CAR on the Beat
NICAR 2012

Dan Catchpole, Yakima Herald-Republic
Contact: catchpole@gmail.com, 206-355-4319
Twitter: @dcatchpole

WHY USE DATA ON YOUR BEAT?

By quantifying an issue, you change it from a matter of opinion to a matter of fact. That changes the conversation.

Never forget the human angle. Data analysis is useful to journalists so long as the data affects or sheds light on people and their world.

Using computer-assisted reporting (or data-driven reporting) for beat stories frees you from relying as much on official or expert sources. It also allows you to expand on story leads (e.g. adding context), find new stories and angles, and allows you to visualize complex stories.

A note from a handout written by Peter Aldhous (a great resource in the CAR community):

“Note: data can be used in both reporting and storytelling. But think carefully about what you need to show to your audience. Some of the best data-driven stories actually contain little in the way of numbers or graphs.”

>> That last point is huge. Remember: CAR is a means to an end. Always ask yourself, how do these data affect people? Why will readers care?

OKAY, HOW DO I DO THAT?

First, join the NICAR listserv, and ask a question when you’re stuck.

To successfully use CAR skills in your beat, you need to think about two elements:

• You (and your newsroom)
• Your beat

YOU

Data Mindset

Always look at your beat (and the world) and ask yourself (and sources and other journalists, too) what is quantifiable? Here’s a Zen-data exercise: Walk down a city block or go for a hike, and ask yourself, what do I see that’s quantifiable? Imagine looking at the world like the character Neo does in The Matrix (He sees it in binary code, we want to see it in terms of sortable data.)

The St. Paul Pioneer Press’ MaryJo Webster offers some ways to help do that:

➢ Think of data as people too. Interview the data. It can give you tips, answer questions, raise questions and mislead you (and confuse you).

➢ Define your story ideas as questions, not statements. Statement: I’m doing a story on unsafe bridges. Question: What percent of bridges in the state are not safe?
➢ **Tune your data radar to pick up opportunities to quantify.** Look for issues, statements and other things that can be quantified.

For an example of tuning your data radar, I had to write a budget story about a rural fire district on my beat. The district fire chief told me that the district needed to hire more full-time firefighters because its number of volunteer firefighters was declining. I did some research, and found that this is a national problem, but I couldn’t find an example of any reporter who had quantified the problem.

So I found the data for each fire district in the state over the past decade. That showed a slight decline in my paper’s county. I did more reporting and found that fire districts were receiving more emergency calls. So fewer volunteers were responding to more calls. I got the call data (which had to be pulled from each district b/c of a historic quirk with 911 dispatch in my paper’s coverage area), and found that calls had gone up by about 20%, which produced a much more dramatic story:

---

**Not so dramatic:**

**Way more dramatic:**

---


---

**Running with scissors**

Beat reporters are usually pressed for time, but rushed data analysis can produce errors. Don’t rush. Don’t run with scissors.

➢ **Know your data:** How was it collected, entered and processed? What do the values represent? (It might seem self-evident, but that doesn’t mean it is.) Talk to the database manager to learn about the data’s provenance.

➢ **Use trusted methodology:** Don’t make it up as you go. Run the data and then run it from scratch again. Have others double check your data. Consult experts (stat profs and teachers, data journalists, NICAR-L) about your methodology.

➢ **False precision:** Beware of your results, especially when venturing into new ground. Analysis can produce results that are precise but not really correct or don’t match the ground truth, so turn to notebook reporting to verify. For example, Washington state’s Employment Security Division calculates industry sector employment by rounding to the nearest 100. So if a county’s agriculture jobs go from 900 to 1,000, did the county gain 100 jobs? Not necessarily. It might have really gone from 949 to 951 jobs.

➢ **Correlation vs Causality:** Just because two variables move in relation to each other does not mean that one causes the other, or that if one moves, the other will definitely react in a
particular way. As journalists, we don’t have the time (or academic training) to do the rigorous analysis it takes to establish causality. Also, there might be other factors at play influencing the variables. So, we can typically only say that there is an observed trend, that A tends to move in relation to B. (For most readers, that is more than enough.)
  - For example, I analyzed pitching data from 1921 to 2011. I measured the average number of strikeouts per nine innings and the average number of complete games thrown by starting pitchers, and found that the two have a Pearson correlation coefficient (or Pearson r) of -0.96. That is a super strong relationship for something affected by so many variables, like pitching. The results of a Pearson r test can only be -1 to 1. So, yeah, -0.96 is incredibly strong. But the most I can say is that there is a strong correlation between the two. I can’t prove that the increase in strikeouts led to fewer complete games being thrown.

