Uplink

THE FORUM FOR COMPUTER-ASSISTED REPORTING

January 1993
Volume 4, Number 1

Read this before scanning text

Optical character reader turns analysis into 4-month nightmare

By Barb Pearson
USA Today

A look at how Congress spends its money was *USA Today*'s first experience with scanning. Perhaps it was the size of the project (7,500 pages) or the number of records (325,000) or the complexity of the subject — whatever, it turned out to be a nightmare.

USA Today and Gannett News Service wanted to look at what it costs to run Congress and to identify, among other things, the top spenders. We started by asking Congressional leaders for a computer tape of all 1991 expenses by House members' offices, committees and support departments. They refused, saying all the information was publicly available in four books kept by the Clerk of the House. At least one person acknowledged that providing the information on tape, rather than paper, would make analysis easier.

In order to proceed with the project, we chose to scan the four books into the computer. The plan was to have the tabular data scanned and converted into fixed-position ASCII files, which we could then read into Paradox.

We quickly dismissed buying a desktop scanner and doing the work ourselves because of the volume, and the slowness and inferiority of portable models. Instead, we hired a company that estimated scanning on its professional machine would take four weeks. Try four months.

The company used the Xerox Imaging Systems K5200, a Kurzweil hardware and software system priced at $14,500. Stacks of pages are automatically fed into the scanner, much like a high-volume copying machine.

Among our problems:

- On many pages, the Government Printing Office had not printed the type square on the page. The software has a tolerance of about 5 percent skew, anything more than that requires an operator to manually place the page in the machine. This is a very time-consuming process.
- The book's agate type stymied the splitter. The splitter software is designed to separate touching characters, making it possible to "read" them. If it splits improperly, all sorts of strange things can happen. For example, the letter O could be read as a C. The splitter has problems with small and bold text and with characters that are angled or italic. Well, the House books are all of that.
- Any stray or unwanted marks on a page, such as ink smudges and page numbers, introduced errors.

The result: Lines of garbled or wrong characters: 7's appeared as 1's, 8's as 3's. (Some scanning errors were comical — shamed employee, instead of shared; resistant executive, instead of assistant.) Data from two lines often converged into one. Plus, the tabular structure of the original data was often lost, creating errors later when we imported the fixed file into Paradox.

It took months of extensive work to clean up the ASCII files. It was crucial that every number be double-checked with the original. And, in order to do any collapsing of data and analysis, we needed to correct any errors in staff names, job titles and government contractors so they were consistent.

Although it's hard to estimate the overall error rate caused by scanning, on some pages more than a fourth of the characters or numbers were wrong.

Before you proceed with a scanning

• • • • •
continued on page two

Date	Voucher No.	Payee	Service dates
LBJ INTERNS, MEMBERS CLERK HIRE AND OFFICIAL EXPENSES OF M: OFFICE OF THE HON. DAVID O'B MARTIN—Con.			
EXPENSES			
10-08	1270460005	CARY R. BRICK	09/20/91-09/22.
10-08	1270460006	Do	09/20/91-09/22.
10-08	1275660023	AQUA COOL	08/31/91
10-08	1275660002	FEDERAL EXPRESS CORP.	08/28/91
10-08	1275660001	FORT COVINGTON SUN	09/01/91-09/01.
10-08	1276300002	AMERICAN INTERNATIONAL	08/29/91-09/03.
10-08	1276300003	Do	09/13/91-09/15.
10-08	1276300005	AT&T INFORMATION SYSTEMS	08/10/91-09/09.
10-08	1276300004	Do	08/12/91-09/11.
10-08	1276300001	BELL ATLANTIC MOBILE SYSTEMS	08/04/91-09/04.
10-11	1282650003	AMERICAN INTERNATIONAL	09/19/91-09/22.
10-11	1282650002	Do	09/26/91-09/29.
10-15	1283670014	RINGAMERICA, INC.	09/01/91
10-22	1291750001	TELEPHONE ANSWERING SERVICE OF WATERTOWN	10/01/91
10-22	1291750002	THOMAS J LANKFORD	09/16/91-09/24.

