SAGE was founded in 1965 by Sara Miller McCune to support the dissemination of usable knowledge by publishing innovative and high-quality research and teaching content. Today, we publish more than 750 journals, including those of more than 300 learned societies, more than 800 new books per year, and a growing range of library products including archives, data, case studies, reports, conference highlights, and video. SAGE remains majority-owned by our founder, and after Sara's lifetime will become owned by a charitable trust that secures our continued independence.

Los Angeles | London | Washington DC | New Delhi | Singapore | Boston

DATA LITERACY

A User's Guide

David Herzog
Missouri School of Journalism

Los Angeles | London | New Delhi
Singapore | Washington DC | Boston

SECTION II

IDENTIFYING AND OBTAINING DATA

CHAPTER 2 CLUES FOR UNCOVERING DATA

One of the big challenges in becoming data literate is being able to quickly identify data and determine which government agency keeps them. If the data are posted on the Internet, we can find and download them. If data are not posted online, we can request them from the agency.

To meet that challenge we need to develop a data state of mind, one that opens our eyes to the proliferation of data online, in government agency mainframes and servers, and in university libraries. After developing a data state of mind—or data antennae—it will be easier for us to see the possibilities. Data are everywhere.

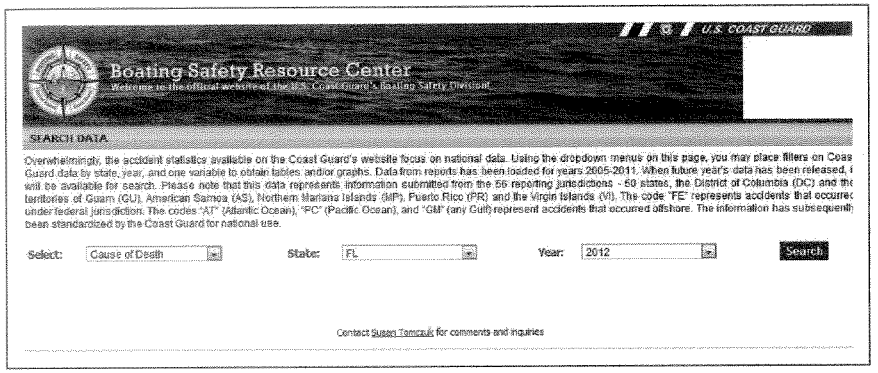
Understanding how and why government agencies process data is a good first step toward developing that data state of mind. Specifically, we'll examine how and why agencies collect, analyze and publish data. To look at it another way, we'll examine the data input and output.

HOW AND WHY AGENCIES COLLECT, ANALYZE AND PUBLISH DATA

Broadly speaking, agencies collect, analyze and publish data for one or more of three primary reasons: (1) they're required to do so by law, (2) agencies believe the data will help them execute their missions or (3) agencies are participating in an open-government effort.

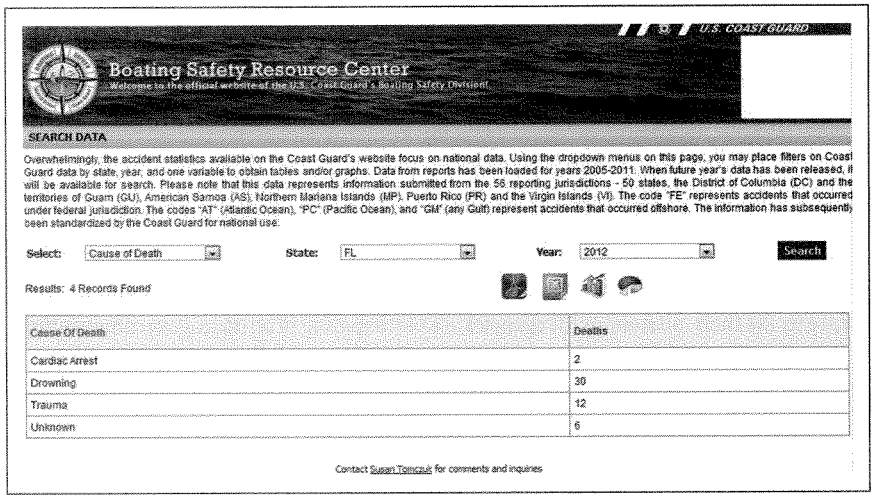
Laws are one of the biggest motivations for state, local or federal government agencies to collect data. Sometimes the laws specifically require the creation of a database. Other times agencies use databases so they can comply with their responsibilities laid out in the laws.

The U.S. Coast Guard's Boating Accident Report Database holds data about recreational accidents reported to state law-enforcement authorities. In the most recent year available, 2012, the Coast Guard recorded 4,515 accidents that killed 651 people and caused 3,000 injuries. The Coast Guard uses the database, which anyone can obtain by making a U.S. Freedom of Information Act request, to track boat accident trends and to generate boating statistics publications that are posted on the Web (Coast Guard, n.d.). In addition, the Coast Guard provides a search interface that allows users to query the database and generate reports in tabular form that they can then export to an Excel spreadsheet.



Source: Coast Guard. Retrieved from https://bard.cns-inc.com/Screens/PublicInterface/Report1.aspx.

Note: Boating accident search page. The U.S. Coast Guard allows the public to search its database of recreational boating accidents with this Web form.



Source: Coast Guard. Retrieved from https://bard.cns-inc.com/Screens/PublicInterface/Report1.aspx.

Note: Cause of death search results for the state of Florida, 2012.

Publishing these data on the Web clearly is a benefit to law enforcement; boating safety advocates, such as the National Safe Boating Council; and members of the public. Regardless of these benefits, the Coast Guard is doing so because of a law. Congress in 1983 passed legislation establishing the State Marine Casualty Reporting System within the U.S. Department of Transportation. The Coast Guard was part of the Department of Transportation and now is part of the Department of Homeland Security.

The law, as amended the next year, says,

(a) The Secretary shall prescribe regulations for a uniform State marine casualty reporting system for vessels. Regulations shall prescribe the casualties to be reported and the manner of reporting. A State shall compile and submit to the Secretary reports, information, and statistics on casualties reported to the State, including information and statistics concerning the number of casualties in which the use of alcohol contributed to the casualty.

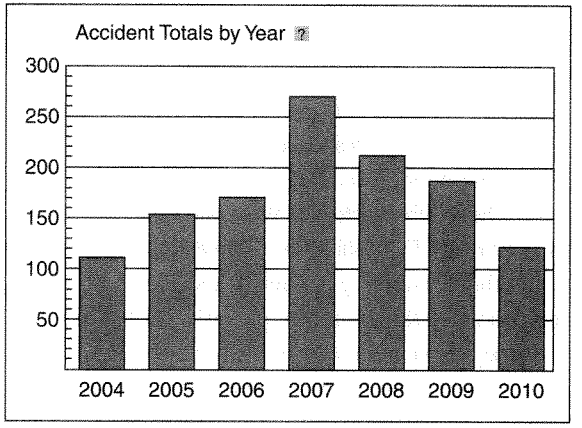
(b) The Secretary shall collect, analyze, and publish reports, information, and statistics on marine casualties together with findings and recommendations the Secretary considers appropriate. If a State marine casualty reporting system provides that information derived from casualty reports (except statistical information) may not be publicly disclosed, or otherwise prohibits use by the State or any person in any action or proceeding against a person, the Secretary may use the information provided by the State only in the same way that the State may use the information. (U.S.C. Title 46, 2011)

We can see from the letter of the law that the Department of Transportation was under orders not just to compile these data from state agencies in a consistent format, but also to generate reports and make them public.

Although the law does not give instructions about what data elements need to be collected for the reports, the department later developed regulations that determine when boating accidents must be reported, by whom and within what time frame. In addition, the regulations specify the minimum data that should be collected, including location, time and date, weather and water conditions, the availability and use of personal flotation devices and more (Government Printing Office, 2001a, 2001b).

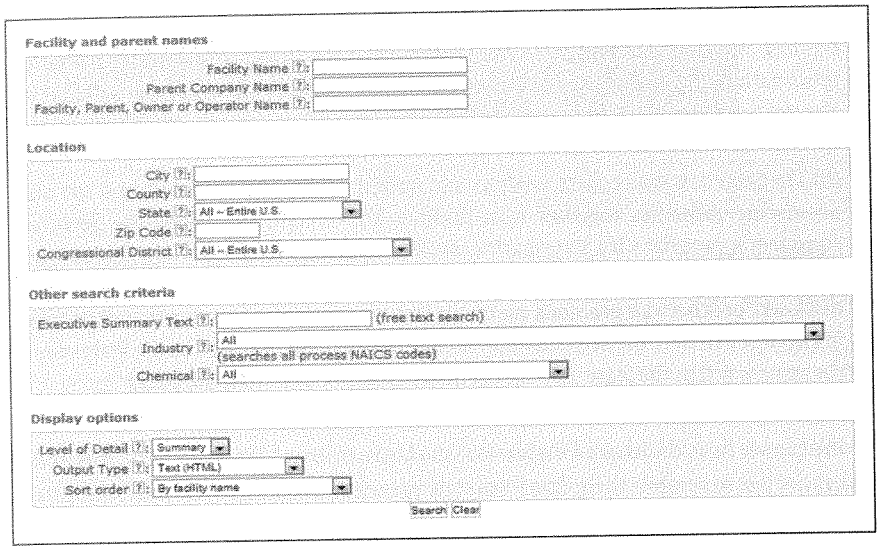
Another federal database that was created to comply with the law is the U.S. Environmental Protection Agency's Risk Management Plan database. The EPA's RMP database holds data about certain chemicals that are produced, stored or distributed at facilities across the country. These RMPs are supposed to be a tool to help local emergency responders, such as firefighters and paramedics, to better plan for and respond to fires and explosions, like the deadly one at the West Fertilizer Company in the town of West, Texas, in 2013.

The EPA does not post the database online itself: the Right to Know Network, a project of the nonprofit Center for Effective Government in Washington, DC, posts it. Visitors to the RTKNet website can examine data that have already been summarized, such as this chart showing the number of



Source: Risk Management Plan Database. (n.d.). The Right-to-Know Network. Retrieved July 10, 2013, from http://www.rtknet.org/db/rmp/search.

Note: U.S. hazardous chemical accidents by year. The nongovernmental Right to Know Network generates charts using data from the U.S. Environmental Protection Agency.



Source: Risk Management Plan Database. (n.d.). The Right-to-Know Network. Retrieved July 10, 2013, from http://www.rtknet.org/db/rmp/search.

Note: Right to Know Network's U.S. Risk Management Plan search screen.

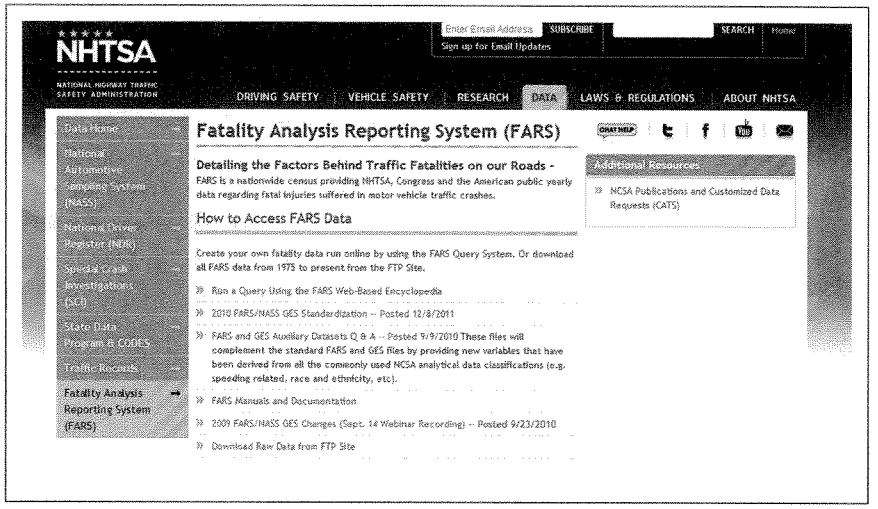
reportable accidents by year. Site visitors can also run more-complicated queries using a form to drill down and get details about particular facilities (Center for Effective Government, n.d.).

Established in 1970 by President Richard Nixon, the EPA is in charge of enforcing federal air, water and ground pollution laws. (In some states, the EPA has delegated duties to state-level agencies.) The RMP database has its roots in the Clean Air Act of 1970, which sought to curtail air pollution caused by urbanization and industrialization.

Twenty years later, President George H. W. Bush signed into law amendments with provisions intended to prevent chemical accidents. The EPA developed rules and required facilities to comply by 1999 (Environmental Protection Agency, n.d.b).

The regulations specify which 77 toxic and 63 flammable substances facility owners must report, and at what threshold quantities. For example, facilities that have 10,000 pounds of anhydrous ammonia must report to the EPA (Chemical Accident Prevention Provisions, 1994). The fertilizer factory in West had 54,000 pounds on hand in June 2011, according to the most recent EPA data on the RTKNet website (Center for Effective Government, n.d.).

In other instances, government agencies create databases on their own initiative, to help them meet their own strategic goals. The National Highway Traffic Safety Administration is supposed to reduce deaths, injuries and economic losses caused by motor vehicle accidents in the United States (National Highway Traffic Safety Administration, n.d.). NHTSA, part of the U.S. Department of Transportation, developed the Fatality Analysis Reporting System database in 1975 as a tool for monitoring fatal traffic accidents across the



Source: National Highway Traffic Safety Administration. Retrieved from http://www.nhtsa.gov/FARS.

Note: Fatality Analysis Reporting System home page.

country and giving researchers data that they could use to examine the causes. State agencies collect the data about the fatal crashes and then provide them to NHTSA, which compiles FARS and releases the data annually to the public. Transportation planners, safety advocates, journalists and attorneys all have used the data for research. NHTSA itself has used the data over the years to generate dozens of analytical reports available for download.

NHTSA allows site users to query its data using a series of Web forms. More-advanced data users, such as those who have experience using database managers, can download raw data files back to 1975—the inaugural year for the database.

We can see that governments have different motivations for collecting, using and making data public. That's important for us to consider later when we're testing our data, using integrity checks. Our goals for working with the data can differ greatly from the goals of the people inside the agencies that collect them. For instance, we may be interested in determining which industry group's employees contributed the most money to candidates running for governor in our state. However, our state probably records the occupation and employer—but not industry category—of each contributor. That means we won't be able to answer that question without a lot of additional research and changes to our campaign contribution data.

CLUES FROM DATA ENTRY

As we begin to develop a data state of mind, we should keep our eyes open for clues that government agencies are creating databases. Agencies have a multitude of ways to enter data into a database, some of them high tech, others rooted in the paper-bound ways of the 20th century.

