Acquiring Records:
Dealing With Data

Society of Professional Journalists
and the Sigma Delta Chi Foundation
David Cuillier, Freedom of Information Committee
May 2011

Contents

About the Trainer: David Cuillier, Ph.D., is a member of the SPJ Freedom of Information Committee and has been an SPJ newsroom trainer since 2005. He is an associate professor of journalism at the University of Arizona in Tucson, where he teaches public affairs reporting, computer-assisted reporting and access to public records. Before entering academia he was a public affairs reporter and city editor for a dozen years at daily newspapers in the Pacific Northwest. He is co-author with Charles Davis of *The Art of Access: Strategies for Acquiring Public Records* and researches the psychology of access. He provides news and tips about FOI at www.theartofaccess.com and http://blogs.spjnetwork.org/foi/, and he can be reached at cuillier@email.arizona.edu.

Data Lingo

Terms to Know When Requesting Data:

- **Database.** A database is a collection of information, usually displayed in a table laid out in a grid - like what you would see in a spreadsheet program like Excel.
- **Data.** This is a plural noun, so you would say “Let’s see what the data tell us.” If it looks awkward in your writing, simply write around it (substitute “database”).
- **Field.** The fields in a database are the columns of information going up and down. Typically in a database it would be a category of information that is collected for all records, so in a database of pet licenses you might have a field labeled “Name” for pet name. Make sure to know exactly what “fields” are in a database that you request.
- **Row.** Each row in a database - the information that goes across from left to right, represents a different record. So in the case of pet licenses, a row would represent an individual licensed pet, like Max the black lab. Make sure to know how many “rows” of data, or records, are included in a database you request.
- **Record layout.** This is a description of the fields contained in a database, and the type of field it is. For example of Date of Birth field would be a “date” field. A Name field would be a text field. And a Salary field might be a numeric field - something you could calculate. Always ask for a record layout so you know what's in a database.
- **Code sheet.** A code sheet, sometimes called a data dictionary, provides a key to any codes in the data. For example, sometimes in a Gender field a 1 might represent females and a 2 males. You need the code sheet to interpret those numbers. Always ask for it. Sometimes they call it a data dictionary. Or just ask for a list to translate any codes in the data.
- **Tab-delimited text file.** This is the most easily transferable format of data because any software program can import the data. When requesting data, ask for it first in a simple format that you have software for, such as Excel. But if the agency can’t provide it in those formats, it should be able to export it in a tab-delimited text file. You can also ask for it as a comma-delimited text five, also called a comma-separated text file (csv file). These are much easier to work with than “fixed-format data,” which requires you to do more work when importing into Excel or Access. Just say “tab-delimited text file” and they will assume you know what you are doing.

Overcoming Data Denials

If the agency denies your request, check with experts to see if the agency has a right to keep the data secret. Here are some common agency denials and how to respond:

IF THE AGENCY SAYS…

"WE DON'T HAVE A DATABASE LIKE THAT"

Often when you ask a clerk or PIO about data they won't know what is feasible or available. You might even have them say they don't have it in the computer and then offer to print out paper copies for you – from their computer. You can try to call or make an appointment with the official processing your request to get more help. Meet with the techies and have them show you the data they keep.

If you are reasonably certain the data you've requested do exist, and if your request letter was clear and informative, you should try to do more research. Are there news reports, congressional hearings or court records that describe the information you want more clearly? Rewrite your request, giving the agency more guidelines and clues for where they might find it. Try to be as patient and understanding as you can; some agencies are short staffed or have disorganized data systems.

"SOME OF THE INFORMATION IS EXEMPT FROM DISCLOSURE, SO WE WON'T GIVE YOU ANY OF IT"

The agency can't withhold an entire data file because some portion(s) of it is exempt from disclosure. The agency must release any non-exempt material that can be reasonably extracted from the exempt portion(s). It's usually pretty easy to delete a field (such as social security number). It's a little more time-consuming to read through notes/memo fields that have narrative text. But it's no different than redacting a paper record.

