Roundtable data discussion during the Data Journalism session

Enter a “data set of mind.” Always assume relevant data is out there, even from non-environmental sources when working on an environmental story. Example: If EPA can’t tell you about Superfund problems following a natural disaster, Coast Guard data may offer insights.

Get agency forms. If you know forms, you know their data sets.

When you FOIA for data, also ask for the data dictionary or data key. You don’t want to assume you know what the fields mean.

Do your best to talk with an internal source who understands what everything within the data means so you can learn about its strengths, weaknesses and vocabulary. Just say: *I have the data. I don’t want to misrepresent anything.* You want the PIO to understand that a background conversation with the keeper of the data is in everyone’s interest.

Agency gives you data in a PDF? Don’t be shy to call back and ask for it in a machine-readable format, such as csv or xls. If the PDF has a table full of data, that data is almost certainly from a spreadsheet or another dataset that can export to a spreadsheet-friendly format.

If the PIO says, “We only have PDFs,” or “Our database can’t export,” or “Our dataset includes non-public information, so you can’t have any of it,” politely but persistently ask for a background discussion with either their data person or their IT person. The agency almost certainly does not only have PDFs, because the data came from somewhere. The agency’s database almost certain can export, and if it truly can’t because it’s incredibly old, that’s a story you might want to write -- because they’re keeping important information in a pretty useless way. And it’s easy to exclude certain columns with Social Security numbers or other sensitive information -- here’s an example of how to do it with SQL, a common database language: https://www.sqlservercentral.com/Forums/Topic667861-338-1.aspx. The PIO may not understand this, but the data or IT person does.

When working with data, talk to experts, agencies and/or companies early on. They know if there are issues that aren’t visible in the data. Never rely on the dataset alone without reporting it out. The information can mean something other than it seems like it does. What do the headers really mean, for instance? What are the caveats because of the way the information is collected? Talk to people who deeply understand the subject. Does the data reflect reality? If not, why?

If the agency/group producing the data issues reports about the data, read them all. It’s a good place to learn about shortcomings/opportunities about the data.

Having problems getting data out of an agency? Contact a journalism organization such as the Reporters Committee for Freedom of the Press, for advice/help. Appeal FOIA denials. Consider partnering with another newsroom to get more people on the case.

Some data sources:

Investigative Reporters and Editors' data library: https://www.ire.org/nicar/database-library/

PublicData.google.com

Advanced Search on Google: Remember, you can search by filetype. .PDF .XLS .CVS etc.

EPA has a lot of datasets, including the Toxics Release Inventory (air, water, land pollution); Greenhouse Gas Reporting Program; and the Emissions & Generation Resource Integrated Database, or eGRID (power-plant data).

U.S. Energy Information Administration has a lot of data that can help environment & energy reporters. Call PIO Jonathan Cogan if you're not sure how to find specific data, or whether the dataset you're looking at is the best one for your story.

U.S. Bureau of Labor Statistics: How many jobs does a specific industry have? How many of those jobs are in your state or region? How has that changed over time? Any of these questions might be useful context if you're writing about coal, wind, solar, manufacturing, environmental services, etc. Both the BLS and the Bureau of Economic Analysis also have wage data of various sorts.

U.S. Geological Survey. Datasets range from climate change to water to invasive plants to the locations and causes of grizzly bear mortality.

Global Forest Watch: Remote-sensing data.

ESRI, the mapping-software firm, collects mapping data.

Consider using data from scientific publications. You need to clear with researchers and journals that publish it. But be mindful that you want to be cautious. You only want to use good data. How big is the sample size, for instance?

Some trade groups collect data that could be relevant for the beat. The Outdoor Industry Association, etc. Just keep in mind that an interest group collects data to further their interests, so you want to understand the collection methods and potential shortcomings. (You want to know that about any dataset, really.) Surveys are a good example: The questions may be written to try to get a specific answer.

U.S. Census Bureau. Demographic data.

Remember, data can have flaws. Vital to ground truth when using a data set to tell a story. Treat data like a human source -- don't assume it's right just because it's full of numbers.

Visualization: Maps, timelines, bar charts, etc. If you're just trying to show numbers in a simple graph, a bar chart is probably the go-to. Start the graph at zero or explain why you're not. (Better than not starting at zero, if you're dealing with a large number getting larger: Show the change, e.g. 4 percent in 2014, 8 percent in 2015, etc., or 5,000 additional in 2014, 21,000 additional in 2015.)

Here's a memo that Lisa Song put together on the basics of data journalism: https://docs.google.com/document/d/1NA6F7u0WTt1IvY8wu8tubk1kAn_p9Rt5U5QScW5_pQk/edit

The data is never the story. It's a good backbone. But stories are about people (or, say, other species and people, or plants and people). People want to read about people.

Questions when looking at pollution data: Who are the biggest emitters? How has that changed over time? Who does not need to disclose due to loopholes in reporting? How does that impact your region?