We may see government workers collecting data in the field using handheld computers as part of their jobs. For instance, parking enforcement officers in your city and on your campus might be using devices that allow them to enter violation data and to print out, right on the spot, a ticket that looks like a store or ATM receipt. Some units, which look like big, rugged calculators, have number and character buttons for the data entry. Others have touchscreens with styluses. Some even come with GPS receivers so the officers can record the precise location of where a parking violation occurred. In a similar vein, other government agencies, such as fire departments and health departments, have been using tablet devices to input onsite inspection data.

When we see this data collection, we can assume that the data do not stay in the individual handheld unit or tablet. In fact, these data later are transferred to a centralized database that holds all of the inspections.

Likewise, police officers do data entry in the field when they respond to calls from 911 dispatch centers. With the aid of their ruggedized laptop computers mounted next to them in their cruisers, they use templates to enter distinct pieces of data about the report. For instance, an officer might enter data about the location, nature and outcome of an incident, and the name and contact information of any witnesses interviewed at the scene. When the officer has completed the report, he or she can send it to the police department's centralized incident reporting system.

Most data entry, however, is nowhere near as advanced. A lot of data entry is done using paper or Web forms, so we should look for those, too, as clues. It's hard to believe, but a lot of data at government agencies are entered manually. A clerk sits at a computer and keys in data that someone has entered on a paper form. So, if you see someone entering data in a government building, casually ask them what they're doing and about the data they work with. Sometimes government agencies scan their forms and use software to digitally extract the data.

If you're interested in a database, get a copy of the form that's used to feed data into it. Forms tell a lot about what you can expect to find in databases. You should assume that all of the data collected on the form will be entered. You may be incorrect, but you should start with the assumption that more data are available.

Here's a good example of a government form used to collect data. The U.S. Bureau of Alcohol, Tobacco, Firearms and Explosives licenses firearms dealers, which include big box retail, sporting goods and hunting shops. Whenever a federal firearms licensee (or FFL) loses or has firearms stolen, the licensee must report that loss or theft to the ATF's National Tracing Center within 48 hours. The ATF enters the data about the stolen firearms into the database and later searches them when law enforcement agencies want to trace guns that were used in crimes (Bureau of Alcohol, Tobacco, Firearms and Explosives, n.d.).

U.S. Department of Justice
Bureau of Alcohol, Tobacco, Firearms and Explosives

OMB No. 1140-0039 (07/31/2012)

Federal Firearms Licensee Firearms Inventory Theft/Loss Report

All entries must be in ink. Please read notices and instructions on reverse carefully before completing this form.

Section A - Federal Firearms Licensee Information

Federal Firearms License Number	Federal Firearms Licensee Telephone Number *(Include area code)*
Trade/Corporate Name	
Street Address of Federal Firearms Licensee	City / State / Zip Code / Telephone Number *(with area code)*
Full Name of Person Making Report	
Street Address of Person Making Report	City / State / Zip Code / Telephone Number *(with area code)*

Section B - Theft/Loss Information

	Date	Time	Description of Incident
Date of Theft/Loss Discovered			☐ Burglary ☐ Robbery
Police Notification			☐ Larceny ☐ Missing Inventory
ATF Notification			ATF Issued Incident Number

Name of Local Authority to Whom Reported (For burglary, larceny or robbery, include the police report number and officer/ detective name)

Street Address of Local Authority			Theft Location if Different from FFL Premises		
City	State	Zip Code	City	State	Zip Code

Name and Telephone Number of the ATF Representative Notified *(If this report is the result of an ATF compliance inspection, provide the name and telephone number of the ATF Inspector.)*

Brief Description of Incident *(e.g., How firearms were stolen, etc.)*:

Section C - Description of Firearms

Acquisition Date	Type	Manufacturer	Model	Caliber/ Gauge	Serial Number

Certification

I hereby certify that the information contained in this report is true and correct. I also understand that failure to report the theft or loss of a firearm from my inventory or collection within 48 hours of the discovery of the theft/loss is a violation of 18 U.S.C. § 923(g)(6) punishable as a felony.

Signature of Licensee | Date

ATF Form 3310.11
Revised September 2009

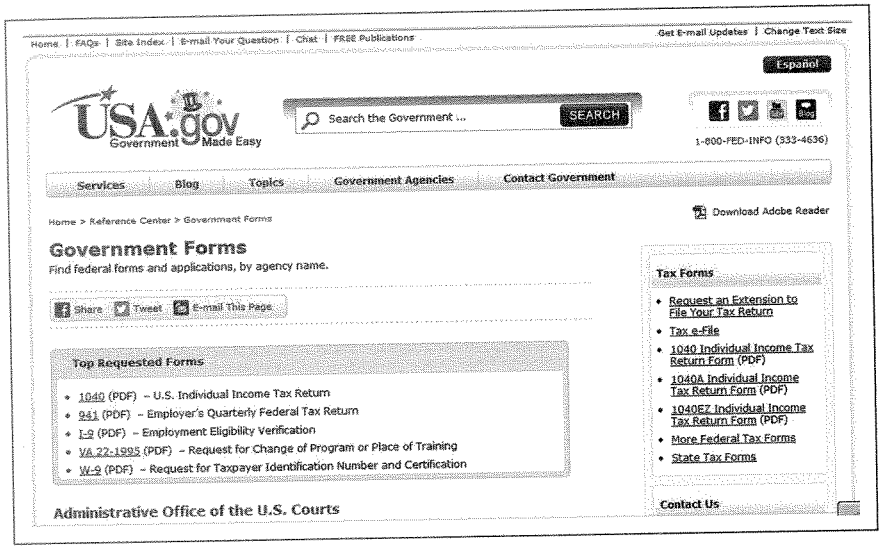
Source: Bureau of Alcohol Tobacco, Firearms and Explosives, Department of Justice. Retrieved from https://www.atf.gov/files/forms/download/atf-f-3310-11.pdf.

Note: Federal form used to report lost or stolen firearms.

The boxes or fields on the form tell us exactly what gets collected. We see details about the FFL, including license number and street address. The form asks for the contact

information of the person making the report; details about the theft or loss, including when local police were contacted; type of incident; and a brief description. Also, there's space for data about the firearms stolen or lost, including the manufacturer, model and serial number.

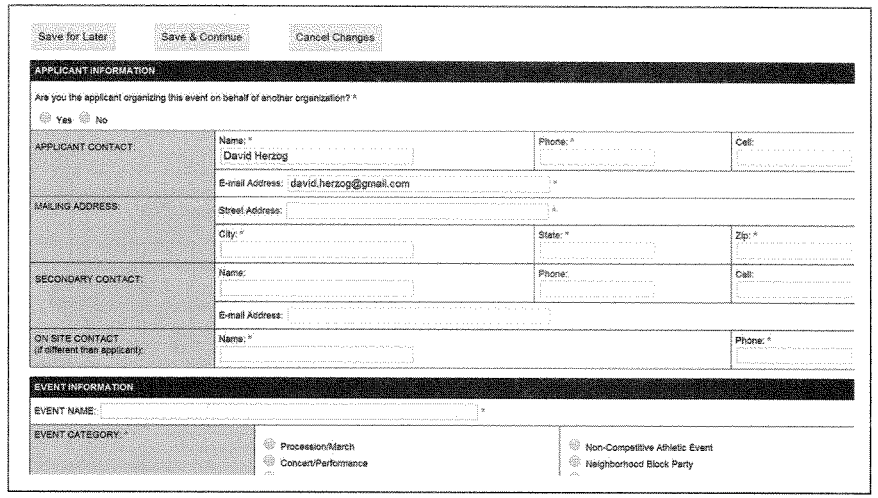
It might be an overstatement to say government agencies love forms, but they certainly do thrive on them. In fact, if you type http://forms.gov in your browser, you'll land at the federal government's portal for forms.



Source: Retrieved from http://www.usa.gov/Topics/Reference-Shelf/forms.shtml.

Note: Website for forms used by federal agencies.

Some of the forms accessible from this portal are Web forms, not documents such as Microsoft Word or Adobe Acrobat files. More agencies, large and small, are employing these online forms to collect data. For example, anyone who wants to have a parade, run or other event using the streets in the city of Columbia, Missouri, must complete the following Web form, then click the Submit button. These data then go from the form into a database administered by the city. The form's URL, which has a .php extension, is our clue to the presence of a database. **PHP** is a programming language that lets Web forms pass data to databases. Also be on the lookout for **.cfm** (Adobe ColdFusion) and **.asp** (Microsoft Active Server Pages), which can also pass data from a Web form to a database. By examining the city's form, we know that it collects data about the person making the request and the nature of the event, such as the time, date and location.



Source: City of Columbia, Missouri. Retrieved from https://www.gocolumbiamo.com/CMS/special_events/step1.php.

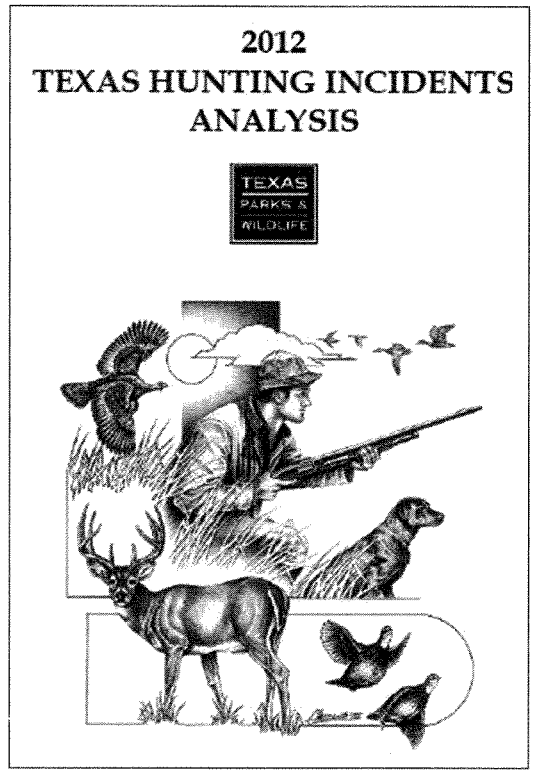
Note: City of Columbia, Missouri, form for parade permit applications.

CLUES FROM REPORTS

So far, we've been looking at how data collection can help us uncover clues about databases. Now we're going to look at how the products of databases, such as reports, can help. Another way of thinking about this is that data collection provides the input, reports provide the output.

As we saw earlier, agencies may collect data to fulfill legal mandates or help them meet strategic goals. The same applies to agencies issuing reports. In addition, executive agencies—at the federal, state or local level—often generate reports using data at the request of the legislative bodies that oversee them.

The Texas Parks and Wildlife Department regulates hunting in the state and issues licenses to hunters. It also collects data about hunting accidents that it



Source: Texas Department of Parks and Wildlife. Retrieved from http://tpwd.state.tx.us/publications/pwdpubs/media/pwd_rp_k0700_1124_2012.pdf

Note: Texas annual report for hunting accidents.

uses to produce annual reports (Texas Parks and Wildlife Department, n.d.). These reports summarize accidents for the year and, in some cases, provide data from other years for comparison. The summarized tables, like the ones below, provide a good clue that the underlying data for this report came from a database.

EQUIPMENT	Percentage in parentheses (%)			
Rifles	16 (55)	8 (32)	11 (48)	11 (44)
Shotguns	11 (38)	16 (64)	9 (39)	11 (44)
Handguns	2 (7)	1 (4)	2 (9)	3 (12)
Muzzleloader	0	0	0	0
Bow	0	0	1 (4)	0

Source: Texas Department of Parks and Wildlife, Retrieved from http://tpwd.state.tx.us/publications/pwdpubs/media/pwd_rp_k0700_1124_2012.pdf

Note: Table in hunting accidents report.

We get an even stronger indication that these data came out of a database when we arrive at the section that provides details about the individual hunting incidents. (Two fatal accidents from 2012 are shown below.) Each has a header for the date, county, shooter's age, gender, firearm, animal hunted, whether the accident was self-inflicted and whether the shooter had taken a hunters' education course.

2012 FATAL INCIDENTS FIREARM/BOW HUNTING RELATED (A)*

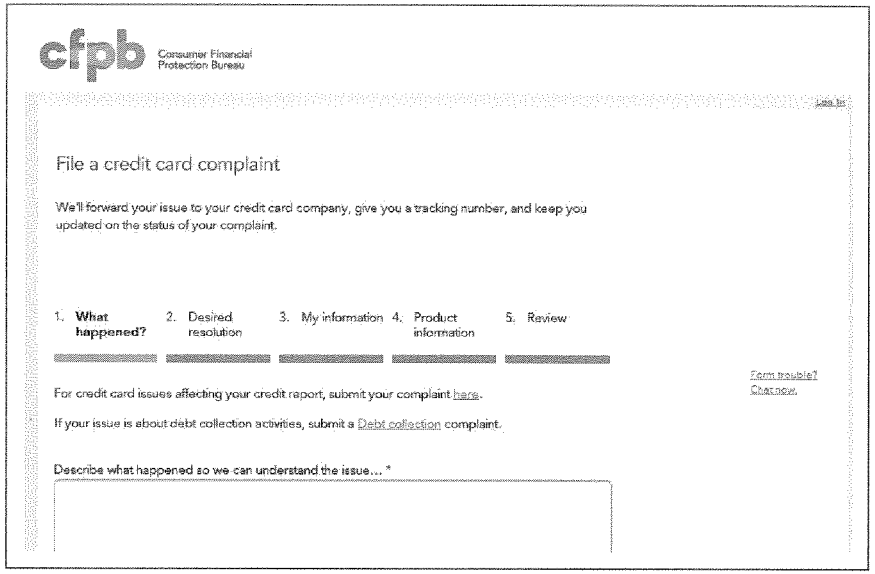
*A. Firearm/Bow & Hunting Related—An accident/incident resulting from the discharge of a firearm or bow while hunting, which causes the injury or death of any person(s).

Date	County	Shooter's Age/Gender	Firearm	Animal Hunted	Self-Inflicted?	Hunter Ed? (Shooter)
4-2	Bexar	36/M	Rifle	Opossum	No	No
Comments: Shooter shot at an opossum after dark and did not see that his friend was in the line of fire.						
Prevention: Always point the muzzle in a safe direction; always stay within a safe zone of fire; communicate with hunting companions; know where others are positioned at all times; never fire from behind others; complete hunter education.						
11-10	Polk	42/M	Rifle	Deer	No	No
Comments: Shooter was unloading his rifle at the rear of his truck after hunting. Gun discharged through the rear of the vehicle, striking victim in lower back as he sat in the cab of the truck.						
Prevention: Always point muzzle in a safe direction; treat every firearm as if it is loaded; handle firearms carefully; complete hunter education.						

Source: Texas Department of Parks and Wildlife. Retrieved from http://tpwd.state.tx.us/publications/pwdpubs/media/pwd_rp_k0700_1124_2012.pdf

Note: Detail section of Texas hunting accidents report. This section of the report includes data about each hunting accident reported to the state.