"WE CAN'T GIVE IT TO YOU BECAUSE AN EXEMPTION SAYS WE HAVE TO KEEP IT SECRET"

FOIA exemptions are generally discretionary, not mandatory – an agency is not required to withhold all information. Agency officials can choose to waive the exemptions and release the material, unless another statute specifically restricts that disclosure. One exception is FERPA, but note that FERPA doesn't cover everything and they can release the records if identifying information of a student is removed.

"OUR PROPRIETARY SOFTWARE DOESN'T ALLOW US TO COPY DATA"

I don't know of any software that can't copy or export data. Maybe it exists, but it must be rare. Usually the person saying that is unfamiliar with

the software and needs to confer with the agency computer technicians. If, after talking to their techies, they still stick to that story, find out the software maker and call them up. No doubt the company will want everyone to know how useful and versatile the software is and explain how to copy the data.

"COPYING THE DATA WITH FIELDS REDACTED WOULD CREATE A NEW RECORD, AND WE ARE NOT REQUIRED BY LAW TO DO THAT"

It is true that most laws do not require government agencies to create new records, only let you see or copy existing records. But you aren't asking for a *new* record – just a copy of their existing records with some information redacted. It's no different from getting a copy of a paper file with some information (fields) redacted with a black pen. Just because they blot out a name on a piece of paper doesn't mean it's a new record. The same theory applies to data. Copying data with some fields redacted, or even combining fields from different databases, is not creating a new record. It's copying existing data.

"OK, OK. HERE IS YOUR DATABASE. THAT WILL BE $1 MILLION, PLEASE."

Make them justify their programming expenses with a line-by-line explanation. You might be able to narrow your request to get a simpler database that would still serve your purposes. Arm yourself with what other agencies charge for data copies, including for computer programming time. If many other agencies charge nothing or very little, then make that known, including by writing a story about it.

Look at the agency's FOI logs to find out if others have been receiving the same data. Ask for an extra copy. Ask for a backup copy of their data if they make backups. Get an outside expert to scrutinize their time estimates. The estimates are usually inflated and unreasonable.

"WE ONLY PROVIDE THIS INFORMATION TO RESEARCHERS. YOU CAN HAVE IT IF YOU SIGN THIS CONTRACT WITH US."

Some agencies give information to researchers provided they sign a contract with use restrictions, such as prohibiting identification of individuals in the data. Few reporters are willing to sign such agreements. The problem is you might want to use the information for something else later and won't be able to. Also, fundamentally it designates journalists as above average citizens with special access, and it creates a new category of "public information." Either it's public or not. Some journalists advise only considering such agreements when the information is clearly not public but the agency is willing to release it for your story.

"WE DON'T KNOW WHY YOU WANT IT OR HOW YOU MIGHT USE IT. YOU MIGHT USE IT IN A WAY WE DON'T LIKE."

Tough noogies. In most states a data request cannot be denied based on who the requester is or how the information will be used (except in the

case of commercial mailing lists in some states). If they ask why you want the information you can tell them: “I wouldn’t want to determine the story before I have all my facts. I’m just doing my job at gathering information.” If you request data routinely from an agency (weekly), then it will be no big deal and they are less likely to question you.

"WE JUST DON’T WANT TO GIVE IT TO YOU"

The agency must explain its legal statutory reasons, usually in writing, for determining that an exemption applies to any particular information.

- You have the right to contest any exemption claim.
- The exemptions must be narrowly applied, since the FOIA was created to maximize public access to agency records.
- You can file an administrative appeal to a higher agency official. And if this fails, you can file a lawsuit. The federal court must conduct a full judicial review of the agency's claims and it is up to the agency to justify its denial of your request.
- Even if the agency releases substantial portions of the material you've requested, you can appeal the decision to "sanitize" the rest. You can also request a detailed justification for each deletion.
- While you are haggling with the agency, try to get the information from another agency. Some records are kept by multiple agencies (for example, boating accident data kept by state agencies and the Coast Guard).