House Clerk records, actual size. **Reading all these numbers would be daunting to anyone but virtually impossible for an optical scanning machine.**

Tech tips

Working around a software bug

Solving some problems with delimited files in XDB

By Elliot Jaspin
Cox Newspapers

Anyone who has worked with software for any length of time will soon encounter a bug. It may be large or small, but once you hit the enter key and watch a program sail off into cyberspace, it can be awfully irritating.

My personal favorite involved software for aircraft navigation. The software worked perfectly until the plane flew across the equator. Then the onboard computer commanded the plane to fly upside down. Hmmmmm.

While every programmer will try to produce software with as few bugs as possible, the chances of producing a perfect piece of code seems remote. The trick for the user is to find an acceptable "workaround" when a problem is encountered.

A particularly irritating bug in some versions of XDB involves importing delimited files. Each field in a delimited file is separated by a character, usually a comma, and text fields are enclosed within quotation marks. Thus, a record might appear as:

"Doe, John", "123 Maple St.", "Bristow, VT"

Although there are four commas in our example, software should be able to see this as a record with only three fields, because two commas are within quotation marks. When the software encounters the first set of quotation marks, it should know that anything that follows is data. If it encounters a comma it should not treat it as a delimiter. Once it reaches the terminating set of quotation marks, the software should consider the comma that follows as the end of one field and the beginning of another.

Alas, versions of XDB prior to release 2.41 see the quotation marks but keep on looking for the comma as a delimiter. The results are usually commastrophic (**ouch**).

While the problem is potentially fatal, the workaround is simple. XDB allows the user to specify what the field delimiter is when importing a delimited file. By the same token NineTrack Express allows the user to pick any character that can be entered at a keyboard as a delimiter. Instead of using a comma, transfer a file in NineTrack Express using a "^" or a "|", both characters rarely or ever are found in a file. Once you have transferred the file from tape, specify the character in XDB, and the file should import perfectly.

A few closing notes, this particular b XDB was corrected in version 2.41. But a similar bug has popped up in the database program from Microsoft Access. The software gets confused when it encounters a record such as:

"Bartowski, "Buddy"", "123 Maple St.", "Bristow, VT"

As I said at the start, software bugs are everywhere, like. . .

continued from page one

• • • • •

project carefully consider the quality of your originals. Is the type clear, preferably not agate, smudge-free, and printed straight on pages? Testing is crucial. You should have several dozen, representative pages scanned and then doublecheck the electronic data with the original. How many errors were introduced and how difficult will correction be? Is scanning worth it if the cleanup is extensive?

In researching this article, our contractor told me that scanning technology has improved dramatically, even since we began our project a year ago. We hope so. The qua the conversion we got was so poor that, h we known at the time, we wouldn't have go the scanning route. Instead we'd have cut back on the data we wanted from the House books and had it double-keyed.

Series leaves politicians with no one to blame for recession but themselves

By Chris Feola
The Republican-American
Waterbury, Conn.

Like too much of what happens around here, it all started when a bureaucrat said something stupid.

In front of a room packed with businessmen and journalists, a Connecticut Department of Economic Development official stood up and laid all the blame for the state's lingering recession on the shoulders of the banking industry.

The credit crunch, he said, had taken $2 billion away from the private sector of Connecticut's economy. Do you know what that does to an economy? he asked.

We thought that was a good question, especially because the government he worked for had taken just over $2 billion out of the economy in record tax hikes.

If taking money out of the economy one way was bad, wouldn't the same hold true for taking it any other way?

This was not some obscure academic question. Connecticut was hemorrhaging jobs and wracked by the recession and defense cutbacks. The state lost more than 180,000 jobs — almost 12 percent of its non-farm workforce, which it reached in the late '80s.

So the question of what had gone wrong with the economy — and how to fix it — was of vital importance to our readers, especially since every seat in the state Legislature, all six spots in the U.S. House of Representatives, one U.S. Senate seat and the White House were all up for grabs in the November election.

We decided to limit our project to the effect taxes have on the economy. We made that decision for a couple of reasons: It kept the project manageable, and taxes are a hot issue in Connecticut. Not only had the state seen two record tax increases in three years, but legislators had just passed the state's first income tax.