The U.S. Consumer Financial Protection Bureau began operations in 2011, the year after it was established by the Dodd-Frank Wall Street Reform and Consumer Protection Act. Dodd-Frank sought to stem many of the mortgage, credit card and other consumer lending predatory practices in the 2000s. The bureau takes and investigates consumer complaints. Consumers can even use an online form, such as this one, for credit card complaints.



Source: Consumer Financial Protection Bureau. Retrieved from https://help.consumerfinance.gov/app/creditcard/ask.

Note: Consumer Financial Protection Bureau online complaint form.

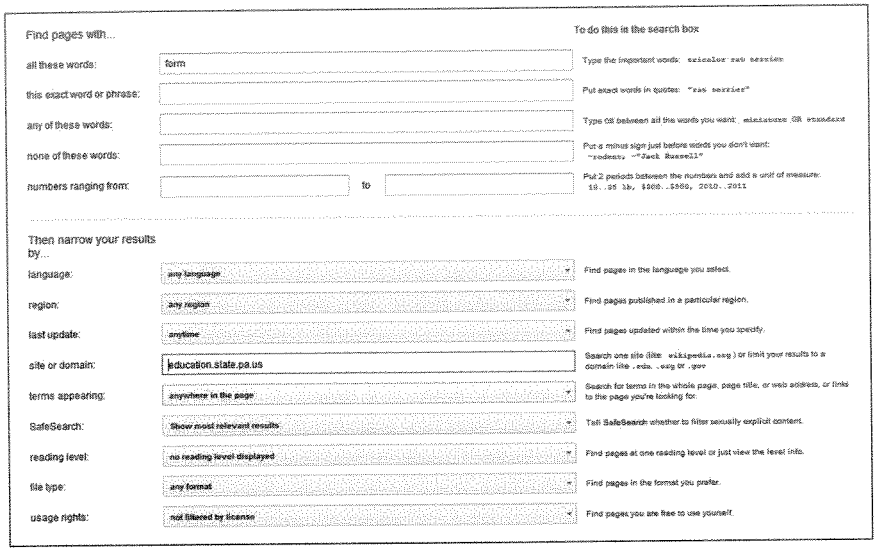
In the summer of 2013, the CFPB released a 19-page report that provides a snapshot of complaints and how they are handled. The report relies heavily on summarized numbers and charts to communicate data about the complaints. Also, the report tells us that data were derived from a database of the complaints (Consumer Financial Protection Bureau, n.d.a). So no guesses here!

TRICKS TO UNCOVER FORMS AND REPORTS

Finding forms and reports on government websites can be a challenge. Many times, agencies scatter them about rather than provide a centralized home for them, as the federal forms portal strives to do. Fortunately, we can use some Internet sleuthing tricks, using Google Advanced Search.

Point your browser to http://www.google.com/advanced_search and you'll see a user interface that is much busier than the serene, default Google search page. The advanced search allows us to limit our searches to specific websites, document formats or both. So this means we can look for PDF forms on government websites (those having

a .gov) domain, or even for forms on the Pennsylvania Department of Education website (education.state.pa.us). Go ahead and enter "form" in the Find pages with . . . all these words box. Then enter "education.state.pa.us" in the Then narrow your results by . . . site or domain box. Now search and look at your results.



Source: Google search.

Note: Google Advanced Search screen.

You can also use the advanced search to restrict your results to PDF or Word files by using the file type drop-down list.

After you get more experience running advanced searches, you can run them right from the main Google search box by mimicking the syntax that appears after you run the advanced search. For our query, this says, "form site:education.state.pa.us". Google translates that as instructions for it to look for the word "form" anywhere inside the education.state.pa.us domain. So if you want to look for forms on the EPA site, you'd type "form site:epa.gov". To get only PDF forms, try, "form site:epa.gov filetype:pdf".

Now that you are equipped with some tips for developing a data state of mind, we're going to tackle the world of online databases and develop some strategies for locating and downloading them. Being able to find relevant online data and download them is one of the key data literacy skills that you'll need to succeed.

ON YOUR OWN

Choose one of the two federal databases and find the law that led to the creation of the database: U.S. Food and Drug Administration's Operational and Administrative System for Import Support or U.S. Department of Education's Campus Safety and Security

database. Cite the law and write a few paragraphs explaining how the law made the database possible.

Every state has an agency whose mission is to collect and disclose political campaign finance data. Identify that agency in your state. Cite and summarize the law that gives the agency the authority to collect these data.

Find a form that a federal, state or local government agency uses to collect data. Specify where you found the form, and include a URL if you downloaded it from the Internet. What data does the agency collect with the form? Try to find a corresponding database online and provide its URL if you are successful.

CHAPTER 3 ONLINE DATABASES

In a perfect world, we would be able to easily download all of the data that we need from the comfort of our computer keyboards. All government agencies, from the federal down to the local, would post all of their public data on easy-to-find websites that have been well-indexed by search engines. The agencies would make their data available in formats that could be easily opened in our spreadsheet or other analysis and visualization programs. They would also provide copies of any documentation that we'd need to help understand the data, and make sure that the documentation is complete.

That perfect world, of course, does not exist and never will. Although it's impossible to know the exact percentage of data that are available, it's a safe bet that government agencies in general post less than half of their data online, despite the high-profile efforts that are part of the **open government** movement. Agencies may post the data in formats that are difficult to import into—or are too large for—spreadsheet programs. Agencies also may post data without providing the documentation that users need to understand them. There are all kinds of real-world challenges when it comes to finding the right data online and using them effectively.

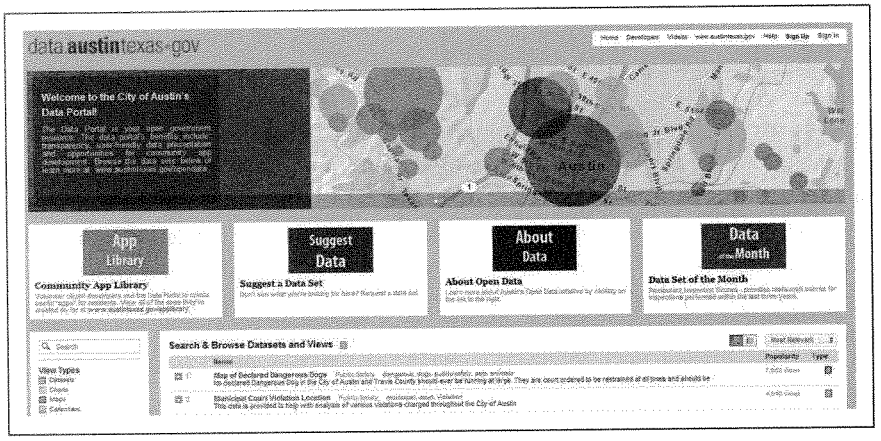
In this chapter, you will learn effective techniques for quickly finding, understanding and using data sets from government agencies. Because the Internet changes daily, no one can create a list of must-know sites that we'll be able to use ten or even five years from now. So the best approach is to understand how government agencies store data online and adopt the best practices that will help you find what you need in a reasonable amount of time. Also, when you find sites that you'd like to revisit, make sure you bookmark them so you can reference them easily later.

DESTINATION: DATA PORTALS

Government **data portals** have come into vogue with some agencies and members of the public, thanks to the **Government 2.0** or open-government movement. The movement got a big boost in May 2009, when the Obama administration launched Data.gov as a new public destination for federal government data sets. The federal government's effort was in part modeled on efforts of early innovators, such as the D.C. Data Catalog.

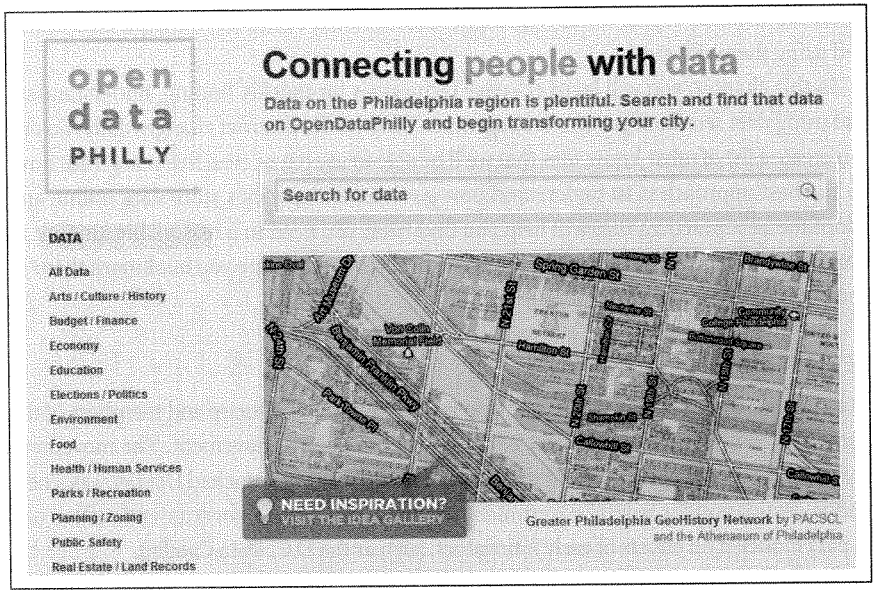
Sometimes, data portals can be frustrating to use. Agencies sometimes only link to existing data, or they post only a limited number of data sets that might not be useful.

Besides the federal government, many other government entities have since launched open data portals, such as Chicago; Austin, Texas; Montgomery County, Maryland; and Oregon. Some cities, such as Philadelphia, have partnered with nongovernment organizations to provide portals.



Source: Dangerous and Vicious Dogs. (n.d.). *City of Austin.* Retrieved August 16, 2013, from austintexas.gov/department/dangerous-and-vicious-dogs

Note: Data portal for city of Austin, Texas.

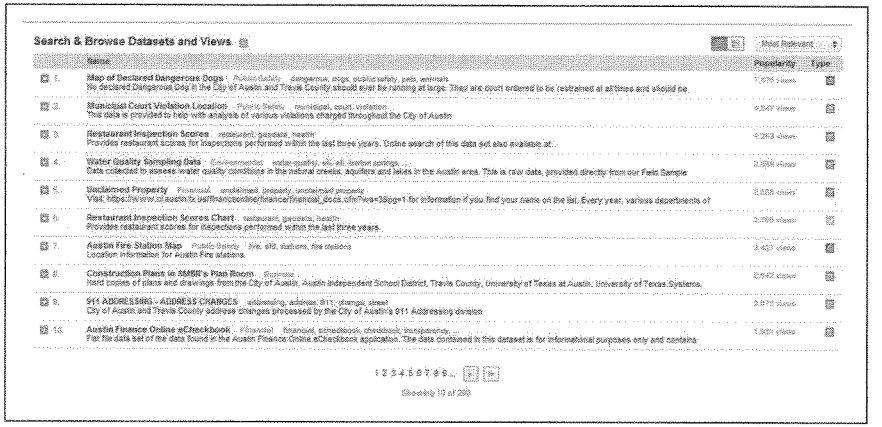


Source: Open Data Philly. Retrieved from http://www.opendataphilly.org/.

Note: OpenDataPhilly portal. The city of Philadelphia partners with a nongovernmental organization to make data available to the public.

Spend some time on the so-called **open data** sites and you notice some similarities in terms of the functions and appearances. That's because two major data-hosting Web platforms—**Socrata** and **CKAN**—dominate. Socrata is a Seattle-based company that sells its open-data platform services to government agencies, numbering at least three dozen in the summer of 2014 (Socrata.com, 2014). CKAN, on the other hand, is an open-source data catalog program that government agencies are allowed to deploy and use with no licensing costs. CKAN was developed by the Open Knowledge Foundation, a nonprofit organization based in the United Kingdom.

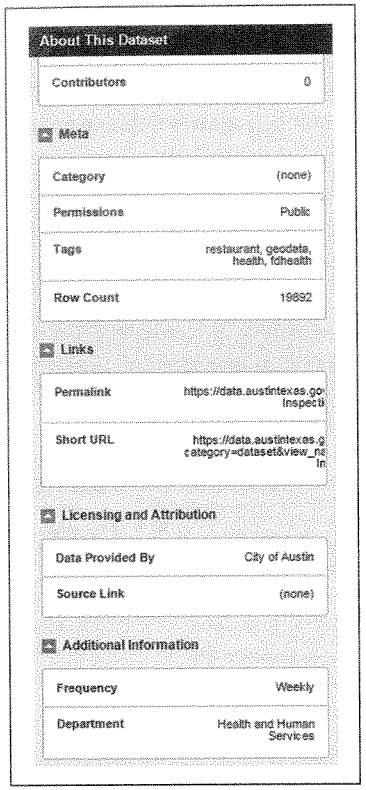
Let's explore one of the Socrata-powered sites: data.austintexas.gov. This is the official open data portal of the City of Austin, one that's pretty manageable in terms of size and ease of navigation. The bottom half of the landing page lists data sets posted by the city.



Source: Retrieved from Data.austintexas.gov.

Note: Austin, Texas, data set listing on portal.

After we click on the link for restaurant inspection scores, a list of the inspections appears. Click the About button at the upper right and we get information about the data that will help us understand them better. For instance, we see that the data are for inspections dating back three years, that they have 19,964 rows and that they are updated weekly by the city Health and Human Services Department.



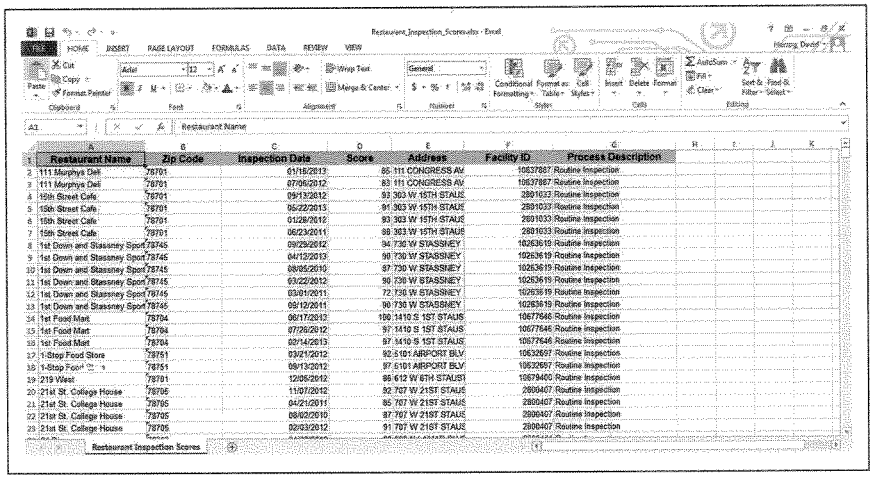
Source: Retrieved from https://data.austintexas.gov/dataset/Restaurant-Inspection-Scores-Chart/hqa6-stx4

Note: About box on Austin data portal. This box provides some important information about the data provided.