Data-Driven News Pegs

Computer-assisted reporting is more than knowing spreadsheets and databases. It is a way of thinking. Below are 10 basic strategies for how stories can be improved or discovered with data. For every story you are writing about, ask yourself if it could be enhanced or proven with data that might be available or created based on the ideas below.

1. **Extreme.** Find the biggest, smallest, highest, lowest, richest, poorest. Example: Highest paid city employee.

2. **The Letterman list.** Instead of pinpointing the top or bottom, a ranked list is provided for readers. This helps readers find their state, city, resident hall, football team, etc., on the list and see the relation to similar units.
   Example: Money magazine's annual "100 Best Places in America to Live" list.

3. **Year-to-year.** Look for change in a unit from one year to the next.
   Example: Auto thefts dropped 12 percent in 2010 as compared to 2009.

4. **Long-term trends.** Look at the big picture by examining units over a long period of time, such as five, 10, 20 or 100 years. The numbers work well in a line chart.
   Example: Burglary rate has risen 36 percent over the past 20 years.

5. **Individuality.** Look for well-known individuals who might be of public interest.
   Example: Check databases of tax evaders, heavy water users or deadbeat dads with the name of prominent leaders to see if there is a match.

6. **Linking.** Link two different databases to see what matches come up.
   Example: Link DUI records with transit drivers to see who is driving the buses.

7. **Counting.** Add up numbers in a database to get interesting figures.
   Example: Check the county dog-bite database to find out how many people were bitten this year.

8. **Grouping totals.** Add up numbers for different groups and then rank them.
   Example: Add up the political contributions given to the mayor by the type of contributor and then rank them. Maybe developers gave the most money.

9. **Averages.** Find the average or mean of a set of numbers.
   Example: The average income of high school football coaches compared to other teachers.

10. **Comparison.** Take local numbers and compare to state or national numbers.
    Example: Average GPA for athletes compared to universities nationwide.

Posting Data Online

Here are some programs you can check out for posting data online for the public, from the simple, cheap and easy, to the complicated and expensive:

1. **Google Fusion Tables**

   http://www.google.com/fusiontables/

This is a free service for posting data online and making it look fancy through maps and other visualization techniques. You can upload data, share it, and allow people to update it if you want. Anyone can upload a basic CSV text file.

2. **Socrata**

   http://www.socrata.com/

   This is cheap and easy, and is actually used by the White House and some media organizations. For no cost you can post online really fast and share it with the world. Anyone can do it. The databases are kept on the Socrata server, though, and it is limited in space. You have to pay to put up large files.

3. **Tablesorter**

   http://tablesorter.com/

   This program gives you a little more control over your content and looks professional. It's also free (they ask for a donation if you like it). The downside is you have to know html to be able to integrate it into your website, but any media organization's Web person should be able to work with it.

4. **Caspio**

   http://www.caspio.com/

   This is a really nice program, and it even uses point-and-click interface for ease of use. It's very slick and has some great features. Probably the best out there, used by a lot of journalists. It can be a little more expensive, though. It ranges from $39.95 per month for the "value" package to $349.95 per month for the corporate package (more data can be stored and accessed).

5. **Django**

   http://www.djangoproject.com/

   This is a high-end program that requires a person to know Python programming language. It appears to be growing in popularity, including in journalism.

6. **Other programs**

   Other programs that you can check out include DataGrid, DataTables, Tableau public, Simile exhibit, and Tablesetter (open-source software created by ProPublica).

Data Ethics Checklist

Gathering and disseminating government data online raises ethical concerns that require some conscientious thinking. Here is a checklist to review when dealing with data (also, see the SPJ Code of Ethics, http://www.spj.org/ethicscode.asp):

1. **Is the database accurate?**

Often an agency will provide data that have been compiled in a hurry, changed from year to year, or just garbled for some technical reason. Take an excerpt of the data and verify it is accurate. Count the number of records to make sure it makes sense. Clean up the data. Some people say it's OK to take a government database and post it online, and if there are inaccuracies it is the government's fault. That is true, legally, but journalists have an ethical duty to verify and publish the truth.

