Where to find data

Find data through Google

- Google has a bunch of advanced functions that can help you find data and/or documents quickly.
- Search for specific phrases with quotes ""
  - Example: **"new york population"**
- Exclude search words with a hyphen -
  - Example: **"new york population" -city**
- Search for something either/or with OR
  - Example: **budget OR appropriations**
- Find websites that are similar to others by using related:
  - Example: **related:propublica.org**

Find data through Google

- Search only within a certain domain by using site:
  - Example: **site:propublica.org; site:nyc.gov**
- Search for a specific file type by using filetype:
  - Example: filetype:pdf
  - Other useful filetypes:
    - filetype:ppt (PowerPoint)
    - filetype:xls (pre-Excel 2007)
    - filetype:xls (post-Excel 2007)
      - **Note: check for both xls and xlsx!**
    - filetype:doc (Word documents)
    - filetype:txt (text files
    - filetype:csv (comma-delimited text files)

Find data through Google

- You can use this to find data or documents from specific agencies
  - For example, searching **filetype:doc site:nyc.gov** will bring up any Word documents on the New York City government website.
- And you can use the operators we discussed earlier to further drill down your search
  - Example: **subway OR bus filetype:pdf site:nyc.gov** will bring up any PDF on the New York City government website that mentions the word "subway" or "bus."

- There's plenty more where that came from
  - Find a full list of advanced search operators at https://www.google.com/advanced_search

Getting data from the government

- Your first step: calling the agency
    - Never send out a "blind" records request

- Describe what you're looking for, or what you're trying to find
    - Get specific names of database or documents that you want

- Filing a formal records request should be your last resort
    - The easiest route is to ask the agency for the records
    - "Let's cut down on the paperwork here."
    - Formal requests take time!

Getting data from the government

- Laws govern the release of data and information to the public

- Federal: Freedom of Information Act (FOIA)

- State: Each state has its own open records law.
  - **Know your state's law!**
  - What's allowable in one state might not be in another.
  - Different states might have different response deadlines.
  - States often have First Amendment coalitions/advocacy groups. Find yours.

Getting data from the government

- For both federal and state records requests, save time and make sure you're requesting documents from the correct agency before filing

- Since you've done your homework, you should have a specific name or description of the document(s)

- Be sure to ask for electronic fulfillment, unless there's a reason you need paper

Getting data from the government

- Some tips on filing a records request:

- Ask that your request be fulfilled on a "rolling basis," i.e. release the documents as they come

- It never hurts to ask for a fee waiver

Getting data from the government

- Agencies have a deadline to respond to you. Set an alarm or reminder for that date
  - Example: FOIA law says the agency has to send you a response within 20 business days
  - Keep track of requests with a website like [FOIA Machine](#)

- Toe the line between specific and vague
  - Too specific and you might inadvertently exclude some records. Too vague and you run the risk of the agency saying the request is too vast to fulfill
  - This is where your previous phone calls come into play
  - Limit time periods. 5 years of records is easier to fulfill than 50

Source: Jack Gillum, Associated Press, [IRE tipsheet](#). Reprinted with permission from IRE.

Getting data from the government

- If you are charged fees, always negotiate
    - Agencies might overcharge as a denial/stalling tactic
    - Per-page reproduction fee? Don't need it if it's electronic.
    - "It will take $10,000 for 50 hours to fulfill this request..."
        - How much is that employee getting paid an hour?

- Know your rights
    - Denied? You can (and should) appeal
    - Be polite, but don't take "no" for an answer
    - Public shaming

PRO PUBLICA | Journalism in the Public Interest

Foiled by FOIL: How One City Agency Has Dragged Out a Request for Public Records for Nearly a Year

After eight proposed delivery dates, the Administration for Children's Services still has not provided public records we asked for almost a year ago.

by Joaquin Sapien
ProPublica, April 21, 2016, 2:18 p.m.

1 Comment | Print

Getting data from the government

Pursuant to the Alabama Open Records Law § 36-12-40 et seq., I am requesting the following documents:

The equity/revenue-and-expenses report completed by the athletic department for the National Collegiate Athletic Association for the 2014 fiscal year. This report is a multi-page document that had to be submitted to the NCAA by Jan. 15, 2015. It contains 38 revenue and expense categories, followed by specific breakdowns of each of those categories, by sport and gender. I am requesting the full report, including the detail tables and the Statement of Revenues and Expenses that appear at the end of the report.

PLEASE NOTE: The NCAA report is different than the equity report that is sent to the U.S. Department of Education for Title IX compliance.

Please advise me in advance of the estimated charges associated with fulfilling this request.

In the interest of expediency, and to minimize the research and/or duplication burden on your staff, please send records electronically if possible. If this is not possible, please notify me by phone at 415-323-5830 before sending to the address listed below.