Before we download any data, we need to practice safe computing and document our work: Create a word processing document or text file with the name **Data_Notebook**; this is going to be where we take notes about the data we get and what we do with them. Enter the date, then "Austin restaurant inspections" and the URL "https://data.austintexas.gov/dataset/Restaurant-Inspection-Scores/ecmv-9xxi". Don't forget to include details about the data, such as the number of rows, time span, update frequency and source.

Now, on to downloading the restaurant inspections file in Excel format onto our computers. Click the Export button, then Download as XLSX to get the most current Excel file format. It may take a few seconds to download the file, depending on your Internet connection and computer processor speed. Look for the downloaded Restaurant_Inspection_Scores.xlsx file and then open it in Excel. Now we have a copy of the file that we could analyze.

Let's practice safe computing again and make sure we downloaded all of the data: We can see that we downloaded all seven columns. To check the rows, hold the Ctrl and End keys on your keyboard

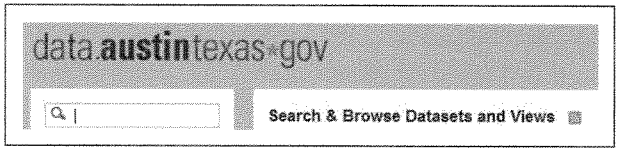


Source: Retrieved from https://data.austintexas.gov/dataset/Restaurant-Inspection-Scores-Chart/hqa6-stx4.

Note: Austin restaurant inspection data downloaded to Excel file.

together at the same time and Excel takes you to the bottom right corner of the spreadsheet (Mac users, use the Command and End keys). We see that we have the same seven columns as at the top, as well as 19,965 rows (one for the headers and 19,964 for the data, just as noted in the About box).

Open data portals also allow us to search. Let's say you're an intern for a local social services agency that wants to better understand affordable housing in Austin. Use the search box at the left of the data.austintexas.gov page to look for affordable housing. The results show us that the city indeed does offer data about affordable housing.



Source: Retrieved from Data.austintexas.gov.

Note: Austin data portal search box.

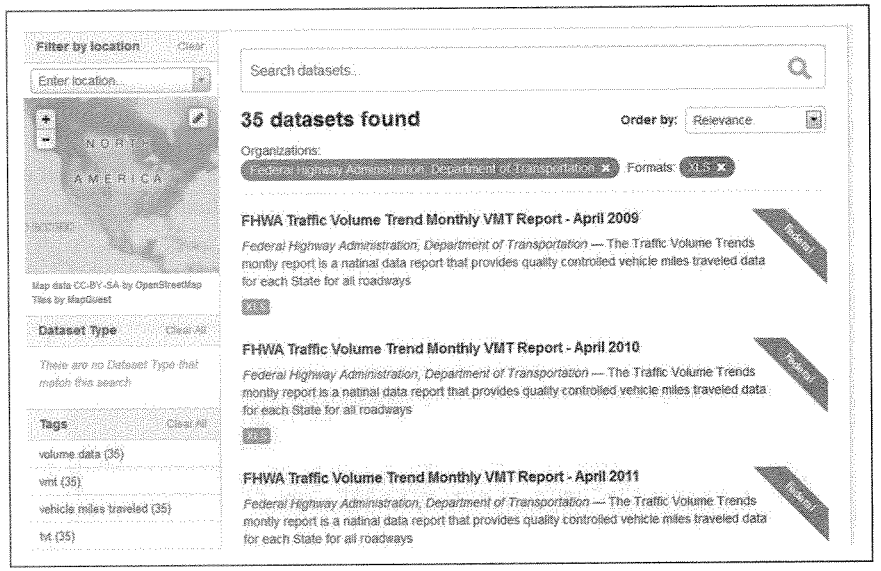
Now let's go to the data portal for the federal government, Data.gov, which is built on the CKAN open-source platform. There are no data directly on the home page, but users can get to the data via the topic icons or the search box.

Source: Retrieved from Data.gov.

Note: Data.gov home page. The federal government redesigned the home page in 2014 to allow users to more easily search or browse for data.

The easiest point of entry is through the Data tab, which takes us to a list of data sets that we can browse, filter or search. In the summer of 2013, Data.gov cataloged more than 161,000 data sets. That number might seem impressive, but it is a little deceptive and overstates the real number. Many of the data sets that Data.gov links to are extracts of larger databases. For example, the EPA provides links to download Toxic Release Inventory data by state (or federal territory) and year. So the EPA considers the 2009 TRI data for Rhode Island as one data set, even though those data are part of a national database covering many years.

One of the best ways to look for data is to use the powerful filtering tools that are part of CKAN. On the left sidebar of the data set page we can filter by data set type, tags, formats, groups, organizations (or specific agencies) and community categories. We can add filters to further zero in on the data we want. Click on Federal Highway Administration in the Organization Filter and get just the data sets posted by that division of the U.S. Department of Transportation. Data.gov updates your results, showing you the new number and



Source: Retrieved from http://catalog.data.gov/dataset.

Note: Search results for Excel files from the Federal Highway Administration on Data.gov.

that you have filtered for the FHWA. Filter more by picking XLS for Excel under the formats filter. The number of results is even smaller and you now have two filters displayed.

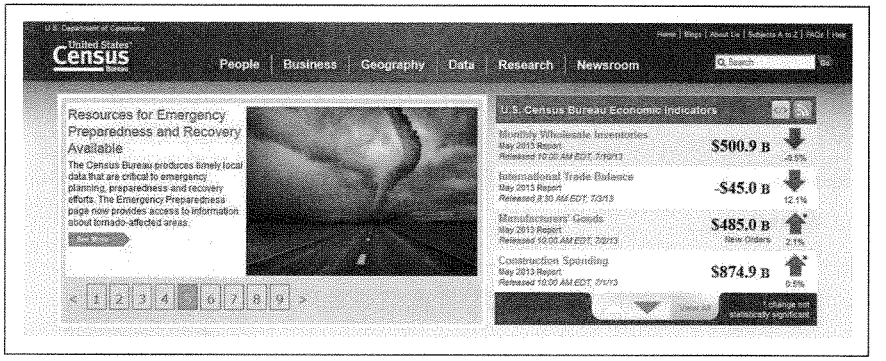
Clear the filters by clicking the Xs next to them both.

Searching is pretty easy, too. Just type a search term in the search box and then click on the magnifying glass and you'll get data sets that meet your criteria. At this point you can also apply filters. Let's say we want to find data that might tell us about the declared disasters that have occurred in our state. So go ahead and filter for just data that come from the Federal Emergency Management Agency, which is part of the U.S. Department of Homeland Security. Some of these data, including the Excel file of declared disasters, could be promising.

STATISTICAL STOCKPILES

Another great way to learn more about online data is to tap into the websites of government agencies whose mission is to provide statistical data. On the federal level, this includes the Census Bureau, the Bureau of Justice Statistics, the Bureau of Labor Statistics, the National Center for Education Statistics and the Bureau of Transportation Statistics. On the state level, you might find similar agencies, such as the Missouri Economic Research and Information Center.

The U.S. Census Bureau is one of the biggest data providers in federal government. In fact, its mission is to provide quality data about people and the economy. Many of us know the Census Bureau because of the **decennial census**, which is an attempt every ten years to count every person in the United States for the purpose of apportioning seats in the U.S. House of Representatives.



Source: Census Bureau, Department of Commerce. Retrieved from http://www.census.gov.

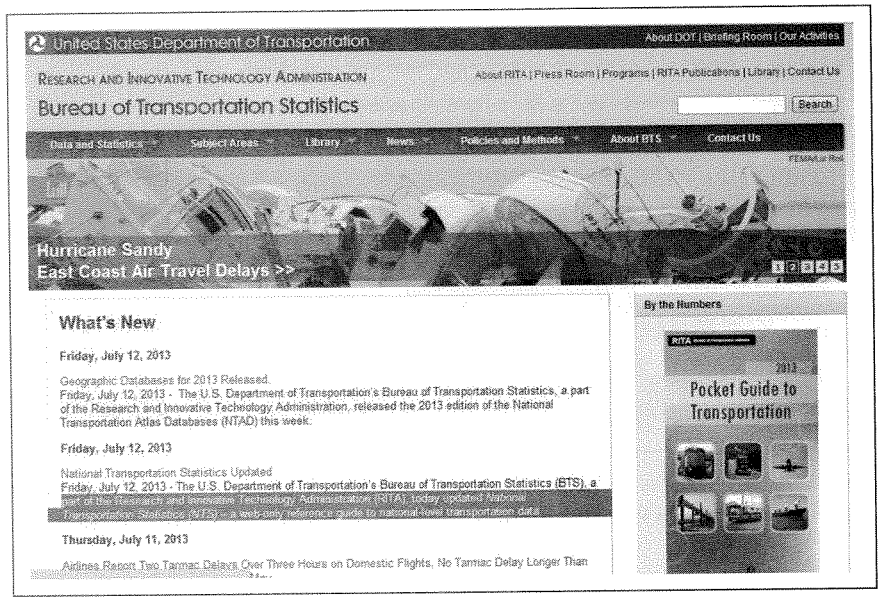
Note: Census Bureau Homepage. The Census Bureau generates a huge stockpile of data about people, business and government.

But that's just a sliver of what's available from the Census Bureau. We can find demographic data that are much more detailed, as well as data about construction spending, retail trade, automobile registrations and home ownership. Dig deeper and you will find data about government employment, payrolls, debt levels, assets and

budgets. Browse the navigation tabs on the Census Bureau website to get an idea of the data it collects and makes available. The bureau's American FactFinder provides an interactive tool that allows users to build customized tables from the decennial census and the American Community Survey, a data set with more demographic detail. The trade-off is that the ACS is based on a sample of the population and is not as statistically reliable as the decennial census.

The Bureau of Justice Statistics is a division of the U.S. Department of Justice and has published dozens of data sets about crime, justice and law enforcement (Bureau of Justice Statistics, n.d.). Criminologists, policy analysts and social scientists analyze these data to help identify trends in incarceration, hate crime, identity theft and human trafficking. The BJS regularly publishes reports that are based on its own analyses of the data, such as one in June 2013 that looked at indicators of school crime and safety in 2012 using the National Crime Victimization Survey (Snyder and Truman, 2013).

The Bureau of Transportation Statistics is part of the U.S. Department of Transportation and compiles data sets about air, highway, rail and waterway travel. It even has data about Canadian and Mexican border crossings and oil and gas pipelines. Commercial



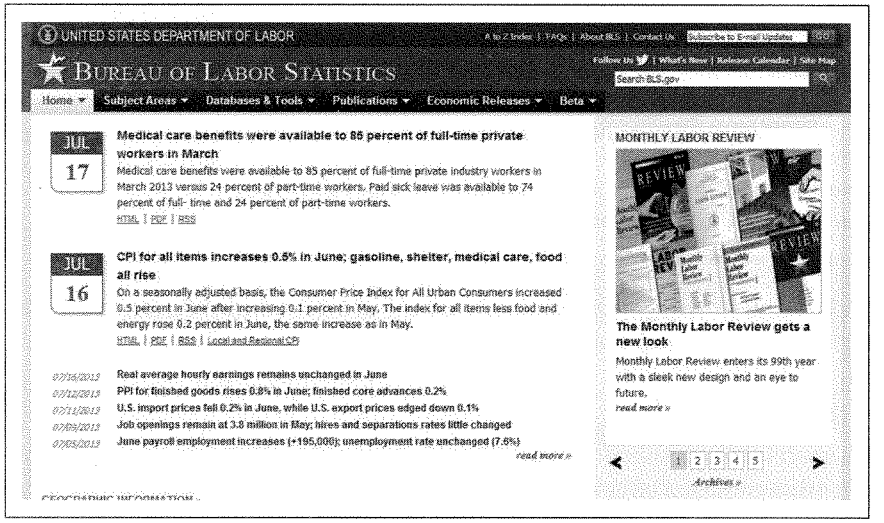
Source: Bureau of Transportation Statistics, Department of Transportation. Retrieved from http://www.rita.dot.gov/bts/.

Note: Bureau of Transportation Statistics website.

airport operators, journalists and airlines use BTS data to evaluate the cost of flying, on-time performance and general airport activity.

BTS's own researchers and others have used the data to create reports about drunken driving, container port activity and the impact of the 9/11 attacks on U.S. travel (Bureau of Transportation Statistics, n.d.).

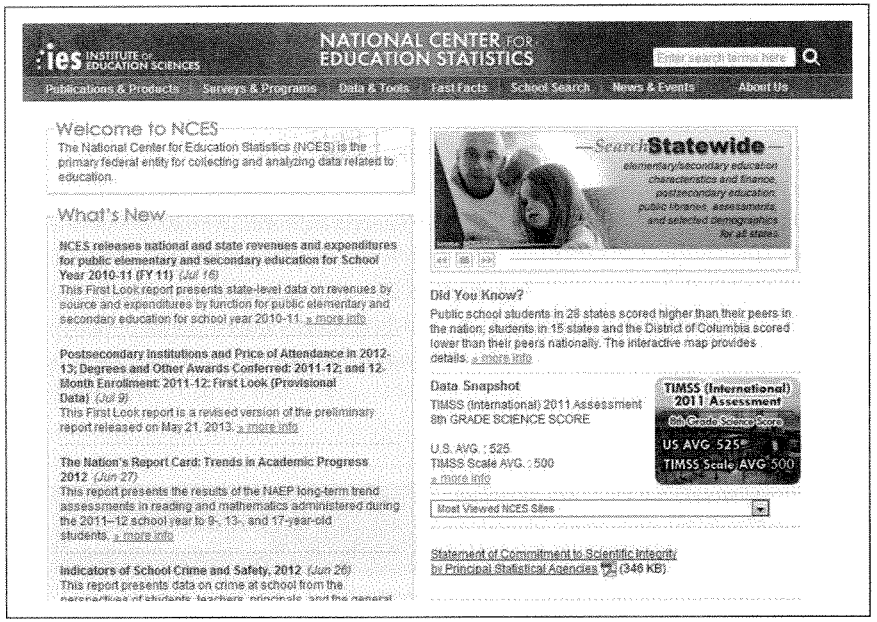
For economic and employment statistics, the U.S. Department of Labor's Bureau of Labor Statistics is one of the best resources. BLS collects authoritative data about employment, the labor force, compensation, mass layoffs and inflation (through the Consumer Price Index and other indexes). In fact, the BLS data are the source of the government's official numbers for unemployment and inflation. Economists and economic development professionals use these data on the job. The best way to find the data is to navigate by either subject area or database name.



Source: Bureau of Labor Statistics, Department of Labor. Retrieved from http://www.bls.gov/home.htm.

Note: Bureau of Labor Statistics website.

Data about early childhood, primary, secondary, higher and adult education can be found on the National Center for Education Statistics site, run by the U.S. Department of Education. The center is a trove of data about costs, enrollment and crime at universities and other higher education institutions. It also provides data about student to teacher ratios at public schools. Recent reports using the center's data include looks at private and



Source: National Center for Education Statistics, Department of Education, http://ies.ed.gov/.