2. **Did I analyze the numbers right?**

   Maybe, but it doesn't hurt to make sure. When you get your findings, run them past the agency to see whether they find errors. You would want to give them the chance to reply, and they can't prevent you from publishing it. No need to provide the entire story to an agency before publication, but it doesn't hurt to present key findings.

3. **Should I withhold some of the data from the public?**

   Just because we have a legal right to acquire public data doesn't mean we have to publish everything. We might get databases that include home addresses and home phone numbers. In some situations that might be important to publish online (e.g., sex offender data). In other situations the harm in publication might outweigh the benefits (e.g., database of child beauty pageant winners). Always weigh the harm vs. the good. There is no shame in reasoned editing.

4. **What would my mom/aunt/grandma/neighbor/child think?**

   Apply the sniff test to determine whether publication of the data might evoke a public backlash that could lead to closure of the records. This has happened many times around the country, particularly when data have been posted that include citizens and home addresses (e.g., concealed weapons permits) or other sensitive information (e.g., 911 recordings). People are worried about privacy invasion and will call for blanket closure of records. That doesn't mean we don't publish data if it has any whiff of personal privacy. But if we do, we must have a clear and defensible reason that we explain up front. What would you tell your mom?

5. **Am I pursuing data aggressively in the public's interest?**

   A lot of journalists don't write about denials or fight for records because they say it's inside baseball, or a conflict of interest. However, it is the ethical responsibility of journalists to fight for records on behalf of the public. Agencies aren't saying "no" to you – they are saying no to the thousands or millions of fellow citizens. Tell those citizens their agencies are keeping secrets. When government breaks the law (including the state public records law), tell people.

Data Access Resources

Investigative Reporters and Editors
www.ire.org

This group has an online resource center (for members) that includes a searchable database of more than 20,000 investigative stories and a searchable database of 2,000 tip sheets. Also check out the Extra! Extra! daily stories: http://www.ire.org/extraextra/ IRE also provides bootcamps to learn computer-assisted reporting.

National Institute for Computer-Assisted Reporting
www.nicar.org
NICAR is a part of Investigative Reporters and Editors, providing training and data for journalists. The database library provides cheap data that is cleaned and ready to go for analysis. Also, see the group's newsletter, Uplink, for story ideas and tips.

NICAR Listserv
Subscribe to the NICAR listserv for getting answers to CAR problems. It's a little heavy and brings in about 15 e-mails a day, but worth lurking for a while. To subscribe, send an e-mail to LISTSERV@MIZZOU1.MISSOURI.EDU. In the message of the first line, write: subscribe NICAR-L your name.

"Computer-Assisted Reporting: A Practical Guide"
An outstanding book that includes data for practicing and nuts-and-bolts examples. Written by Brant Houston, former executive director of IRE. Order online at the IRE Web site: www.ire.org.

Drew Sullivan's Bookmarks of Databases
www.drewsullivan.com/database.html
Drew's compiled a list of interesting free downloadable and searchable databases on the Internet. Good for finding raw data.

IRE Beat Page
http://www.ire.org/resourcecenter/initial-search-beat.html
Shawn McIntosh's now-famous page listing links by beat. Try this one for a broad overview. (Hosted by IRE's Reporter.org)

Experts who know the ropes
Get to know reporters at newspapers that use CAR, profs who use or teach Excel and Access and computer programmers at computer shops. Feel free to contact me (David Cuillier), and I can give you tips or refer you to others who can help you out: cuillier@email.arizona.edu. Also, find a stats guru at a local university who might be able to help with data analysis, as well as local computer programmers who might be willing to help with analysis and data transfer. Join local computer users groups.