Bulletproofing the data project

Sarah Cohen
The New York Times
May 1, 2014

Here are some handouts on avoiding data errors in stories (all from IRE, so you'll need to sign in):

- "Editing the data-driven story," by Maud Beelman of the Dallas Morning News: http://http://ire.org/resource-center/tipsheets/4045/download/?fileid=3950>
- DANGER! Look out for Dirty Data, by Jaimi Dowdell, IRE http://ire.org/resource-center/tipsheets/3999/download/?fileid=3905
- "A Guide to Bulletproofing Your Data," by Jennifer LaFleur, now of the Center for Investigative Reporting: http://ire.org/resource-center/tipsheets/3848/download/?fileid=3745

Here is an email I wrote to Craig Silverman of Poynter in April on the same subject, on what I try to do when possible.

Report out the underlying records

All data comes from some kind of individual record -- it's a survey response, a traffic ticket, or a payment. So I try to track the statistics back to the underlying records and determine how they were collected and how they make their way into some kind of data system. Most still come from some kind of form, which -- even if it is transmitted electronically and in bulk -- still has a detailed description or a manual somewhere. Those forms tell you a lot: How much discretion do people have? What do the instructions say? Is it confusing? Then I talk to people who fill them out and people who process them once they're filled out. I also go through hearings, audits and other criticisms of the system.

Say you're looking at crime stats: what do the underlying incident reports look like? What decisions are cops making, and what are they rewarded or punished for doing that might push close calls one way or another? What gets counted or ignored? What problems have they had in the past in responding to incidents or writing them up? I also look for items that never get into the system at all. In this case, it might be crimes that got downgraded, or people who don't ever call the police (maybe like domestic violence victims who might lose their lease for repeated complaints.)

This is a reporting job, not a data analysis job -- one that too many people skip.

I just did a story on deportations, and ICE only maintains the "most serious" criminal conviction -- It took about 3 weeks to nail down what that meant, who assigns it and how it matters to the agents on the ground.

Data definitions & codes

I try to go through each field in the database and be sure I know what each one actually means, not just what it says, even if it seems irrelevant to what I'm doing. The names and the descriptions of codes are usually some kind of bureaucratic or computer-ese shorthand, and they often aren't what they seem. There's a famous example of this: old versions of the federal contracts database had a field called "obligation type". If it was "A", it meant the amount listed was a payment by the government to the contractor. If it was "B", it meant that there should be a minus sign next to the amount -- it was money the government got back from the contractor. Yikes.

I try out a few sentences using the data and see what might be wrong with them. In the immigration example, a single person could be deported many times, but the agency wouldn't give us anything that would let us track an individual. It meant we could never characterize the data as "people", only "cases."

Integrity checks

I try to find some kind of benchmark I should be able to hit pretty closely (or exactly). In the immigration example, the agency had published a few key statistics from the same data, and once I got the definitions down, I could match them almost exactly. There was a good explanation about why they would have changed since they published the initial figures, and they were only off by about 100 out of 300,000. Most data reporters start laughing when they see a spreadsheet of 65,536 rows - it's the limit of older versions of Excel and usually indicates that dataset isn't complete. (Early in the Wikileaks story, the reporters were stumped because they were missing the later months. The problem was they'd imported it into a spreadsheet with those limits. They realized it quickly, but they might not have if they weren't diligent.)

I also try to do simple frequencies on any field I care about to see how often they're filled out and whether they seem to have any oddities. A lot of times you'll find out no one has ever used the code you think is so interesting, or even that they have redacted the cases with a code you might care about. The Washington Post won a Pulitzer in 1999 for a series of stories about police shootings that originated from a missing code -- the

FBI had been removing justifiable homicide by police (I think it was "77") from its Supplemental Homicide Reports for years, and Jo Craven McGinty fought to get those missing records. The story, of course, was far more than the data, but the data sparked them to look at the problem of police shootings in DC.

I also look for things like missing months, zip codes that don't exist, impossible combinations like 10-year-olds with court records, or codes that abruptly start or stop. When I find these problems, I have to report them out.

Find a rabbi or a sherpa

Others know this data better than I do and I try to convince them to guide me and look over my shoulder. Ideally, there are 3-4 people who you can trust to vet your work along the way. This usually requires convincing them you're worth their time by doing a lot of the steps above and by reading anything they've written on the subject. It's also useful to find any academic literature on your topic and contact any researchers who have touched on it. They often know of projects they would have wanted to do, but couldn't get funded.