Note: National Center for Education Statistics website.

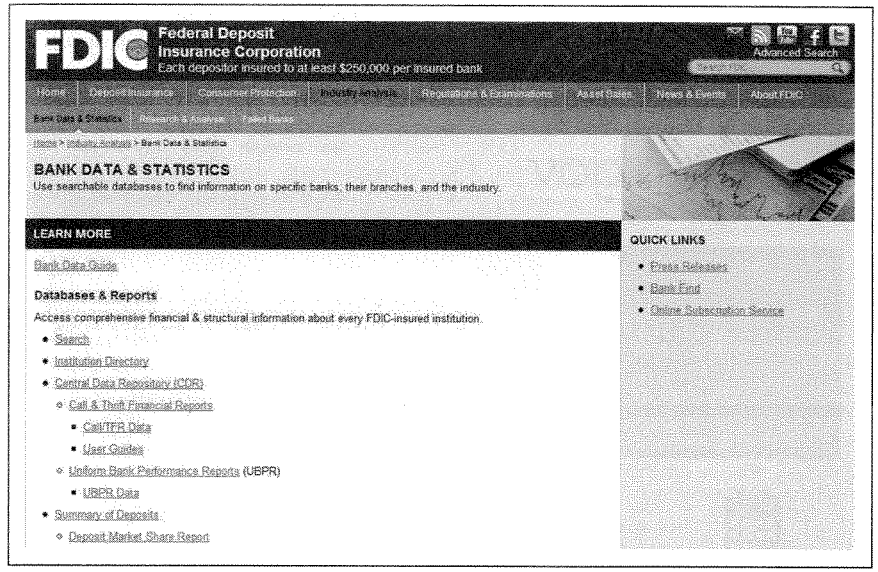
public schools, and the most popular majors for bachelor's degrees (National Center for Education Statistics, n.d.).

AGENCY SITES

It's smart to become familiar with how agencies store data on their own sites, because only some data are hosted on or linked to data portals. This is where the hunt for data can get challenging because agencies sometimes scatter data around. Use website navigation tabs to look for words like Data, Open Data, Transparency and Statistics.

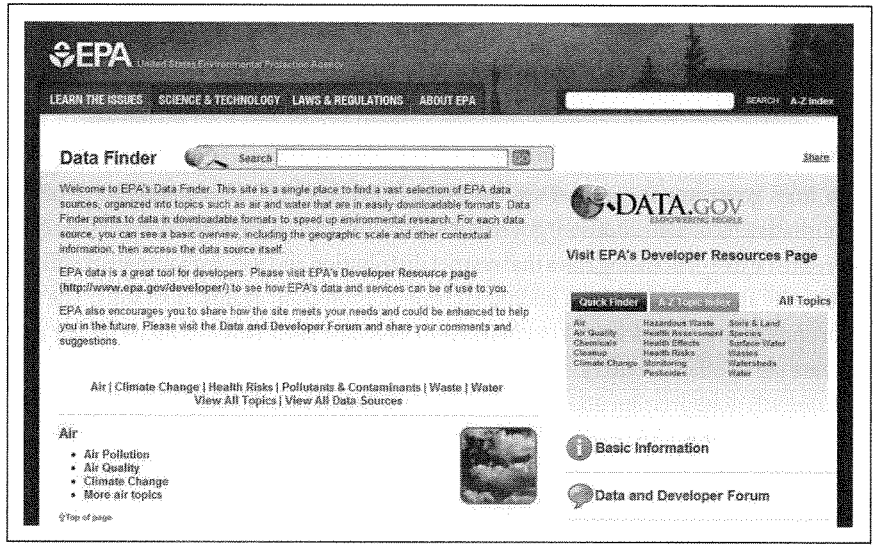
The Federal Deposit Insurance Corporation, for example, links to its data through its Industry Analysis navigation tab, then Bank Data & Statistics. The data include the Summary of Deposits, which has data about the amount of deposits taken in at each bank branch location. Marketing analysts and regulators use the data to examine the level of competition within banking markets.

See if the agency has its own data portal or page by typing the following in your browser: www.agencyname.gov/data. For example, entering "www.epa.gov/data" takes you to the EPA's Data Finder page.



Source: Federal Deposit Insurance Corporation. Retrieved from https://www.fdic.gov/bank/statistical/index.html.

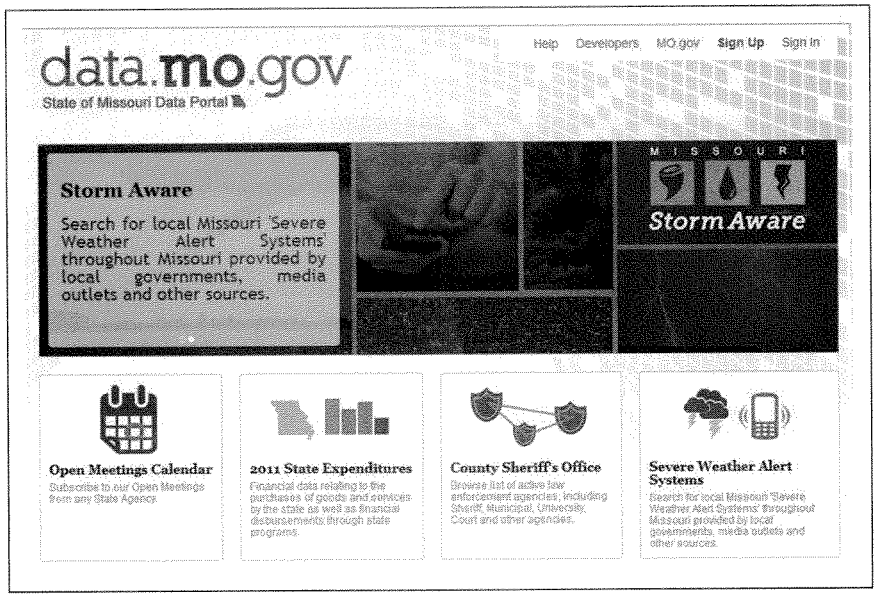
Note: Federal Deposit Insurance Corporation data about banks.



Source: Environmental Protection Agency. Retrieved from http://www.epa.gov/data/.

Note: Environmental Protection Agency data page. Try to find agency data pages by typing "/data" after the agency's main URL.

Or you can try typing data.agencyname.gov. To look for the Missouri state open data portal, you would type "data.mo.gov".



Source: Retrieved from Data.mo.gov.

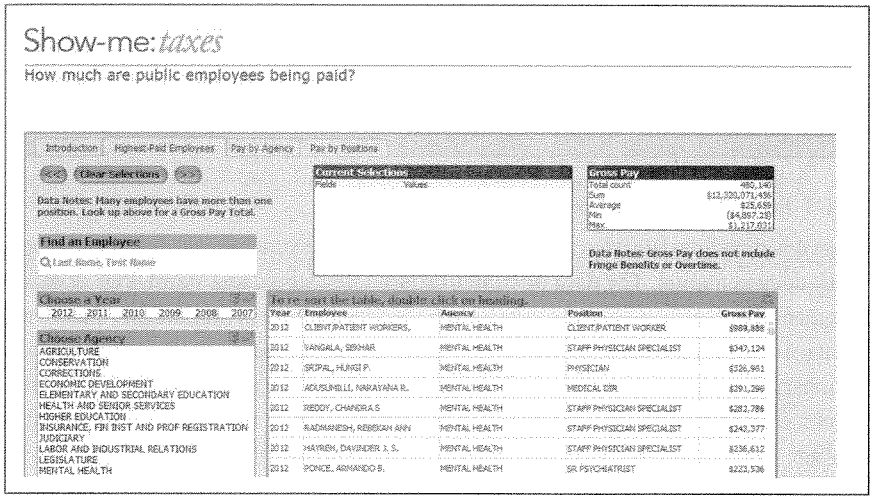
Note: State of Missouri data portal. Try to find data portals by typing "data" instead of "www" in a URL.

NONGOVERNMENTAL RESOURCES

Data from government agencies are used widely because they are considered to be an authoritative record. In addition, government data (at least in the United States) are considered to be free of licensing rules that restrict distribution and use. For journalists, official documents and data provide an additional benefit: protection against libel suits. The fair report privilege provides a shield to journalists who base their news accounts on fair and accurate use of official sources.

However, nongovernmental organizations (or NGOs) also offer online data that we can use for our analyses. The Right to Know Network, which we visited in Chapter 2, posts several databases from the EPA. The data are from the EPA, but are distributed by the Center for Effective Government through the RTKNet site. Aside from promoting government transparency, the center advocates for progressive revenues (Center for Effective Government, 2013) or higher tax rates for people with higher incomes (and vice versa). On the other side of the political spectrum, Missouri's Show-Me Institute posts salary data for public employees that site visitors can download into an ASCII text file. The institute is a

nonprofit organization that promotes free market and libertarian solutions to social and governmental problems. The institute gets its payroll data from the State Department of Administration.



Source: Show-Me Institute. Retrieved from http://www.showmeliving.org/payroll.

Note: Missouri state data posted on the nongovernmental Show-Me Institute's website.

Before using data from a nongovernmental source, ask,

- How did the NGO obtain the data?
- What did the NGO do to process or change the data?
- What interest does the NGO have in making the data available?
- Can you get the data directly from the official government source instead?

Some NGO sites collect data that are not from governmental sources. The Roper Center at the University of Connecticut and the Gallup Company both offer access to archives with public opinion polling data. The Inter-university Consortium for Political and Social Research at the University of Michigan archives data from social science researchers. If your college or university is a member of the consortium, you'll be able to download the data, assuming your computer is connected to your institution's network (Inter-university Consortium for Political and Social Research, n.d.).

DATA SEARCH TRICKS

Just as we used Google Advanced Search to uncover clues about data, we can use it to find data files that we can download. The trick is to think of words that appear on webpages

with downloadable data and plug those into our search form. Let's revisit the EPA's data page at http://www.epa.gov/data for some ideas. We see that "downloadable" and "data" appear multiple times, which leads us to believe those terms will work well for our search. Go ahead and try "downloadable data site .gov" to look for pages with "downloadable" and "data" on websites that have .gov extensions. Your results list should have a mix of federal, state and local websites. Of course, you could try to narrow your search further: for example, if you wanted to find only Excel files add "filetype:xls". Try different search terms, such as "download data," and see what happens.

DON'T FORGET THE ROAD MAP

When downloading data from a government website, make sure you also download a copy of any documentation. The **data documentation**, often stored as a Word, PDF or other document file, usually is essential in helping understand what's in the data file. There's no standard name for the documentation, so it might be called record layout, file layout, data dictionary or something entirely different. If you can't download the documentation, try to get a copy of it from someone at the agency, using contact information that's on the website.

Whatever the name, the documentation usually provides some key pieces of information about the data set:

- Table names, along with record counts for each table.
- Column or field names in each table, along with a field description, type of field (character, number, date, etc.) and width.
- Codes and their meanings. Data are often stored in codes, so the documentation should explain these.

For example, the Federal Aviation Administration routinely collects data about licensed pilots and releases some of these data in the Airman Directory Releasable File. On the same webpage where the FAA releases the data in two different text file formats, it also provides documentation for each (Federal Aviation Administration, n.d.). This example from the nine-page documentation for the **CSV** (comma-separated value, or delimited text) file says we have a table called Pilot Basic, which has 13 columns or fields, starting with UNIQUE ID and ending with MEDICAL EXPIRE DATE. All of the fields are formatted as A, or alphanumeric. (**Alphanumeric** data use characters to represent numbers and text.) The lengths tell us how many characters can be stored in each column. The remarks provide information about codes used (see the remarks for MEDICAL CLASS) and that the dates are stored in the MEDICAL DATE and MEDICAL EXPIRE DATE columns as MMYYYY (or month, month, year, year, year, year).

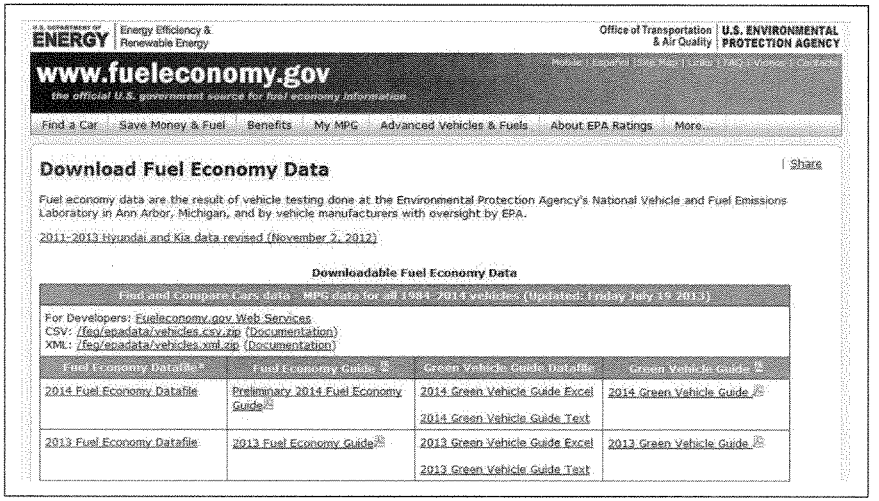
FIELD NAME	FORMAT	LENGTH	REMARKS
Pilot Basic record file			
UNIQUE ID	A	8	1st position = 'A' or 'C' followed by a 7-digit number
FIRST & MIDDLE NAME	A	30	
LAST NAME & SUFFIX	A	30	
STREET 1	A	33	
STREET 2	A	33	
CITY	A	17	
STATE	A	2	Blank if foreign address
ZIP CODE	A	10	
COUNTRY-NAME	A	18	
REGION	A	2	
MEDICAL CLASS	A	1	1=First 2=Second 3=Third (Certificate Type "P" only)
MEDICAL DATE	A	6	MMYYYY (Certificate Type "P" only)
MEDICAL EXPIRE DATE	A	6	MMYYYY (Certificate Type "P" only)

Source: Airmen Certification Database. (n.d.). FAA: Home. Retrieved July 22, 2013, from http://www.faa.gov/licenses_certificates/airmen_certification/releasable_airmen_download/

Note: File documentation for the Federal Aviation Administration's airmen database.

DOWNLOADING, UNZIPPING AND INSPECTING DATA FILES

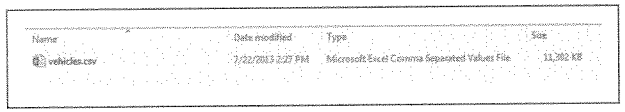
In the final part of this chapter, we are going to download, unzip and inspect a delimited text file. We'll work with fuel economy data from the U.S. Department of Energy kept at http://www.fueleconomy.gov/feg/download.shtml. Under Find and Compare Cars Data, download the CSV file to a location on your computer where you can easily retrieve it. Then click on the Documentation link, which is our guide to understanding the contents of each column. (Make sure you record the details about the data in your data notebook. Do this for all of the data you download.)