So we sat down and made a list. All we needed to do, we figured, was crunch 10 years of federal budgets; 10 years of budgets from three states; 10 years of budgets from 22 towns; and 10 years of budgets from 16 school districts.

Then we needed to find a dozen families, persuade them to let us look at their finances, then crunch their budgets to see how much they were paying in taxes and link it all together with the government budgets to show where their money was going.

OK. The project took six months. *The Republican-American* has been doing computer-assisted reporting since 1988, so the resources to handle such big piles of data were on hand. The paper has a series of 386's with big hard drives, plus a flat-bed scanner hooked to a very good optical character recognition program.

We acquired a filing cabinet full of data available only on paper, and 25 megabytes directly from government computers via modem. Since this was primarily a numbers project, we moved everything into a spreadsheet — Borland's Quattro Pro for Windows.

We used Quattro to build 3-D spreadsheets with each topic — the breakout on federal spending, for example — in a separate sheet that cranked its results into the master sheet. This kept the master sheet simple enough to be read by person other than Werner Von Braun.

"The Tax on Living" ran every day for more than two weeks, ending the day before the election. It showed that the average Connecticut family was paying a little over 50 percent of their income in taxes and how those taxes made it financially attractive for businesses such as UPS to pay to move their operations to other states.

It showed that 42 percent of the cost of a bottle of liquor and more than 50 cents on every gallon of gasoline went for taxes. It showed taxes where people expected — on gas — and where no one would have ever guessed — one on gas pump nozzles and another for pump gauges.

It also showed the money was being spent less than wisely. Not only did Congress increase its budget 17 percent last year, the money provided for things such as the House and Senate photo studios and radio and television studios.

And then there's overhead. We obtained government figures that showed more than $300 billion went for overhead — rents, telephone and utility bills and airfare, among other things. (That's the total for the entire federal budget, not a separate appropriation.) That doesn't include a dime for programs or salaries.

Perhaps the biggest fuss occurred after we obtained an internal Congressional document that revealed the ugly truth about the Social Security Trust Fund: There isn't one.

All Social Security taxes go straight into the treasury, where they are spent just like the rest of the taxes. The money is replaced with "IOUs," the document went on to say.

Sen. Joseph Lieberman, D-Connecticut, called the system a sham, a fraud, and said it amounted to nothing more than flim-flam.

You always wonder if a project like this has any effect. We know it was read. We were deluged with calls and letters, and many readers took the time to fill out and mail us a do-it-yourself total tax form we printed.

The key to this series was putting families and faces on the numbers. It is one thing to explain that X amount of dollars are going out in taxes. It is another to show photos of the family paying thousands in taxes and have them explain they can't afford allergy shots for their child.

"The Tax on Living" showed that the average Connecticut family was paying a little over 50 percent of their income in taxes.

Projects at the San Francisco Examiner and the Lawrence Eagle-Tribune uncover (what else?) political waste and greed

Some of our far-flung friends recently wrote us with details about their computer-assisted reporting projects.

From Dick Rogers
San Francisco Examiner

It began with a telephone call in February to Lisa M. Krieger, the *Examiner*'s medical writer. The caller, a health professional in town, described what he believed was waste and incompetence in the use of city, state and federal funds. The task of evaluating the 100-plus AIDS programs carried out in San Francisco was beyond our resources, but it turned out that the city was about to conduct individual evaluations of every publicly funded grant program.

The process took months. It wasn't until October that we were able to go through each file folder. There were 124 — each one containing scores on at least 11 performance categories, as well as five or more other entries and written comments. None of it was computerized.

I had been doing some computer-assisted reporting over the last two years, most of it on my home PC. In September, the paper acquired a Compaq 486 with a 200-megabyte hard drive, FoxPro database software and a spreadsheet. So the AIDS story was my first official opportunity to demonstrate the number-crunching virtues of the computer.

Some database wizards won't bother with a computer story unless the information is available on disk or tape, but I saw this as a perfect argument for the build-your-own-database approach. Besides, it was a way for me to start getting familiar with FoxPro.

