COLUMBIA JOURNALISM REVIEW
Strong Press, Strong Democracy

Reports — November / December 2010

Serious Fun With Numbers

We're drowning in data, but few reporters know how to use them

By Janet Paskin

The story was already great, even before Daniel Gilbert opened his first spreadsheet. Thousands of citizens in the southern Virginia area Gilbert covered for the *Bristol Herald Courier* (daily circulation: 30,000) had leased their mineral rights to oil and gas companies in exchange for royalties. Twenty years later, they alleged, the companies had not paid, adding up to potentially millions of dollars owed. As Gilbert learned, the complaint was complicated. It involved esoteric oil and gas practices and regulations, a virtually unknown state oversight agency, the rules of escrow accounts—and finally, some very angry people and a handful of very big companies. With these facts alone, he could have written a stellar story giving voice to citizens' complaints, and shining a light on a little-known regulatory agency. That, in many newsrooms, would have been plenty.

But Gilbert, who officially covered the courts for the paper, wasn't satisfied simply to raise the specter of noncompliance. Whenever a well produced natural gas, the energy company was supposed to make a monthly payment into a corresponding escrow account. These payment schedules were public. So were the production records. All Gilbert had to do was match the production records with the payment schedules to see who had—and had not—been paid.

Easier said than done. Gilbert requested the information he needed and received spreadsheets with thousands of rows of information. In Excel, a typical computer monitor displays less than a hundred rows and ten wide columns. Gilbert's data was much too massive to cram into this relatively modest template. So he started with one month's worth of information, using the program's "find" function to match wells and their corresponding accounts. One by one. Control-f, control-f, control-f. It was tedious and time-consuming. There was a story there, he was certain. But control-f would not find it.

What would you do? Could you navigate, process, and make sense of thousands of rows of data? If you have not yet had to ask yourself this question, there is no time like the present.

Most journalists are just like Gilbert, with daily computer skills that include Internet searches, word processing, and maybe some basic calculations in Excel, none of which enables journalists to truly mine large collections of data. Meanwhile, the amount of raw data available to journalists has mushroomed. At the federal level, the Obama administration's "open government" initiative has given rise to new sources like Data.gov, a website devoted to the aggregation and easy dissemination of national data sets. State and local governments have followed suit, making much of the data they collect available online. More elusive tranches of

data have been pried loose by nonprofit organizations courtesy of the Freedom of Information Act; an inquisitive journalist can download them in minutes. “I’m constantly amazed and surprised about what’s out there,” said Thomas Hargrove, a national correspondent for Scripps-Howard News Service who often leads data-based research projects for the chain’s fourteen newspapers and nine television stations.

Against this backdrop, the ability to find, manipulate, and analyze data has become increasingly important, not only for teams of investigative journalists, but for beat reporters. It is hard to conceive of a beat that doesn’t generate data—even arts reporters evaluate budgets and have access to nonprofit organizations’ tax returns. What’s more, because the universe of data is vast and growing, and the stories that use it are rare, data-based journalism has become a powerful way to stand out in the crowded news cycle. “When you acquire a certain level of data skills and literacy, you can punch way above your weight,” says Derek Willis, a web developer at *The New York Times* and author of the computer-assisted reporting blog, The Scoop. “Simply put, you can do things others can’t.” And last but certainly not least, readers *like* data. They like charts and interactive graphics and searchable databases. At The Texas Tribune, which has published more than three dozen interactive databases and usually adds or updates one a week on average, the data sets account for 75 percent of the site’s overall traffic.

Of course, news-gathering organizations have to some degree understood the value and power of data for more than twenty years. Bill Dedman’s 1989 Pulitzer-winning investigation into the racist lending practices of Atlanta banks relied heavily on database reporting and was widely seen as a validation of computer geeks in the newsroom.

But even after many organizations hired computer-assisted reporting specialists, using data for stories has usually been limited to big investigations and projects. And with good reason: years ago, data-driven stories were almost prohibitively inefficient to write. A reporter had to identify what data he needed and which agency collected them; it often took a FOIA request to secure the data, which tended to arrive in sheaves of dot-matrix-printed paper. It was then up to the reporters to build their databases—by hand.

That’s not the case anymore. Agencies maintain and disseminate their data electronically. While there are still plenty of data sets that require diligence, persistence, and FOIA requests, many can be accessed without even speaking with a human being. And in the newsroom, every reporter has a spreadsheet program like Excel or can find one for free online. The logjam, these days, has more to do with reporters’ and editors’ interests and aptitudes—with their capacity for number-crunching—than it does with technology.