Source: Department of Energy. Retrieved from http://www.fueleconomy.gov/feg/download.shtml.

Note: Federal government website for vehicle fuel economy data.

Find the file using Windows Explorer (or Finder if you are using a Mac). Right-click on the file and pick Extract All . . . to launch the Windows unzipping utility. Change the destination of the extracted files if you wish and then click Extract. You now have a folder called vehicles.csv that contains a file of the same name. The compressed files usually are called **zip files** and have a .zip extension.



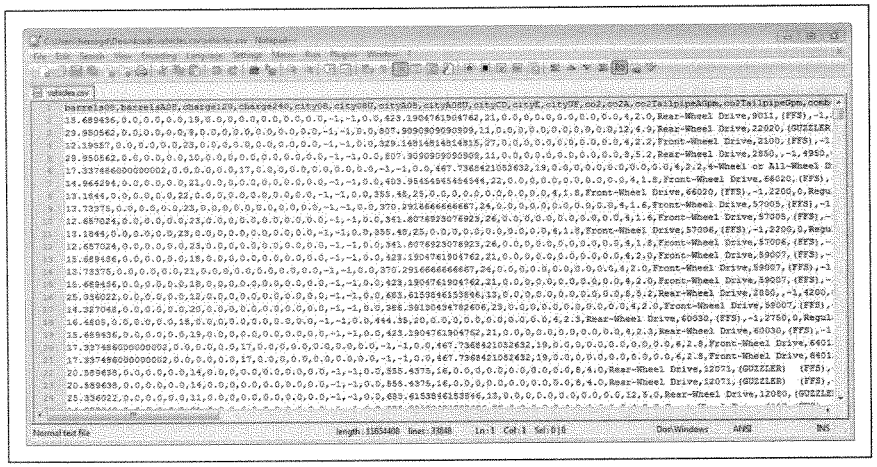
Source: Department of Energy.

Note: Comma-separated values file containing data about vehicle fuel efficiency. Windows computers usually display CSV files as Excel files.

Windows identifies the file as a Microsoft Excel Comma Separated Values File. Because the .csv file extension is associated with Excel, you can double-click on the file to open it. Use Ctrl-End to navigate to the end of the file; you'll see it has 33,847 rows. Close the file now and don't save any changes, if prompted.

Vehicles.csv is a text—not an Excel—file, so we can view it in a text editor program. On Windows, we'll use Notepad++, which is a free and open-source program that we can download from http://notepad-plus-plus.org/. Notepad++ has many more features than the stock Notepad that comes with Windows; it can open larger files than Notepad can, also. (Mac users, install TextWrangler, which is a free program available at http://www.barebones.com/products/textwrangler/.)

Open vehicles.csv with your text editor and you should see something like this:

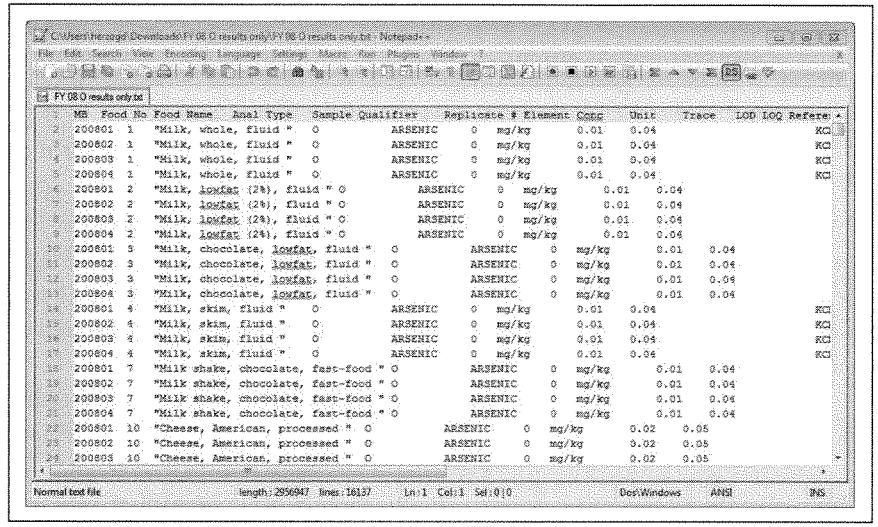


Source: Department of Energy.

Note: Comma-separated values in a text editor. Note the commas, which are used for column delimiters, and the compressed data.

It looks like a mess, as if someone took our data and squished them together. Actually, this is how comma-delimited text files are supposed to look. The first line contains the labels for our columns. Each comma denotes a column break to Excel and other programs. The first line of data starts on the second line. Use Ctrl-End to navigate to the end of the Notepad++ file and you'll see we have 33,847 rows—the same number we did when we opened it in Excel.

Now, let's download, unzip and inspect some tab-delimited data that are produced by the Food and Drug Administration for its Total Diet Study, which is supposed to monitor the levels of contaminants in food. The Total Diet Study was started in 1961 as a program for detecting radiation. Since then, it has expanded to look for the presence of pesticides and industrial chemicals in food (Food and Drug Administration, n.d.). Navigate to the analytical results page at http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm184293.htm and download the O 2008 file under the Elements heading. Find the file named FY 08 O results only.zip and unzip it. You now have a text file called FY 08 O results only.txt. Excel does not recognize the .txt extension as a native file, so we can't just click on it to open it. Instead, open it using your text editor. The file should look something like this:



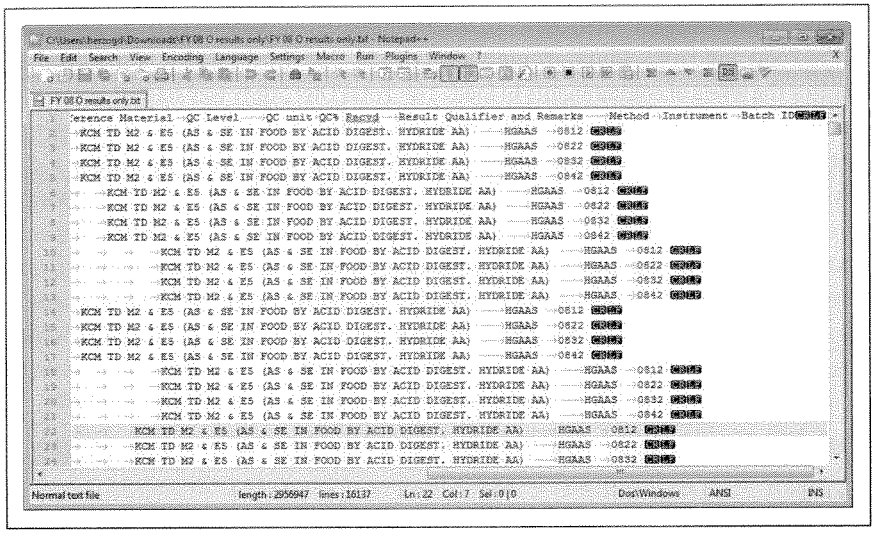
Source: Food and Drug Administration.

Note: At first glance, no delimiters are in this delimited-data file.

Again, because this is a delimited text file, it looks like a mess—with squished-together data. We notice that the first row contains the column headers. The data themselves start in the second row. Also, some of the data have double quotation marks. These are called **text qualifiers** and are sometimes used in delimited text data to denote

text that should be kept intact inside a column. Single quotation marks also can be used as text qualifiers.

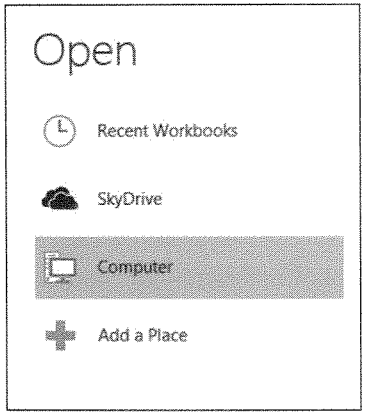
But where are the delimiters—the tabs? Tabs are considered ASCII characters, but they are hidden, so we normally don't see them. In Notepad++, we can make them appear by going to View | Show Symbol in the menu, then selecting Show All Characters.



Source: Food and Drug Administration.

Note: Tab characters used as delimiters, no longer hidden.

Aha! Now we can see reddish arrows. Those are the tabs that serve as the column delimiters. We also see some periods, which represent spaces. Let's now scroll over to the right side of the file and we can see two black blocks with white text. They say CR and LF and mean carriage return and line feed. These are hidden ASCII characters that are used to denote an end of the data record.



Source: Microsoft Excel 2013 for Windows.

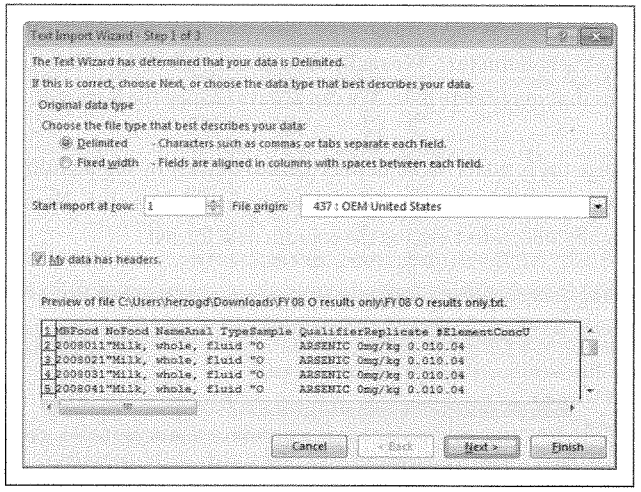
Note: Opening files in Excel

One last check: Let's use Ctrl-End to see how many lines of data we have. We should see 16,136, including header row. Close the text file now.

Time to open the file in Excel. Start Excel and then Open Other Workbooks in the pane on the left. Select Computer and then use the Browse button to the right to look for the extracted text file.

We won't see any files because Excel is looking only for native Excel files. Change the File type option at the bottom right to All Files and the text file appears. Open it, and Excel launches a text import wizard that walks us through the process of opening the file.

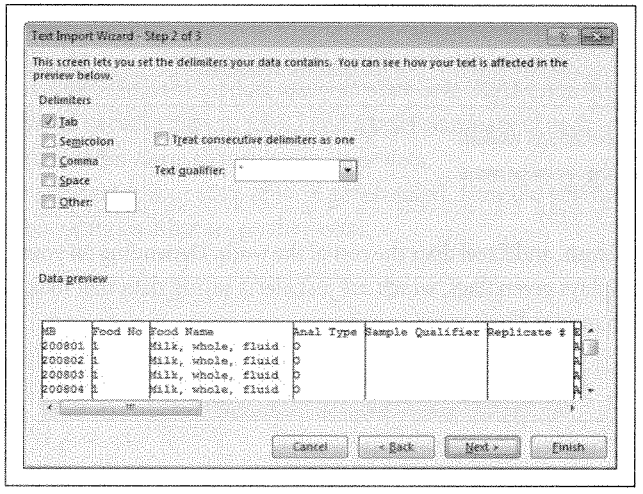
In Step 1, we need to tell Excel what kind of text file—delimited or fixed-width—we're importing. Excel has correctly guessed delimited. (Excel sometimes gets this wrong, so make sure to check this.) Our data have headers in the first row, so we need check the box that says My data has headers and start the import at Row 1 to capture these data.



Source: Microsoft Excel 2013 for Windows.

Note: Step 1 in the Excel text import wizard.

In the next step, we need to specify the delimiter character, so make sure the box for Tab is checked. (If we need to specify a character other than any of those listed, just check the Other box and enter the character in the box next to it.) At this step, we also need to tell the wizard that we're using double quotation marks as text qualifiers.

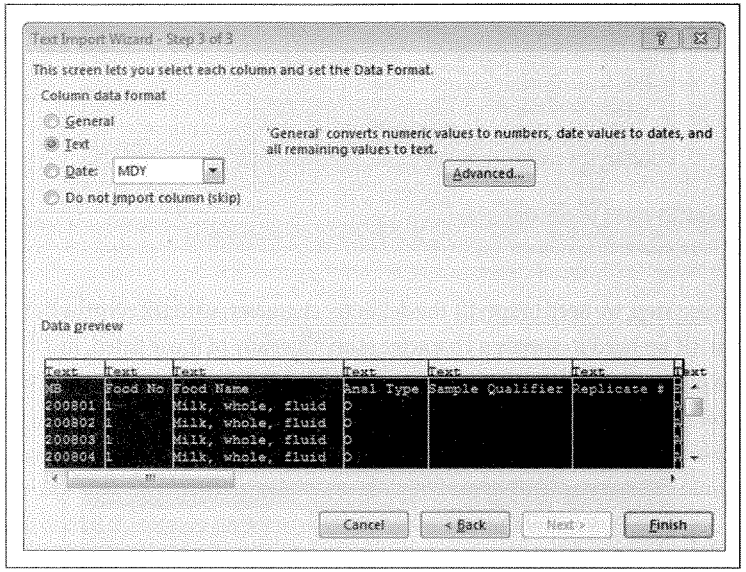


Source: Microsoft Excel 2013 for Windows.

Note: Step 2 in the Excel text import wizard.

Last, we need to tell Excel what data types to apply to each column. The default often is General, which means Excel takes a look at the contents of each column and makes an educated guess about the format. This can be a problem because we might have a column that includes zip codes, some of which have leading zeros. Take 02818 in East Greenwich, Rhode Island, for example. If Excel imports this as general, it will treat it as a number and lop off the leading zero, leaving a zip code of 2818. Fortunately, we have a simple workaround: making the column text. We really should make all of the columns text to avoid any import flubs that may delete data. We can always change the format of the column in Excel later.

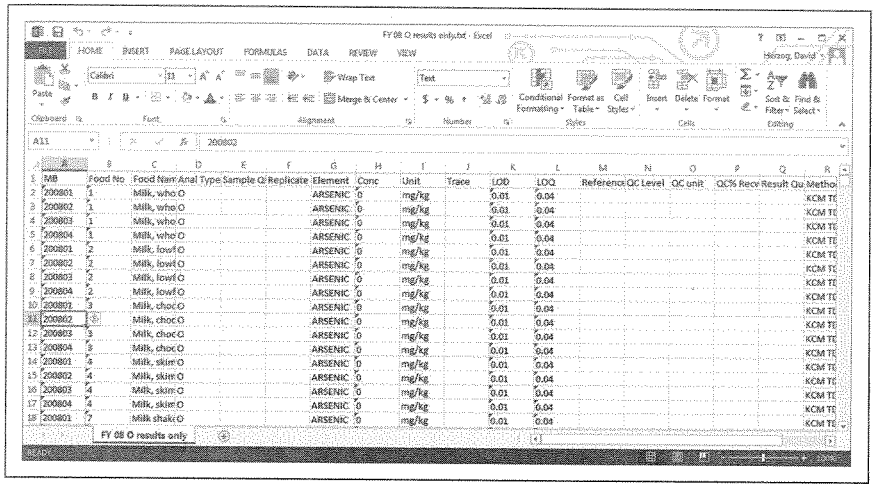
We can change the formats for all of the columns quickly by highlighting the first column, then scrolling to the final column and selecting it with Ctrl-click. Now that all of the columns are highlighted, select Text as the column data format.



