FINDING DATA

Finding data simply requires some good-old-fashioned reporting: ask a lot of questions. It's the same as hunting for a human source. If it falls within your beat, ask sources you already know, perhaps they can point you in the right direction. These same sources will also likely help you figure out what questions to ask the data

These days some stuff is readily available on the Internet, but don't assume that something will be there – or even that it will be complete or as broad as you need it to be.

If it's something you haven't covered before, here are some tips:

Look for what others have written on the topic

Look for reports from govt agencies or online data on the topic (even if it's national in scope or pertains to a different place…it can still provide clues)

Take an educated guess about what government office might be responsible for the data, call them up and ask for help

Getting "summarized" data from a govt agency is usually easy – and many times the employees you deal with will simply assume that's what you want because that's what most journalists ask for. However, is the summarized data sufficient for your needs? Most of the time, the answer will be no.

Let's say you're writing about crime in the city and the police department offers to give you data showing crimes broken down by type (murder, rape, robbery, car theft, etc), going back a number of years. This will allow you to ask questions like – which type of crime has increased the most over time? Which has decreased the most? But it won't allow you to ask things like: are certain types of crimes more common in one part of the city than others? When do the crimes most often occur (night? Weekends?)? How many of the city's murders are "random" versus committed by someone who knew their victim?

If you asked for the raw data, with one record for each crime, you would be able to get fields with all these details and you can summarize or aggregate in whatever fashion you want.

Key thing to know going into an analysis: what does each record in my dataset represent? You'll have the most flexibility if you can get data that represents the lowest level possible.

Asking for data:

Being able to speak at least a little bit of their "language" and convince them that you know your way around spreadsheet or database software will go a long way in making it easier for you to not only get the data, but also get help working with the data once you have it.

Who you get the data from in a particular government agency depends entirely on how the department is run. Your ideal situation would be to deal directly with the computer people who maintain the database and know what all the fields mean. But that is often hard to do. Some agencies require that you make your request via the media or public relations person. Others have a dedicated public records/FOIA person. Other times you simply have to go to the head of the department. If you already have a source in the agency that you've been dealing with, start with them. If they can't help you, ask them who can.

Rule of thumb....sometimes you will ask someone in an office if a database exists and they might tell you "I don't think so." Don't assume that means "no." Ask them to refer you to someone else in the office who might know the answer. Keep moving to other individuals until you get a definitive answer.

Once you find someone who knows something about the database, do some "reporting" on the database itself. Some things you want to ask about in advance:

What kind of software is the data stored in now? (the answer will likely be a program you've never heard about; probably something designed specifically for this agency or for this type of recordkeeping. Occasionally it will be "SQL Server" or "Oracle" or something you've heard about. If it's a very simple dataset, it might even be in Excel or Access)

Do you have a standard process for exporting data for other public requests or for making reports? Or would you have to create a new export process?

What fields of information does it contain? And can I get a copy of the record layout in advance?

Are there any fields of information that would be considered private/non-public (that would be redacted)?

Are there any fields that are coded that would require a codesheet? Can I also get a copy of that?

What do you use the database for? Are there any reports that are routinely generated based on this data? Do you do any analysis? (this question is good for finding out whether materials exist that you might be able to use to double-check your own findings or to get some basic summary analysis, if that's all you need)

How does the information get into the database? Does it start from a paper record and get key-punched? Or does someone put the information directly into the database? (for example, inspectors in various agencies these days have started carrying laptop computers or mobile devices and typing up the results of their inspections directly into a program that is then uploaded to the main database when he/she returns to the office. In the past, inspectors filled out paper forms and a clerk would type the results in...often leading to data-entry errors)

What does each record represent? (one incident? One inspection?). This is important to know because sometimes it’s not what you’d expect. For example, I worked with a fire department’s response data. Each record represented a piece of equipment sent to the scene. So there could easily be multiple records for each incident.