At the Bristol paper, Gilbert clearly needed help. His editor, Todd Foster, had been Gilbert’s champion and mentor on the story thus far, but he knew little about managing thousands of rows of data. Neither did anyone else in the newsroom. Gilbert, however, knew who did: Investigative Reporters and Editors. For years, this journalism nonprofit has been running computer-assisted reporting workshops, called Boot Camps, on the University of Missouri campus in Columbia and around the country. At the six-day workshop, Gilbert would learn how to use spreadsheets and a more sophisticated database management program—the two fundamental tools he needed to manipulate the data he had. The only issue was getting Foster to say yes.

That was hardly a slam dunk. Of course, Foster wanted Gilbert to nail down the story. But as one of seven reporters on staff at the *Herald Courier*, Gilbert typically generated three or four stories a week. His colleagues would have to scramble to fill the hole during his absence. Then there was the cost. The *Herald Courier* and its parent company, Media General, were suffering the same economic hardships as the rest of the newspaper industry. In 2009, Media General mandated fifteen furlough days for most of its 4,700-plus employees, equivalent to a 5.8 percent pay cut. Sending Gilbert to Missouri, in this climate, was not an easy sell: tuition for the workshop was $560, plus travel to and from Columbia, lodging, and meals for a week. The total came to around $1,240, and the reporter would need to use his vacation days to attend.

Still, a potentially important story and six months of work hung in the balance. That weekend, Foster called on the paper's publisher at home, with a few cans of Red Bull and a bottle of vodka in hand. They covered a variety of business issues, and "at the end of the night, I sprung the Boot Camp on him," Foster recalls. "He said, 'Is it worth it?' I said, 'It's worth it. And in April, it might really be worth it.' " Soon Gilbert was on his way to Missouri.

Foster never told Gilbert they expected him to win a Pulitzer for their trouble—at least not in so many words. But the reporter understood that the expectations were high. "They didn't send me there saying, 'Go have fun,' " he notes. "It was more like, 'This better be worth it.' I felt a good deal of pressure to make it count."

This is a fairly standard expectation. Most newsrooms assume that journalists will immediately put their new skills into practice. When Reuters recently sent six beat reporters to one of the ire Boot Camps, they were all required to pitch a story to work on while they attended the session. "We want to see the stories," said Claudia Parsons, Reuters' deputy enterprise editor for the Americas. "That will be the test."

At the same time, making database skills and training a priority can be tough for overburdened reporters and editors. Nor do journalism schools necessarily give such skills pride of place—in fact, many teach them piecemeal, if at all. At the graduate level, New York University requires students in its Science, Health, and Environmental Reporting (SHERP) concentration to obtain a solid grounding in numeracy. In other concentrations, however, these skills play a smaller role. The Columbia University Graduate School of Journalism offers a handful of relevant classes, including investigative reporting, a course called Evidence and Inference, and a new addition, Digital Media: Interactive Workshop, which stresses storytelling through data and interactive presentation. But there is no data course that all students must take in order to graduate. "We don't require every student to know how to use Excel in the same way we require them to know how to use FinalCut Pro or a digital camera," said Bill Grueskin, Dean of Academic Affairs at Columbia. As a result, many students remain stuck at control-f.

What Gilbert learned in Missouri turned out to be indispensable. He took his spreadsheets with him, and learned how to transfer the data from Excel to Microsoft Access, a database management program better suited to large searches. (Funnily enough, Gilbert actually had a copy of Access on his desktop back in Bristol; he just didn't know what it was for.) And he absorbed a basic programming language called Structured Query Language, or SQL, which allowed him to search for specific patterns in his data.

Eventually, Gilbert got his data cleaned and organized enough to be able to write his fundamental query: Show me the accounts that correspond to wells where oil or gas has been produced, but royalties have not been paid. What he found was damning. “Of about 750 individual accounts in escrow, between 22 percent and 55 percent received no royalty payments during months when the corresponding wells produced gas over an 18-month period,” Gilbert wrote in the first of an eight-part series. As for royalty payments that *had* been made, $24 million was lying in escrow, in dispute. Over the course of the series, Gilbert explained the history of the dispute, took the state gas and oil board to task, and showed that citizens who were allegedly owed thousands were being told they were entitled to less than a dime. His series spurred the Virginia legislature to investigate ways to distribute the money in escrow to the people who own it. In April, Gilbert won the Pulitzer Prize for Public Service.

After the prize was announced, Foster told Gilbert that the *Herald Courier* had been hearing about the escrow fund and the government mismanagement for years. “Two prior managing editors had spiked the story,” Foster said. “Royalties, methane gas, escrow accounts—it’s not the sexiest story.” In these earlier cases, nobody had been able to break through the data roadblock. Gilbert, who moved to Houston in October to cover the oil and gas industry for *The Wall Street Journal*, says that he thought it was a “pretty good story” to begin with. “But the data changed it,” he adds. “Instead of just asking the question, I was able to answer it.”