Source: Microsoft Excel 2013 for Windows.

Note: Step 3 in the Excel text import wizard.

Now click Finish, and Excel does the rest of the work. Depending on your software settings, Excel displays green flags, which are just alerts saying data that look like numbers have been formatted as text.



Source: Food and Drug Administration.

Note: Text file successfully imported in Excel.

We haven't yet saved this as an Excel file, so let's do that by clicking on the File tab, then selecting Save As. Browse to a location where you are storing your files, and select Excel Workbook (*.xlsx) as the file type. That does it!

In this chapter we've covered a lot of ground, starting with identifying useful data on the Internet. We also learned how to download, unzip and inspect ASCII files in a text editor.

In the next chapter, we'll turn our attention to how we can learn about all of the data that government agencies keep offline and how we can obtain them informally or formally through open-records requests.

ON YOUR OWN

Use Google Advanced Search to find data that your state or local government keeps online. Give three examples of data that you found and a brief summary of each example. Also, provide the search syntax you used to uncover these data.

Find the data portal for your state or local government. How did you find it? Provide summaries of three data sets that are on the portal.

CHAPTER 4 IDENTIFYING AND REQUESTING OFFLINE DATA

As we saw in the previous chapter, government agencies are posting a growing number of data sets on the Internet. However, most governmental data sets are held offline for a variety of reasons, such as lack of financial or human resources, political sensitivity and the perception that the public might not be interested in the data. Finding online data can be tough, but it can be even more challenging to identify and successfully obtain databases that government agencies store offline. In this chapter we will learn some approaches for further developing a data state of mind to become more aware of these offline data sets. We'll also learn how to use federal and state open-records laws to make formal requests for these databases, which are public information. Knowing the name of the database and how it's kept is key to making a successful request using the federal Freedom of Information Act or similar state laws.

In Chapter 2, we learned that some government agencies collect and use data because they're required to by law or because it helps them meet their strategic goals. We learned that we can uncover clues about databases by looking for data entry with handheld computers or with paper and electronic forms and in statistical reports. In Chapter 3, we learned about data documentation. We can use all of these clues to help uncover offline databases.

OTHER CLUES FOR OFFLINE DATA

Some tools are particularly well suited for helping identify offline data. These are resources such as records retention schedules, audit reports, federal agency major information system listings and nongovernmental organization sites.

Records retention schedules provide details about records held by state and local government agencies, and guidelines about how long those agencies need to keep the records. Each state issues its own records retention schedules for state records, and sometimes issues a separate schedule for local records. Unfortunately, there is little uniformity from one state to another about what's in the schedules and how to access them. The Georgia Secretary of State provides a lookup form for its schedule. The Florida Department of State offers the information in PDF, Word or Excel format. The Utah Department of Administrative Services offers the information as webpages that you need to click

through until you've found what you're after. The example below shows that Utah state agencies are supposed to keep records about the sale of state-owned real estate for six years. The entry fails to mention whether any of the information is available as data, but at least we know there's a possibility.

REAL PROPERTY SALE FILES (ITEM 14-3)

Records which document the transfer of state owned real estate to non-state ownership, whether by transfer, trade, sale, or donation.

RETENTION

Record copy: Permanent. Retain by agency for 6 years after a deed of sale is recorded and then transfer to State Archives with authority to weed.

Duplicate copies: Retain until administrative need ends and then destroy.

SUGGESTED PRIMARY DESIGNATION

Public.

(Approved 07/90)

Source: Utah Department of Administrative Services. Retrieved from http://archives.utah.gov/recordsmanagement/grs/stgrs-14.html.

Note: Entry from Utah's records retention schedule.

Most records retention schedules do not distinguish whether a record is kept in a database or on paper. The schedules for Texas, Delaware, Ohio and New York State are some of the exceptions, in that they mention computer files.

Find records retention schedules by searching or by going to the Council of State Archivists website, which has links for state (http://www.statearchivists.org/arc/states/res_sch_genlst.htm) and local records (http://www.statearchivists.org/arc/states/res_sch_genlloc.htm) schedules.

Audit reports generated by state or federal authorities examine the operations of government. The Government Accountability Office, which is the investigative arm of Congress, monitors the performance of agencies and issues in-depth reports with results and recommendations for improvement. The reports can easily run dozens of pages, but you can get to the meat quickly by reading, in the full PDF report, the Summary of Findings or What GAO Found section, as well as the Scope and Methodology. If a GAO auditor reviewed any databases for the report, he or she will list those in the Scope and Methodology. In the states, auditors who are elected in statewide elections conduct audits and generate similar reports we can use for getting clues about data.

Federal inspectors' general reports might also be helpful. Inspectors' general offices are independent units within federal agencies and are supposed to investigate possible fraud, waste and other wrongdoing. Look for references to databases in these reports, too.

For example, this audit report from the U.S. Department of Agriculture examines the Farm Assistance Program payments by the Farm Service Agency in the federal government's 2012 fiscal year. The inspector general says the FSA failed to adequately document payments. In the Scope and Methodology section of the report, the inspector general noted that it used a sample of program payment data for its audit (Department of Agriculture, 2013). So we know that there's a database of payments that we could request.

Scope and Methodology

We obtained the universe of payments from FSA and statistically selected 80 payments made from October 1, 2011, through September 30, 2012.[12] FSA provided four data extracts that included program payments for all of fiscal year 2012, totaling $759 million. FSA made the payments through 11 programs.[13]

Source: Fiscal Year 2012 Farm Service Agency Farm Assistance Program Payments. (n.d.). U.S. Dept. of Agriculture. Retrieved July 24, 2013, from www.usda.gov/oig/webdocs/03401-0002-11.pdf.

Note: Methodology from federal inspector general report. The methodology shows that the inspector general obtained four data extracts for its examination.

Governmental and nongovernmental websites have a lot to say about data that are kept offline by agencies. Under the Electronic Freedom of Information Act Amendments (**E-FOIA**) of 1996, federal agencies are supposed to provide an index and description of their major information system webpages (Department of Justice, 1996).

The Act, signed into law by President Bill Clinton, directed federal agencies to create electronic reading rooms on their websites where the agencies could post documents and data requested repeatedly under FOIA. The electronic reading rooms are supposed to be the spot where we can find the lists of major information systems as well.

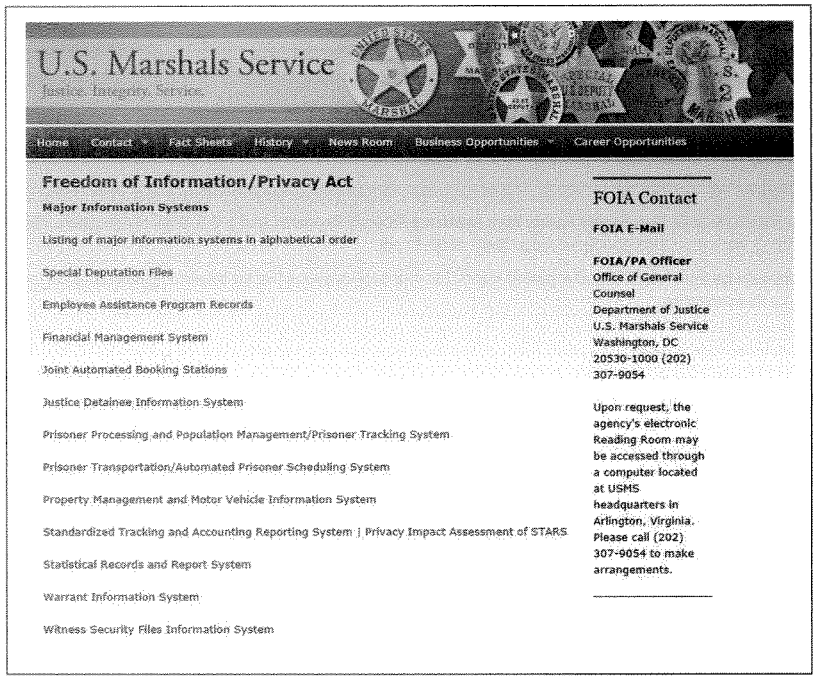
Some agencies have been better than others at disclosing details about their major information systems.

In fact, in 2007 most federal agencies were failing to comply with the law's requirement to post indexes and descriptions of their major information systems, according to a study by the National Security Archive at The George Washington University. As of that year, roughly one out of every three agencies had posted detailed information:

> Contrary to Congress's intent to make agency record-keeping more transparent, the manner in which agencies present record indexes and guides varies widely and is more confusing than helpful for requesters. Many agencies have not attempted to describe their record holdings in a systematic and comprehensive way. The indexes and major information system descriptions that are available vary widely in format and usability. . . . Unfortunately, this congressional mandate has failed, at least with respect to providing the public insight into agency record-keeping and publicly available information. (National Security Archive, 2007)

More recently, some federal agencies, such as the Drug Enforcement Administration, began removing the major information system listings on their own sites and referring site visitors to the Federal IT Dashboard at https://itdashboard.gov/ (Drug Enforcement Administration, n.d.a).

Some other agencies, such as the U.S. Postal Service and the U.S. Marshals Service, continue to have their own listings. For instance, the Marshals Service lists the Warrant Information System as one of its dozen databases that it uses (Marshals Service, n.d.a.).

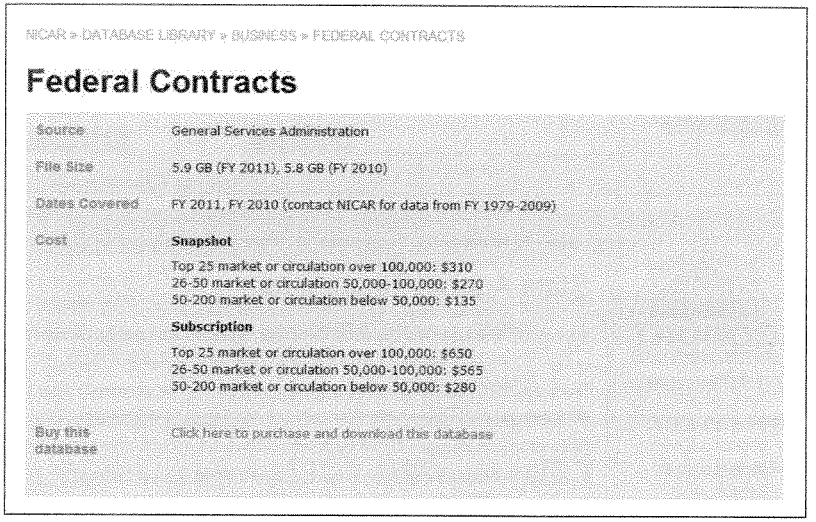


Source: Marshals Service. Retrieved from http://www.usmarshals.gov/readingroom/titles.html.

Note: Marshals Service major information systems.

A new initiative by the Obama administration could plug some of these holes. Under the White House's 2013 Open Data Policy, federal agencies are supposed to create data inventories and post public versions of them (Sinai and Van Dyck, 2013). If implemented properly, these inventories could help to identify the offline databases.

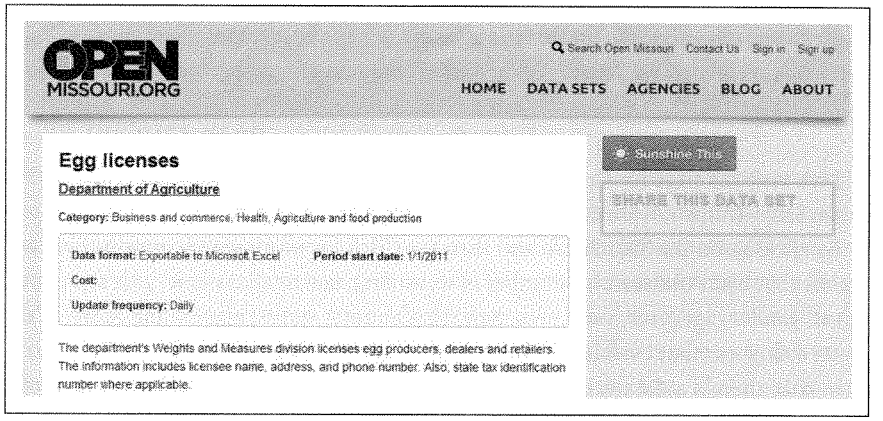
Nongovernmental websites can also tell you about what data these agencies have. The National Institute for Computer-Assisted Reporting Database Library, which is run by the Missouri School of Journalism and Investigative Reporters and Editors, lists more than 40 data sets from the federal government. Only journalists who are members of IRE may purchase the data, but anyone can scout for details about databases at the site. As an example, the database library (which the author has helped direct) provides data about federal contracts. Anyone can see detailed information about this database, noting that it is produced by the U.S. General Services Administration.



Source: National Institute for Computer-Assisted Reporting. Retrieved from http://www.ire.org/nicar/database-library/databases/federal-contracts/.

Note: NICAR Database Library entry for federal contracts data available to journalists.

Then there are sites aimed at informing the public about data held by state and local agencies. OpenMissouri (launched by the author in 2011) provides details about some 250 databases held by state agencies. This example from the state Department of Agriculture provides details about an egg license database that is exportable to an Excel spreadsheet (OpenMissouri.org, n.d.).

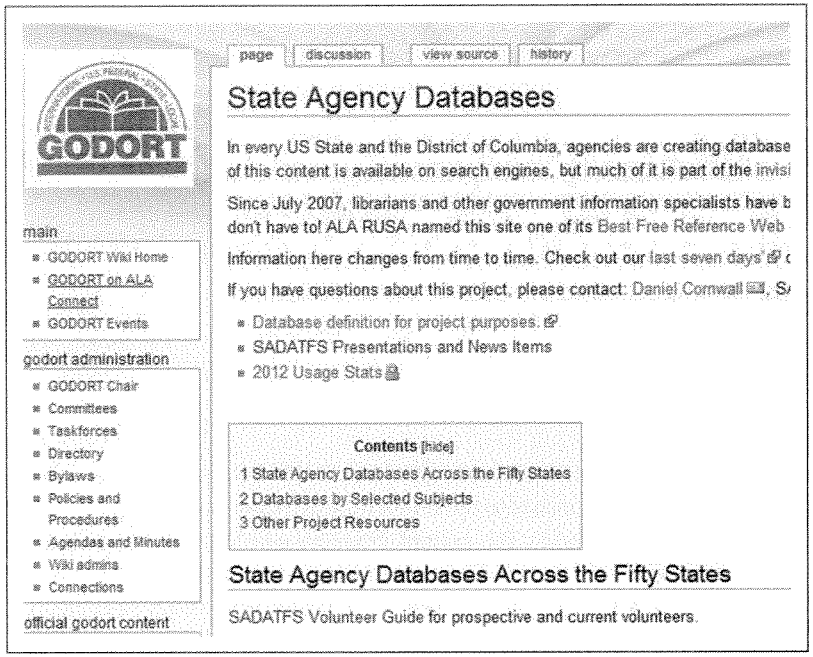


Source: Egg licenses. (n.d.). Open Missouri. Retrieved July 25, 2013, from http://openmissouri.org/data_sets/51-egg-licenses

Note: OpenMissouri entry for egg license data from the Missouri Department of Agriculture.

