. The price is currently $395, and you may be able to get it cheaper from discount software houses.

Bits, bytes and nibbles

You might have thought your April issue of *Uplink* got lost in the mail, but from now until the end of the summer, we'll be publishing on a bimonthly basis. That's because our student help drops off over the summer months.

One-year subscriptions to *Uplink* still cost $20. We'll try to pack more stories and tips into each issue.

Current subscriptions will be extended to include 12 full issues.

Please remember, in order for *Uplink* to be the "forum for computer-assisted reporting," we need your success stories, your problems, your ideas and insights into computer-assisted reporting. Write to:

Uplink
120 Neff Hall
University of Missouri
Columbia, MO 65211

• • • • • •

Tom Foster, projects editor at the *Syracuse Post-Standard*, wants to share ideas with other reporters or editors who have developed Paradox applications.

The article on page three of this issue of *Uplink* describes how Tom has successfully made data available to reporters in the newsroom, in part through programs written in Paradox's application language.

If you share Tom's penchant for Paradox, contact him at (315) 470-3071 or on Compuserve 70401,1732.

• • • • •

Jon Schmid will be leaving MICAR next month to work as an assistant database editor for the Raleigh, N.C., *News & Observer*.

Jon has been handling much of MICAR's role with the professional community while pursuing a master's degree in journalism since Elliot Jaspin left the Institute last June.

Matt Reavy, a doctoral student in journalism, will take over Jon's role as editor of Uplink and will help Andy Scott, executive director for Investigative Reporters & Editors, teach the professional seminars. The next TRI-DART seminar is scheduled for May 23-28.

Both Andy and Matt will handle requests for data analysis and data sales until a new MICAR director is appointed.

- **Uplink** becomes bimonthly
- An invitation for Paradox users
- Jon Schmid moves south for a new job

THE MISSOURI INSTITUTE FOR
COMPUTER-ASSISTED REPORTING

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