If there's something specific you are trying to get from the data, be sure to ask if it includes that. You don't want to get the data and discover the key piece you need is not there. (see story below about trouble getting foster care data)

Once you gather this information, you should have a pretty good idea of whether the data is going to serve your needs and you will be in a better position to make a request that will be less time-consuming for the agency and less costly for you.

SOME OTHER THINGS TO CONSIDER:

When you're requesting data you might have to make some decisions early in the game that will be critical to what you're able to do with the data.

The big one is time frame. Do you want the data to be a single snapshot in time (i.e. the 2010 census headcount for each county) or do you want it to represent a time range (i.e. all crimes that occurred between Jan 1, 2011 and Dec. 31, 2013)?

Snapshot in time versus time range is sometimes dictated by the data itself. For example, census data is always a snapshot. But you can get multiple snapshots and be able to make comparisons over time.

Also sometimes you will need multiple years of data to get a big enough chunk of data to be able to say anything definitive. For example, let's say you want to see if certain judges are more lenient (or more harsh) in sentencing for drug crimes, you will likely need many years' of sentencing data and only include judges that have heard a significant number of drug cases (and you need to figure out what that minimum needs to be).

Generally in the news business we like to be able to say if something has changed over time, so having data crossing a fairly large period of time is most useful. But the big question is how much time? Two years? Five years? 10 years?

My rule of thumb is that five years is good, 10 years is great. But sometimes you can't get 10 years of data (although I think this will change in the near future). I frequently run into situations where the keepers of the data only have a few years at their fingertips, while the older stuff is locked away somewhere as a result of a database overhaul or something like that. You might be able to get the older stuff, but it might cost you more money or take more time.

Also think about the topic you're writing about. For example, if I'm writing about income or employment or housing --- topics that were seriously affected by the recent recession — I would want to be sure I had some data that showed the "before" period. So that would dictate the minimum time frame I'd want to include.

You might also have to make other decisions about what to include or not include. Hopefully the people providing you the data will help you navigate this, but it's good for you to ask lots of questions.

The big thing is that you need to understand --- before you get the data — exactly what the data encompasses. The specific questions you need to ask and the various hurdles you're going to encounter are going to be very different from one dataset to another. Here are a few examples to give you an idea:

Example 1: You want enrollment data, broken down by gender and some other factors, for the University of Minnesota. Will that be for the entire UM system? Or will it just be the Twin Cities' campus?

Example 2: You're requesting parking ticket data for the city of St. Paul. But Ramsey County actually keeps the data. Will the data include just tickets written in St. Paul or will it include everything in Ramsey County?

Example 3: You're requesting data on gun-related incidents in the city. The police will tell you that all the incident data is coded by type of incident, but that there are multiple codes that could fall under this vague definition of gun-related. So you'll need to decide which incident codes to include. Then you'll also need to decide, do I get data on all calls that were made to the police — even if the cops didn't write up a report about it? Or should I only get the ones that resulted in the police writing a report or at least looking into it?

Example 4: You want to compare high schools to see which ones have the best/worst graduation rates. Graduation rate data is easy to download off the state education department's website, but there are tons of schools on there I've never heard of. There are ones labeled "alternative learning center" or "juvenile center" or "alternative", etc. You'll see that these schools don't have very many students. They are specialty schools geared toward dropouts or disabled students or students with criminal problems. Do you include them? That's usually a question best answered by an expert on the topic.

Sometimes you need to make decisions like this just because of the time or cost involved. Perhaps you want to get budget data for every city, but it would be too time consuming to do that for all the cities in your news organization's coverage area because you'd have to collect the information from each one separately. So then do you just do the big ones?

Sometimes you have to find the line you're going to draw and just be prepared to defend your reason. For example, at USA TODAY, I wanted to do figure out whether it was true that schools in rich neighborhoods won more state championships in sports than those in poorer neighborhoods. To do this, we had to collect state championship results from every state separately (most were in PDFs). We ended up limiting it to 13 states and only those sports that are commonly found in nearly every school.

By MaryJo Webster
@MaryJoWebster
Mjwebster71@gmail.com