- Cites the law
- Is specific (reporter has done his/her homework)
- Asks for electronic records

Now that you've got 'em...

A note on Socrata

- Many government agencies use Socrata as an online clearinghouse for their data
  - Example: https://nycopendata.socrata.com/
- Socrata has limited functionality (sorting/filtering), but it's just as easy to download the data on your own
- Just choose "Export" in the upper right-hand corner of the screen

- A drop-down menu will appear, allowing your pick of data formats (text files, Excel, PDF)

Bust through PDFs with Tabula

- PDFs are terrible files for storing data

- As long as the PDF is machine-readable, online tools like Tabula can help

- Machine-readable means the underlying data in the PDF is still intact

Bust through PDFs with Tabula

- How can you tell?
  - If you can highlight the data in the PDF, you're good to go
  - If you can't highlight it, i.e. it looks like a scanned document or a photo, Tabula can't recognize the data

- If your PDF of data is not machine-readable, you will need OCR (optical character recognition) software to convert into machine-readable data.
  - Or go back to your data source and ask for a machine-readable file

How to load data

Types of data files

- Data can be stored in a myriad of file types

- Text files -- the simplest data storage type. Basically a text document of data
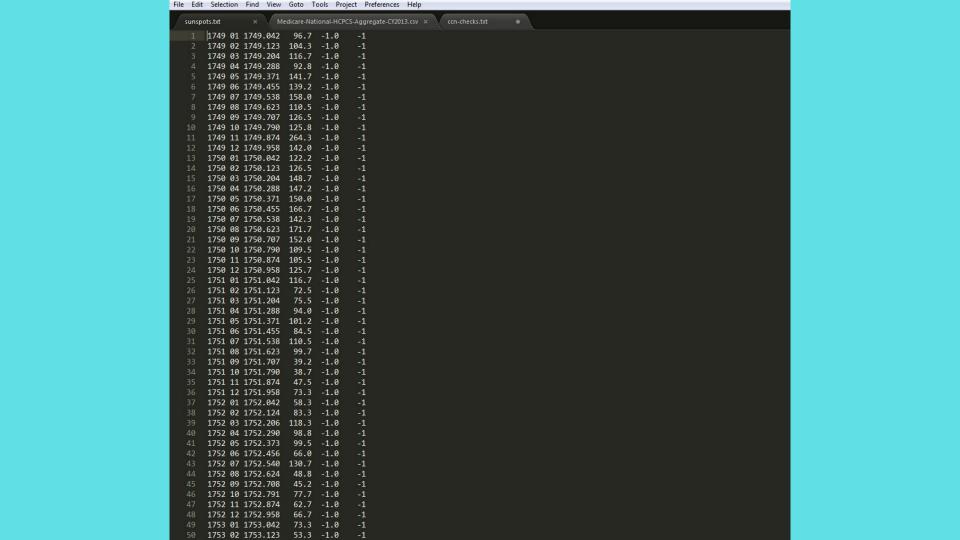
- File type is usually .txt
  - Obviously a 10,000-page document of data isn't useful. We have to separate it somehow
  - Generally two ways to do this

Types of data files

- Check the file name extension (.csv, .txt, etc.)

- Open the file up in a text editor and browse

- Do you have a header row?

- What's separating each field?