[http://public.tableausoftware.com/views/Baseballsdissappearingcompletegame/Disappearingcompletegames]

Triangulate your data

There often is not a dataset out there that answers the specific question you have in mind. But you can often answer it by data triangulation – essentially using two datasets that answer related questions to answer your question.

When you’re hiking (without GPS) and need to find your position with a compass and map, you triangulate your position. You take a compass reading off one landmark, and draw a straight line with that on the map. Then you do the same thing with a second landmark. Where the two lines meet on the map is where you are.

For example, no one is collecting data quantifying if there are enough volunteer firefighters in Yakima County, but by combining two datasets (the number of volunteers and call volumes), I showed that there are about 33 percent fewer volunteers per 100 emergency calls today as compared to 10 years ago.

Or for another example, no agency is probably collecting data to count how many sex offenders work in summer camps in your coverage area. But state agencies have lists of registered sex offenders and of licensed camp staff. By triangulating the two, you can find out how many are working with kids every summer.

Organization

- **Newsroom**: Communicate early with your editors, graphics desk and others. Don’t tell them you have a chart for your story late in the afternoon. If you make your editors’ lives easier and make them look good, they will give you more time to do data stories.
- **Calendar, lists (project and to-do)**: Excel is a great tool for tracking your projects. I keep a Current Project Excel doc in which each project has its own worksheet.
- **Data diary**: Carefully chart what you’ve done with the data. It is critical when you’re covering a beat, and have to pick things up and put them down because of your daily demands.
  - E.g. Steps for cleaning addresses. (Note: There are better ways to clean addresses, e.g. OpenRefine.):
    1. Excel: Ran CLEAN function on addresses in a new column to take out not printable characters. That solved issue of cell contents being on multiple lines.
    2. Copy pasted into Notepad++
    3. Did Find/Replace w/ RegEx to break main address and additional info. FIND: (^.*)([0-9]{5})(.*$) REPLACE: \1\2,\3
    4. Pasted Addresses back into Notepad++.
5. Separated apartment numbers starting w/ “#” from address. Ran F/R: FIND: (^\+.+)(#.+)(YAKIMA.+$) REPLACE: \1\3,\2
6. Pasted (now comma delimited) results into Excel. Copied address column (w/o # part) and pasted into Notepad++.
7. Separated City names (which I should have done before separating apartment numbers). Did FIND/REPLACE with: (^.+)(\sAVE\s)(.+)(YAKIMA.+$) and \1\2,\4,\3
8. I ran it for: RD, AVE, ST, BLVD, WAY, DR, PL, LN, Av, St, Bl, LOOP
9. Pasted results into Excel. Using Filter to display addresses w/o Yakima in them.
10. Separated apt numbers that are only numerical characters and don’t have “#” from address

Tools

First, learn Excel. It has been around since the 1980s. It’ll probably be around for a few more years. (In fact, one paper is still using an Excel tutorial written in 1991 by Rich Gordon, who is now a journalism professor at Northwestern University.) Open it every day. Use it for your personal finances. Get comfortable with it (and pivot tables).

Let your needs dictate what you tools you learn next. Oftentimes, several tools perform similar functions. Figure out which one you like best, and get good at it.

YOUR BEAT

Finding stories:

1. **Proactive:** Mining data for stories
2. **Reactive:** Using data to expand on stories.

Finding data for your beat

Find databases relevant to your beat, and mine them regularly. There are too many beats to give a comprehensive list, so here are some examples of useful sources:

- **General:** Census.gov, FedStats, data.gov, FAA, CDC, National Institute of Health
- **Google searches:** “data”, “database”; Advanced Search >> Search websites by document type (.xls, .pdf, .doc) to find data and documents.
- **School data:** DataQualityCampaign.org (Gates Foundation), your state’s department of education, federal Department of Education. Get school district’s database of school employees’ salaries, benefits, stipends and any other compensation. (If they’re technically state employees, then you can get this data for every district from the state.)

Have a question or want suggestions specific to your beat? Don’t hesitate to contact me, and I’ll do my best to help.

Dan Catchpole, Yakima Herald-Republic
Contact: catchpole@gmail.com, 206-355-4319
Twitter: @dcatchpole