We entered the key information into the database, adding a few fields for certain calculations we had in mind. It took about five hours. Instead of sitting down with a calculator and two big notebooks worth of numbers to look for patterns, averages and totals, we let the computer do the work. The Compaq is a screamer — the typical sorting process was completed by the time I looked up from the keyboard to the screen.

If we sorted incorrectly or failed to include an important category, it didn't matter. We just did it again. Compared to the drudgery of doing all the work by hand, the five hours spent keying the numbers was time well-spent.

The result was a tw-day series built around our analysis of the city evaluations. The punch line: San Francisco has created a far-flung system of AIDS programs that often fall short of their goals and duplicate efforts, wasting money that could be used to improve services. The stories created quite a stir among AIDS groups and provided us with a follow-up story — a vow by AIDS officials to straighten out the system.

If it weren't for the computer, we'd still be fiddling with the calculator and the two big notebooks of numbers.

From Brad Goldstein and Ed Achorn
The Eagle-Tribune
Lawrence, Mass.

For two years, reporters at *The Eagle-Tribune* have been using computers to explore the link between contributors and political candidates.

Last fall, the computer put us on the trail of perhaps the biggest story of the campaign in a local congressional race. A local man, a former congressman, was running for the 5th district again, touting his record as a successful businessman.

When we downloaded his federal election records and sorted on business, we quickly discovered some patterns: members of New investment firms and their wives were plowing tens of thousands of dollars into his Massachusetts campaign. One of the contributors was Robert E. Brennan, who was accused of bilking small investors of millions of dollars during the 1980s.

Further research disclosed that the individual investors and their firms had links to Brennan. Our check with state securities officials revealed that many of them were under investigation in several states on securities fraud charges.

A check of reports filed by the candidate's business ties with the Securities and Exchange Commission uncovered other interesting sidelights. Brennan's First Jersey Securities had sold penny stock in the candidates's company during the '80s; a second firm whose members contributed heavily to his campaign was selling his company's stock now.

The stock sales helped the company provide a $700,000 retirement package for the candidate, although the company had had three years of losses totaling $2.8 million.

We were able to give readers a broader perspective on the candidate's career as a self-made businessman.

The Eagle-Tribune broke the story on the front page of its Sunday paper. The *Boston Herald* followed the next day with information supplied by the candidates' opponent.

Please send reports of your latest computer-assisted reporting projects to Uplink, 120 Neff Hall, Columbia, Mo., 65211.

Bits, bytes and nibbles

■ Here's the latest on upcoming seminars

MICAR's March seminar is already full. But MICAR isn't the only wheel in town when it comes to training journalists in computer-assisted reporting. Here's a list of some upcoming seminars. If we've missed any, please send the information to Uplink.

The Education Writers Association will present "Computer Power in Education Reporting" Feb. 26 in Arlington, Va.

The one-day workshop co-sponsored by The Freedom Forum is advertised as a "day packed with story ideas, tips on getting started and ways to solve problems in your newsroom."

Guest speakers will include Pulitzer Prize winner Elliot Jaspin of Cox Newspapers, Aleta Watson from the *San Jose Mercury News*, Joseph Garcia from the *Dallas Morning News* and Pat Ordovensky from *USA Today*.

The conference is free, but space is limited to 40 participants. For information call Lisa Walker at (202) 429-9680.

• • • • • •

Indiana University's National Institute for Advanced Reporting presents its fourth annual conference, "Computers Equal Power Journalism," March 12-14 in Indianapolis.

Session leaders will show participants how to build computer-assisted projects from the ground floor up. Some of the session topics include: Methodology of Working with Statistics; Dealing with Dirty Data: Ethical Problems; and Management of Computer Resources: In-house Training.

Speakers include Pulitzer Prize winning investigative reporters Don Barlett and Jim Steele; Dwight Morris, editor of special projects at the *Los Angeles Times* Washington, D.C., bureau; Don Fry, writing coach of The Poynter Institute; and Pat Stith, reporter at the *Raleigh News & Observer*, Raliegh, N.C.

The seminar costs $135. For information call Deb Perkins at (317) 274-2776 or write to the Conference on Computer-Assisted Journalism, I.U. School of Journalism at Indianapolis, 902 W. New York Street, ES 4104, Indianapolis, IN 46202-5154.