sunspots.txt   ✕   Medicare-National-HCPCS-Aggregate-CY2013.csv   ✕   ccn-checks.txt   ●

```
 1    1749 01 1749.042    96.7  -1.0    -1
 2    1749 02 1749.123   104.3  -1.0    -1
 3    1749 03 1749.204   116.7  -1.0    -1
 4    1749 04 1749.288    92.8  -1.0    -1
 5    1749 05 1749.371   141.7  -1.0    -1
 6    1749 06 1749.455   139.2  -1.0    -1
 7    1749 07 1749.538   158.0  -1.0    -1
 8    1749 08 1749.623   110.5  -1.0    -1
 9    1749 09 1749.707   126.5  -1.0    -1
10    1749 10 1749.790   125.8  -1.0    -1
11    1749 11 1749.874   264.3  -1.0    -1
12    1749 12 1749.958   142.0  -1.0    -1
13    1750 01 1750.042   122.2  -1.0    -1
14    1750 02 1750.123   126.5  -1.0    -1
15    1750 03 1750.204   148.7  -1.0    -1
16    1750 04 1750.288   147.2  -1.0    -1
17    1750 05 1750.371   150.0  -1.0    -1
18    1750 06 1750.455   166.7  -1.0    -1
19    1750 07 1750.538   142.3  -1.0    -1
20    1750 08 1750.623   171.7  -1.0    -1
21    1750 09 1750.707   152.0  -1.0    -1
22    1750 10 1750.790   109.5  -1.0    -1
23    1750 11 1750.874   105.5  -1.0    -1
24    1750 12 1750.958   125.7  -1.0    -1
25    1751 01 1751.042   116.7  -1.0    -1
26    1751 02 1751.123    72.5  -1.0    -1
27    1751 03 1751.204    75.5  -1.0    -1
28    1751 04 1751.288    94.0  -1.0    -1
29    1751 05 1751.371   101.2  -1.0    -1
30    1751 06 1751.455    84.5  -1.0    -1
31    1751 07 1751.538   110.5  -1.0    -1
32    1751 08 1751.623    99.7  -1.0    -1
33    1751 09 1751.707    39.2  -1.0    -1
34    1751 10 1751.790    38.7  -1.0    -1
35    1751 11 1751.874    47.5  -1.0    -1
36    1751 12 1751.958    73.3  -1.0    -1
37    1752 01 1752.042    58.3  -1.0    -1
38    1752 02 1752.124    83.3  -1.0    -1
39    1752 03 1752.206   118.3  -1.0    -1
40    1752 04 1752.290    98.8  -1.0    -1
41    1752 05 1752.373    99.5  -1.0    -1
42    1752 06 1752.456    66.0  -1.0    -1
43    1752 07 1752.540   130.7  -1.0    -1
44    1752 08 1752.624    48.8  -1.0    -1
45    1752 09 1752.708    45.2  -1.0    -1
46    1752 10 1752.791    77.7  -1.0    -1
47    1752 11 1752.874    62.7  -1.0    -1
48    1752 12 1752.958    66.7  -1.0    -1
49    1753 01 1753.042    73.3  -1.0    -1
50    1753 02 1753.123    53.3  -1.0    -1
```

Types of data files

- Fixed width

- You must manually put breaks in the data to tell your computer where a new field ends and another begins

- If you have a fixed width file, make sure your source gives you a guide or schema that tells you where to put the breaks. Don't guess!
  - Example

Types of data files

- Delimited -- the data fields are separated by a delimiter (often punctuation) that tells our computer where a new field ends and another begins
  - This is the more common of the two
  - Common delimiters
    - Commas (also known as CSV files, or comma-separated values)
    - Most common delimiter
    - This file type can be .csv instead of .txt
  - Tabs
  - Pipes (|)

Types of data files

- But what if you have commas in a field?
  - We use a text qualifier (usually quote marks) to tell the computer to ignore any commas in the field
  - Example: AT&T, Inc. will look like "AT&T, Inc." in the text file

Medicare-National-HCPCS-Aggregate-CY2013.csv ✕ | ccn-checks.txt ⬤

HCPCS Code,HCPCS Description,HCPCS Drug Indicator,Place of Service,Number of Providers,Number of Services,Number of Unique Beneficiary/Provider Interactions,Number of Distinct Medicare Beneficiary/Per Day Services,Average Submitted Charge Amount,Minimum Submitted Charge Amount,Maximum Submitted Charge Amount,Standard Deviation of Submitted Charge Amount,Average Medicare Allowed Amount,Minimum Medicare Allowed Amount,Maximum Medicare Allowed Amount,Standard Deviation of Medicare Allowed Amount,Average Medicare Payment Amount,Minimum Medicare Payment Amount,Maximum Medicare Payment Amount,Standard Deviation of Medicare Payment Amount

A0425,"Ground mileage, per statute mile",N,F,10343,140801494.5,6755692,12671918,15.0939463,0.076552661,408.2278481,8.08359992,7.801376767,0.048680084,10.74028436,0.925850875,6.133360523,0,8.593333333,0.726987114

A0425,"Ground mileage, per statute mile",N,O,16,3458.5,824,1304,12.51162064,6.87,120,2.463704996,7.093404655,6.87,10.74,0.064904357,5.577900824,5.385714286,8.42,0.05160961

comma delimiter

A0426,"Ambulance service, advanced life support, non-emergency transport, level 1 (als 1)",N,F,3000,325543.3,282932,317966,791.3980132,7,28437,458.3650808,262.1513796,7,347.09,19.76073411,203.955559,5.556,277.67,15.76392806

A0427,"Ambulance service, advanced life support, emergency transport, level 1 (als1-emergency)",N,F,8378,4972850.9,3490463,4936591,923.3939244,90.67,28655.35714,460.1374707,417.5188687,90.67,549.57,32.27543877,324.7140335,0,436.73,25.20193278

A0427,"Ambulance service, advanced life support, emergency transport, level 1 (als1-emergency)",N,O,3,480,382,480,589.71875,445,600,6.62807024,421.2877292,397.57,425.45,1.099862878,330.9389375,311.7,340.36,0.977981327

text qualifier

A0428,"Ambulance service, basic life support, non-emergency transport, (bls)",N,F,4973,6812508.1,2053295,4844933,537.1170175,95,28454.625,275.6785558,220.8932109,95,289.24,17.11772737,172.9915218,0,228.65,13.46258727

