The power of building your own databases

Sarah Cohen, The Washington Post, cohensh@washpost.com
Anthony DeBarros, USA Today, adebarros@usatoday.com
David Heath, The Seattle Times, dheath@seattletimes.com

Some of the best reporting in the country comes from homemade databases. As dreary as it sounds, reporters are logging documents, collating notes and recording details. The payoff comes when the analyzed data produces compelling or even shocking findings. Homemade databases often lead to stories that otherwise would go unreported.

Amassing records from diverse sources:

- David Heath amassed documents and bits from other datasets into a database showing that some of the most powerful doctors at the Fred Hutchinson Cancer Research Center had major financial stakes in biotech and drug companies. This database helped document the conflicts of interest at the Hutch, and was part of the Goldsmith and Polk-winning investigation, *Uninformed Consent*, for the Seattle Times. (http://seattletimes.nwsource.com/uninformed_consent/)

- While investigating police brutality, reporters at The Philadelphia Inquirer found that claims were held in five different offices. Putting together the paper records and cross-referencing involved recording up to 77 facts about each case. The database amassed records from many different documents, eventually leading Rose Ciotta and Nancy Phillips to focus on "nickel rides" -- taking handcuffed suspects on harrowing rides in the back of police wagons, sometimes leading to serious injuries. Their findings are detailed in the 2001 series, *Battered Cargo.*

Typing data from paper forms for analysis:

Sometimes, government agencies try to block the release of data by charging exorbitant fees or by simply stalling. Other times, the data isn't computerized or is difficult to extract from a specialized system. In some states, government agencies aren't compelled to provide databases when they exist, but can supply only printouts instead.

Reporters and news organizations often thwart the attempts to keep data secret by typing in the records themselves.

- In Peoria, Ill, Susan Okeson built a database of housing court cases from index cards held in a secretary's notebook. She typed in a little at a time, but eventually was able to show that large property owners were punished less severely than small-time property owners for similar code violations.

- David Herzog, then at the Allentown Morning Call, inventoried the weapons in police property rooms around his circulation area to see whether a proposed ban on assault weapons squared with guns that cops actually saw on the streets.

- Two reporters at the Fort Wayne Journal Gazette in Indiana created a database to follow 10 years' of homicides from incident through sentencing. The database, much like earlier efforts in the Los Angeles Times, The Washington Post and the Chicago Tribune, among others,

tracked information on the age and race of victims, the officers involved, prosecutors, charges and sentences. Niki Kelly and Karen Balsey found that black killers were sentenced to longer terms than whites charged and convicted for similar crimes.

- Tom Loftus of the Courier-Journal in Louisville, Ky., built his own database of gubernatorial campaign contributions long before the state computerized those records. As a result, he reported countless stories where state contracts and government appointments were given almost exclusively to campaign contributors. What's more, he was able to show that one company used its employees as fronts to exceed the limits in campaign finance laws by making illegal contributions. This reporting led to indictments.

Other examples include local campaign contributions and police maltreatment claims, including police dogbites.

Content analysis and document management:

- As part of the team reporting The District's Lost Children, Sarah Cohen of The Washington Post created a database summarizing thousands of pages of heavily redacted documents to identify children, assign culpability and analyze known failures. The database was crucial in that the content analysis of the government's own documents showed that top officials knew children were dying for the same, predictable reasons and did nothing to correct the failures. The series won this year's Pulitzer Prize for investigative reporting and the IRE medal. http://www.washingtonpost.com/wp-dyn/metro/dc/government/lostchildren/

- In Oklahoma City, Ziva Branstetter built a database to keep track of recently open files of juvenile crime cases. She tracked information on each case and the outcome as well as several fields with socio-economic factors such as whether the family lived in poverty, whether the kid finished school and whether the kid had one or more parents in prison. It allowed her to get a clearer picture of the environment the kid was raised in.

- Mark Skertic of the Chicago Sun-Times collected details of tread-separation lawsuits around the country months before the Ford Explorer and Firestone stories hit the national consciousness. The database resulted in an April 2000 series detailing 43 death and dozens of serious crashes related to tires that came apart. (http://www.suntimes.com/tires/).

Systematically recording reporters' notes:

- In the months following the World Trade Center terrorist attacks in September, Anthony DeBarros -- with colleagues at USA Today -- found a way to methodically record notes from dozens of reporters and other sources on each fatality. The newspaper found that few people below a crucial floor died, while almost everyone above that level perished. (http://tribute.usatoday.com/search/default.asp)

- Andy Lehren at Dateline NBC tracked the results of a survey of meat "sell-by" dates conducted by reporters in a recent hidden camera investigation. Reporters entered a store, picked up a package of meat, and imprinted the sell-by date in the styrofoam packaging. They would return the next day, search for their imprinted packages and check. Many -- more than 200 examples across the country at most major chains -- were re-wrapped and re-stamped. (http://www.msnbc.com/news/753195.asp)

- In Charlotte, Ted Mellnik created a database used by 16 reporters and researchers to document racetrack deaths for *Death at the Track,* a 2001 series documenting the deaths of 260 drivers, staff, fans and even reporters at racetracks. Starting with newspaper and other press reports, the paper used the database to document which deaths were verified. It also recorded important details consistently, like cause of death, what the victim was doing at the track, and summaries of the cases for publication. By keeping the thumbnails in the database, the reporters could keep reporting and writing up to the last minute, moving the copy out of the database on deadline. "There were so many people researching this -- we used the Web pages to manage the process of information," Mellnik says. (http://www.charlotte.com/mld/charlotte/news/special_packages/death_at_the_track/)

- Dan Keating, now of the Washington Post, created similar databases while at the Miami Herald during the 1998 investigation of vote fraud that won the 1999 Pulitzer Prize for investigative reporting. Reporters, who were interviewing hundreds of voters who may have participated in fraud, recorded structured questions in the database. It helped Keating identify geographic and voting patterns that were most likely to result in fraudulent votes, such as the name of a witness on an absentee ballot or the frequency of voting in prior elections.

.... And lessons learned about building it yourself

Among the tips passed along by these and other reporters:

- Stay flexible. One of the great things about building your own database is that you can often go back and change what you want to do. By staying flexible, you don't always have to think of and record every possible detail you'll want to use.

- Type consistently. If you want to record heart attack deaths, be sure you record a cause of death each time as "heart attack" or "myocardial infarction", not once each. For people building the database, use as many lookup tables or pick-lists as practical.

- Try sampling your records first. If you're visiting courthouses, try typing in 10 or so records from each courthouse. Then try to do some analysis on the results. You'll find lots of things you'll want to change.

- On a team project, insist that everyone use the database. Half-filled databases are useless for many projects. Make sure your editors know that.

- Remember that most stories can handle one, maybe two, strong numbers. Spend your time figuring out the strongest and most compelling result of the database; avoid reporting out each finding.