• • • • • •

The Transactional Records Access Clearinghouse will hold the TRAC Workshop on Nuclear Regulatory Commission Data May 20-23 at Syracuse University in New York.

The workshop will teach individuals how to use NRC data as the basis for investigations.

Attendance to the workshop will be limited to 30 persons. The cost is $200 for registration and $180 for lodging and meals.

For information contact Randi Maroney, TRAC, 478 Newhouse II, Syracuse University, Syracuse, NY 13244, (315) 443-3563.

The University of North Carolina at Chapel Hill will present a "Workshop in Precision Journalism: Computer-Assisted Reporting for the Next Century," July 12-16 in Chapel Hill, North Carolina.

Topics include the use of statistics for journalists, an overview of statistical software packages, the graphic representation of data and an introduction to the data holdings of major social science archives. The class will also present on-line search strategies for full-text databases such as VU/TEXT, NEXIS and DataTimes.

Phil Meyer, author of *Precision Journalism*, will be one of the instructors. Other teachers include sociologists and librarians from the university.

The seminar costs $750 but early registration, before May 15, is $650. For more information, call Dr. Beverly Wiggins at (919) 966-2350 or write to the Workshop in Precision Journalism, Institute for Research in Social Science, University of North Carolina, Chapel Hill, NC 27599-3355.

• • • • • •

American University will offer "Journalists, Computers and the Federal Government: A Mid-Career Seminar for Reporters and Editors" July 18-23 in Washington, D.C.

The seminar promises hands-on experience working with databases maintained by various government agencies as well as commercial on-line services. Ethics, access and statistical concepts will also be covered at the American University seminar.

While prior computer experience is not required, the seminar will be limited to 10 to 12 journalists who have at least two years Washington reporting or editing experience.

The cost is $650 for early registration before March 1 and $750 after March 1. For more information call Wendell Cochran at (202) 885-2002 or write to the American University School of Communication, 4400 Massachusetts Ave. N.W., Washington, DC 20016.

• • • • • •

Investigative Reporters and Editors and *The News & Observer* will hold "Computing: The News Frontier" October 21-24 in Raleigh, North Carolina.

The seminar promises to put "bytes into your bark" through hands-on training in computer-assisted reporting. Topics include: finding and negotiating for government databases; exploring local, state and federal on-line databases; and using digitized maps.

Participants will return with a budget and plan for starting a computer-assisted reporting program at their own news organization.

For information write to Dan Woods, *The News & Observer*, P.O. Box 191, Raleigh, NC 27602.

POSITIONS OPEN
MISSOURI SCHOOL OF JOURNALISM

The world's first School of Journalism is seeking to fill faculty positions for the Fall 1993 school term. Located in Columbia, ranked by MONEY magazine as the second most livable city in the country, the School offers BJ, MA and PhD degrees and has five sequences which combine a strong professional training program with an active research program. The Missouri School of Journalism is accepting applications for the following positions:

Assistant Professor, full-time, tenure track or renewable contract appointment, to teach computer-assisted reporting and coordinate outreach activities to the profession. Significant media experience and advanced degree preferred. Appointment at higher rank possible for candidate with extensive background. Professor Sandra Scott, Search Committee Chair, School of Journalism, Box 838, University of Missouri, Columbia, Mo. 65205.

Assistant Professor or Instructor, full-time, tenure track or renewable contract appointment, to be teaching editor for the city desk of the *Missourian*, the community daily published by the School. Significant professional experience, creativity and interest in teaching required. MA degree preferred. Computer-assisted reporting skills a plus. Send cover letter, curriculum vitae and list of 3 references to Professor Yves Colon, Search Committee Chair, School of Journalism, Box 838, University of Missouri, Columbia, Mo. 65205.

Screening begins February 1st and will continue until positions are filled.

AN AFFIRMATIVE ACTION/EQUAL OPPORTUNITY EMPLOYER.
WOMEN AND MINORITIES ARE ENCOURAGED

THE MISSOURI INSTITUTE FOR
COMPUTER-ASSISTED REPORTING
120 Neff Hall
University of Missouri
Columbia, Mo. 65211
(314) 882-0684