A0429,"Ambulance service, basic life support, emergency transport (bls-emergency)",N,F,9722,2731110.7,2014926,2716724,669.2708485,76.87969925,31869.9,288.9939803,358.1562734,75.70225564,462.79,29.92258515,278.1289712,59.70285714,370.23,23.76236904

A0429,"Ambulance service, basic life support, emergency transport (bls-emergency)",N,O,3,822,572,822,499.6046594,252.03,500,9.047700509,354.757944,252.03,423,4.301793185,278.7129805,197.59,331.63,3.378427184

A0430,"Ambulance service, conventional air services, transport, one way (fixed wing)",N,F,89,10826,10072,10793,14438.15216,700,29903.92308,5062.165345,4190.477053,700,4731.355452,412.6544457,3275.365067,547.9038462,3652.114057,324.2669331

A0431,"Ambulance service, conventional air services, transport, one way (rotary wing)",N,F,275,56410,54843,56302,17309.74183,360,27709.38536,4177.623604,4599.301837,360,5510.41,296.4455132,3585.952542,282.24,4356.8975,233.0686692

A0432,"Paramedic intercept (pi), rural area, transport furnished by a volunteer ambulance company which is prohibited by state law from billing third party payers",N,F,59,3146,2641,3143,638.4824698,125,1492.44,161.4826762,375.3057629,125,407.41,29.98904694,288.0348601,98,320.7589655,24.78549587

A0433,"Advanced life support, level 2 (als 2)",N,F,5459,111967.8,108047,111419,1101.45624,2.06,5141.55,549.7118639,602.5650325,2.06,795.42,53.82315441,468.8264066,1.62,636.34,42.87535231

A0434,Specialty care transport (sct),N,F,1300,104281.7,79913,97874,1865.055575,40,32209.9,1148.997883,744.0198998,22.12932166,940.05,63.95298324,583.2122558,17.58468271,742.0133333,50.91906057

A0435,"Fixed wing air mileage, per statute mile",N,F,89,1980144.2,10020,10729,104.9972547,6.93,278.7912166,50.98005411,11.75196535,6.93,12.48003565,1.019015093,9.202622582,5.433166667,9.984006024,0.805605927

A0436,"Rotary wing air mileage, per statute mile",N,F,272,3322854,54437,55885,186.5635308,33.32,281.8598527,56.57891524,31.5708137,22.20998926,33.32000599,1.813376021,24.65698949,16.025,26.656,1.410686318

A0999,Unlisted ambulance service,N,F,486,3678.2,2767,3396,698.2032162,12,20283.27,495.3732378,357.4373144,7.089705882,3286.46,127.2913899,280.9220189,5.555,2576.59,100.0887829

A4212,Non-coring needle or stylet with or without catheter,N,O,4,19,18,19,6.236842105,5,17,2.966339322,0.405263158,0.01,6.5,1.454431951,0.324736842,0.01,5.2,1.162530672

A4215,"Needle, sterile, any size, each",N,O,22,182,121,137,6.50967033,0.18,80,8.595171502,0.037417582,0.01,5,0.368865846,0.031153846,0.008194444,4,0.295003223

Types of data fields

- Just like there are different types of data files, there are different types of data fields, and how you upload them can make a huge difference

- Text/characters
  - Short, alphanumeric values. Things you aren't going to need to calculate
    - Names, addresses, descriptions
    - Zip codes -- they're numbers but we won't ever need to add them together

- Numeric/decimals
  - Numeric values. Things we will want to calculate
    - Counts, sums, dollar amounts

Types of data fields

- Dates/times
  - Stating the obvious. If you have a date or time, selecting this option should make it easier to calculate date/time differences in Excel/Google Sheets

- Important! Text vs. numeric will affect sorting order

Types of data fields

- Don't just double-click your data file to open it!

- Excel thinks it's being helpful by automatically uploading your data and guessing at the data types
  - It's almost always wrong.

Types of data fields

- This can cause trouble like dropping leading zeroes
  - Example: New England zip codes.
    - They almost always start with zero, i.e. 02901 (a zip in Providence, Rhode Island).
    - Unless you specifically choose text for a zip code field, Excel assumes it's numeric and "helpfully" drops that leading zero.
    - Thus, 02901 becomes 2901 and can cause problems down the line.

- It will also unhelpfully reformat your number to a date!
  - Example: Say there's a value in your data that is 11-53
  - Excel assumes this is a date (November 1953) and uploads it as 11/1/1953 or Nov-53.