The data diary

On longer stories, I keep a diary of the data work three ways, but different people do it differently. The goal is to make the work reproduce-able and to make sure I've got reasons for every decision I had to make along the way.

1. Use computer programs to do as much as possible, heavily commented and versioned. I haven't gotten in the habit of saving my versions on Github, but I should -- it's a great way to be able to backtrack. It's much more difficult to log every mouse click than to to comment computer code to say what problem I was trying to solve or question I was trying to ask, and what problems I ran into doing it other ways. It's really easy to fix and re-run a computer program than it is to re-do mouse clicks when it turns out there was a mistake early on.

2. Keep a log of my interviews on the data. Invariably, you get conflicting answers or never get satisfactory answers to some of the questions. The problem is, usually no one has used the data the way you're using it, so you have to ask them why what you are doing could be wrong -- what traps are in there waiting for you? I do agree to these interviews on background. I know that others don't, but at this point I'd rather know what I might be doing wrong, and I don't really have any firm results yet.

3. Keep another log decisions I made along the way. Again on the immigration story, I had to decide what a "minor" vs. "serious" crime was, and what to do about drug convictions when it didn't distinguish between sale or possession. I try to make every close call err on the side of the least newsworthy answer to avoid hyping results. In this case, drug offenses were counted as "serious" all the time, because the agency claimed it was focusing on "serious" criminals and we didn't want to undercount them when we said that most deportees weren't.

Report out cases

I describe the process as moving from lab to field and back to the lab, reporting out some cases that seem to reflect a bigger pattern. Those often turn into the anecdotes in the story. In some cases, you can look yourself up or someone you know. In others, you can look up cases that have been in the news. In others, you can go to a place and talk to some actors. When I worked on farm subsidies, we called lots and lots of farmers and asked them to go over what we thought we'd found. We also try to find ex-employees of the agency or a company to go over the records with us. But usually, there was one case that sparked our interest in the records in the first place, so we can look up that case and find others like it.

Last year, I worked on a story on police who committed domestic violence. One set of records I used were disciplinary records for one state, which showed a huge drop in 2012. I kept asking the agency why, and what was missing, but the expert on the data said it was complete. Later, I filed a public records request for a recent case that was in the news. That's when I learned that the entire case is secret until it goes to a hearing, which often happens 6-8 months after the complaint. Those were the missing records, but I would have never known if I hadn't been asking about cases. (It also helps with the public records aspect of the issue -- by giving us enough information to report out cases, the agency can be sure that the data isn't misconstrued, and it reduces the agency's workload in explaining things.)

Vetting results

I want to give the agency or the subject of the time to digest what we've found and seek out as many experts and stakeholders as I can to vet the results. There is always a risk that there is some kind of mistake in logic or processing or some other misunderstanding. I want to hear the reasons I might be wrong, and about anything that surprises them. In most of these projects, no one has ever done what I've done, so they can't fact check for me. One thing that has made this really important step a lot harder is that government agencies now restrict who we can talk with so much that we can't get a real experts to answer questions at the front end. That means we have to try to engage them at least at the back end. It's far from ideal, but even worse when the agency just won't let us talk with anyone but the PIO or the agency head, who may not know enough about the details to give us real feedback. We used to be able to engage with the people who were in charge of the data collection, even if it was just on background.

Editing and fact-checking

I write the sections of the story that I'm responsible for when I work on a team, and go through drafts regularly to find places that the records could strengthen a point. This sometimes involves reworking the data so that it's more precisely matching the sentences we want to write -- for instance, "10 years" not "since 2002". (In other words, cut off a year of the data.)

I'm in the editing sessions or getting copies from the editor on a regular basis, as well as coordinating any data with graphics as we go along. One problem is that there are often old versions of results or spreadsheets floating around among editors, reporters and graphics folks, and they sometimes find their way into drafts over and over, after we've made corrections based on our vetting.

For me, fact-checking involves circling every fact -- whether or not it's a number -- and going back into the data and re-doing the analysis that led to that particular fact. Then I'll pull that out into a fact-check file somewhere. At some point during the writing / editing phase, I try to retrace the entire process from start to finish. For instance, in the immigration story, the most time-consuming thing was to combine 11 spreadsheets into one database containing 3 million records. I didn't re-do that step because I had been able to match my benchmarks after that step. But I do re-do any key tables or queries, check the code that created them, and make sure that they've incorporated any of the comments I got along the way. I'm also checking the graphics regularly to make sure they match what the story says and resolve any differences.