Another nongovernmental source that can be tapped to find information about offline data in all 50 states is the State Agency Databases wiki of the American Library Association's Government Documents Roundtable (http://wikis.ala.org/godort/index.php/State_Agency_Databases). Volunteer librarians who are specialists in government

documents and data have compiled the information since 2007 by scouring state agency websites for searchable online databases. Often, these sites lack data that we can download. But we do know that there are databases behind these search forms that we can request.

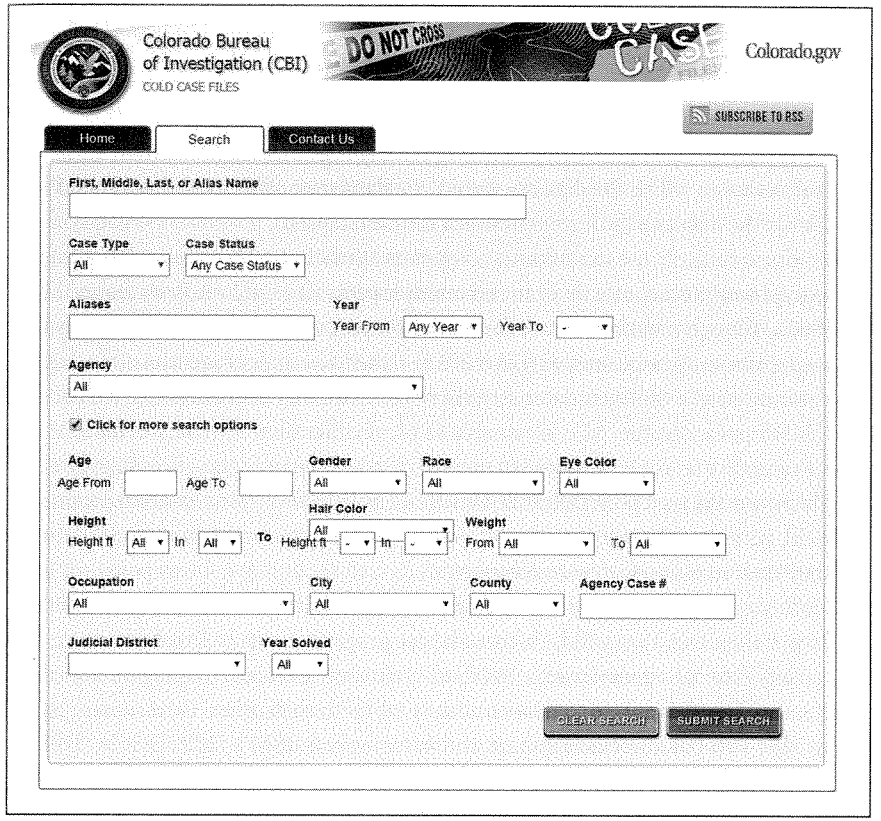


Source: American Library Association. Retrieved from http://wikis.ala.org/godort/index.php/State_Agency_Databases.

Note: State Agency Databases wiki. Volunteers from the American Library Association attempt to track state agency databases on this wiki.

Let's visit the page for Colorado and look for the Cold Case Database, listed under Public Safety (https://www.colorado.gov/apps/coldcase/index.html).

By clicking on the Search tab and then the Click for more search options box, we can see that users can search a number of ways: first, last or alias name, case type, case status, gender, age, year, race, eyes, hair, city, county, status, agency and more. It's a pretty safe guess that because the information is broken down that way on the search form that there's a data table (or tables) with those columns of data. So we could request that table from the Colorado Bureau of Investigation, which is the agency that keeps the data.



Source: Colorado Bureau of Investigation. Retrieved from https://www.colorado.gov/apps/coldcase/index.html.

Note: State of Colorado cold case database search.

FIND THE DATA NERD

After you've identified an offline database, you may need to do some more research before you can request and obtain it. The agency employees who create or maintain the data can be helpful because they are the ones who are most familiar with the system used for storing and retrieving the data. Sometimes you can find technical contact information on the government websites. Other times, you might need to call or email the agency to find the right contact person. In some government agencies, public affairs officers may try to prevent you from talking directly to the employees most familiar with the data. That's unfortunate, because public affairs officers are not as familiar with the data and can make the process more complex than it has to be. As a negotiating tactic, you can always ask to speak to the

data specialist directly, in the interests of making things easier for everyone. That way the burden will be lessened for the public affairs officer.

When you gather information about the offline database, you may need to address a range of technical issues about how the data are stored, processed and formatted. These discussions can get pretty involved, but they usually touch on three common areas: the physical device on which the data are stored, the database software used to process and manipulate the data and the format in which the file is stored.

Some governments use **mainframe** computers for their data processing. Mainframes are large, powerful machines that can run multiple processes and have been around since the 1960s. Though mainframes are costly and sometimes seen as outmoded, they still remain popular in some businesses (Lohr, 2012). Other government agencies often employ computer servers to run their database programs. Computer **servers** are less expensive than mainframes, but they also are unable to process the same amounts of data. Many times a computer server is dedicated to one task. Desktop computers—like Windows PCs and Macs—are the least powerful of the bunch and usually do not host database programs. That said, government agency employees sometimes do create spreadsheets and databases that they then access on their own computers.

Another consideration is the computer software that's used to store and manipulate the data. Sometimes that will be a spreadsheet program, such as Excel. However, it's more likely that an agency uses a database manager. Database managers are often relational—they allow users to relate multiple tables to each other. Government agencies usually run commercial database software, such as IBM's DB2, Oracle, Sybase, Microsoft SQL Server or Microsoft Access. All of those programs run on servers or mainframes, with the exception of Access, which is a Windows desktop program. Open-source database manager programs, such as MySQL or PostgreSQL, are less common in government agencies.

Unfortunately, all of these database programs store the data in their own formats, which are incompatible with each other. If agencies are unable to produce Excel files, they should be able to create ASCII text (delimited or fixed-width), because that is the format common to all computers.

REQUESTING THE DATA

After you know the name of the database and something about how the agency can provide it to you, you'll need to request it. You can make an informal request for the data just by asking for it by phone or in an email. Sometimes that works. Other times, you will need to make a formal request using the federal or a state open-records law.

If you want to get copies of offline data from a federal government agency, such as the DEA or EPA, you need to file a Freedom of Information Act request. If you want to get data from the state or any of the governmental entities within it, you would need to request

those data using that state's open-records law. So, if you wanted data from your local police department or county sheriff's office, you would need to exercise your rights under your state's open-records law.

Both FOIA, enacted in 1966, and state open-records laws start from the premise that all information collected by the government for public purposes is open. Even though FOIA and all of the state laws differ on many details, they all define some key points: response time, acceptable responses, exemptions, data formats and costs. We'll look at how FOIA and one state law—the Missouri Sunshine Law—compare.

FOIA gives the federal agencies 20 working days to respond to requests, compared to 3 working days under the Missouri Sunshine Law. Federal agencies are required only to respond with an acknowledgement that they received a request, so requesters can wait a long time to get data. Some journalists have waited years to get files from the Federal Bureau of Investigation, for example. In Missouri, agencies are supposed to provide the data or deny the request, by providing legal reasons for doing so within the three working days. They are also allowed to take more time to fill the request, as long as they provide a legitimate reason.

All the laws provide exemptions that allow government agencies to hold data that are considered nonpublic. FOIA has nine standard exemptions—including one for law enforcement materials whose disclosure could reasonably be expected to disclose the identity of a confidential source. Another exemption blocks the release of documents that are properly classified as secret in the interest of national defense or foreign policy. The Missouri law has a number of exemptions, including some arrest records and calls to 911 dispatch centers. Both laws have big loopholes in that they allow for records to be excluded by other laws.

Another key consideration is how the laws treat requests for data. Most of the laws—including the Missouri Sunshine Law and FOIA—specifically state that electronic data are public, just as paper documents are. In addition, some laws specify what rights you have in terms of the data format desired. Under FOIA, federal agencies are supposed to provide data in the format that's requested if the agency is able to do so. Under the Missouri Sunshine Law, the format in which the data are stored is considered the public record. Agencies are not required to produce data in other formats, though the law encourages them to do so.

Public-records laws outline allowable costs. FOIA allows federal agencies to pass along copying and reviewing costs. The Missouri Sunshine Law allows agencies to charge copying and staff salary costs. In addition, agencies can pass along programming costs when they choose to create a new record by putting existing records into a different format. FOIA and the Sunshine Law allow agencies to waive costs if the information contained in the request is in the public interest, something that's part of most state laws. Some states even allow requesters to limit their fees to a certain dollar amount.

WRITING THE DATA REQUEST

A good open-records request letter is clear, concise and includes enough detail about what you are seeking. The letter should include

- The name and contact information for the person filling the request;
- Citation of the law under which the request is being made;
- The name of the database, as it's known inside the agency;
- The names of the data columns, if known;
- The time frame of the data;
- The format in which you'd like to receive the data (along with the medium);
- Request for the documentation;
- Request for written explanation, in case of denial;
- Fee waiver request or limitation; and
- Your contact information.

The Reporters Committee for Freedom of the Press has an online service called iFOIA that allows users to build customized federal or state request letters; https://www.ifoia.org/#!/.

Source: Reporters Committee for Freedom of the Press. Retrieved from https://www.ifoia.org/#!/.

Note: iFOIA from the Reporters Committee for Freedom of the Press.

FOIA IN ACTION

As a Washington correspondent for *The New York Times*, Ron Nixon uses data frequently in his reporting on federal agencies. (The following discussion is based on Nixon 2013.) Nixon has used data for stories about how elderly people have been killed or injured by dangerous bed rails in assisted-living and nursing homes, and how companies with federal contracts also do business in Iran, despite U.S. government sanctions.

Sometimes he can get what he needs from U.S. government websites or by making an informal request with someone who works for an agency. If those two avenues fail, Nixon says, he files FOIA requests for the data.

Nixon estimated that he files two or three FOIA requests a month, trying to obtain data for stories that he has on his to-do list. Getting data pursuant to a FOIA request can take a while, so Nixon says he tries to give himself plenty of time.

When he requests data, Nixon says he tries to be as specific as possible; he avoids making requests for everything in the database because FOIA officers in agencies usually kick back those requests. "They always want you to narrow it down," Nixon says. "If you know what you're looking for, it helps."

Nixon recommends that requesters "do a lot of leg work" before filing requests. For instance, Nixon sometimes bases his requests on the information that's collected on federal forms. Sometimes he looks at online federal agency FOIA logs to see what data others have already requested. If someone has already requested and received data, the agency should be able to easily provide a copy of those data.

When making FOIA requests, Nixon says to remember that fulfilling them is not a priority for federal agencies. Agencies do not have enough FOIA staff to do the job well. Remember that the FOIA staff members are overwhelmed, he says, and try to be as nice as possible to them when making or following up on requests.

NEGOTIATING THROUGH OBSTACLES

In real life, getting the data you need can be tough: government agencies can put up many obstacles, some of them legitimate and others not. Some of the common obstacles are privacy issues, cost and ability to produce the data.

Often, government agencies will tell you that they're unable to produce the data because the data are private, which may very well be true. However, under FOIA and the state open-records laws, the agencies are supposed to specify the section of the law that makes these data private. Just because someone thinks data are (or should be) private, doesn't mean that they truly are under the law. Ask agencies to cite the law in writing so you have a copy of the agency's legal reasoning on the record.

If some of the data truly are private, you can probably get the public portion of the data. That's because under FOIA and the state open-records laws, agencies are supposed to disclose what's public. For example, federal law prohibits the unauthorized disclosure of Social Security numbers. If you want to get a database of city employees that includes Social Security numbers, city hall is supposed to redact those private data and give you the rest.

An alphabet soup's list of federal privacy laws—FERPA, HIPAA and DPPA—could bar disclosure of some data stored in government databases.

FERPA is the **Family Educational Rights and Privacy Act**, and ensures the privacy of student educational records. Students have control of these records when they turn 18 or attend a school after high school. Before that, parents have control of the records.

Regulations drafted under FERPA allow school officials with a legitimate educational interest to have access to these records (Department of Education, n.d.). FERPA can become a stumbling block when people seek data from public schools or universities. Public colleges have denied requests for campus arrest records based on FERPA. The Student Press Law Center, which provides assistance to scholastic and university journalists, says FERPA ranks among the most common complaints that it receives and provides a PDF guide to help requesters work through denials (Gregory, 2013).

HIPAA is the **Health Insurance Portability and Accountability Act** of 1996, which protects individually identifiable health information (Department of Health and Human Services, n.d.). No one would argue that our personal medical data should be released without our consent.

DPPA, or the **Drivers Privacy Protection Act**, passed in 1994 after a number of high-profile incidents in which stalkers were able to get home address information from state motor vehicle departments. DPPA restricts states from releasing the personal data contained in a person's driving record (epic.org, n.d.). The law, however, does allow the data to be provided for a number of purposes, including insurance, motor vehicle recalls and court proceedings.

Costs can also become an obstacle. Agencies will sometimes say that providing the data will cost a lot of money. Ask for a price quote in writing, with a breakdown of costs that includes whatever is allowed under the law: copying, research, programming and so on. Even if the cost of the data is reasonable under the law, it still may be more than you can pay. Remember, you can always ask for a fee waiver in your request. Make sure that you research the conditions for the waiver and state how your request meets these conditions.

Another real world obstacle is difficulty in getting the data in a format that you can use. In Missouri, state agencies only must produce data in their original format. So, if the state department of higher education has a particular database that's stored as an Oracle file, it doesn't have to produce an Excel or ASCII file. In other cases, an agency may claim that it lacks the expertise to produce the data in a format that you can use. Or the agency might have outsourced its information technology operations to a third-party vendor, making it difficult to figure out exactly how the vendor could output the data for you. Fortunately, most state laws say that government agencies cannot avoid their obligations under open-records laws by outsourcing their database operations to private vendors.

GETTING HELP

Negotiating for information (whether data or documents) from public agencies truly is an art, one that requires more knowledge and nuance than this chapter can provide. If you want more in-depth guidance, check out *The Art of Access* by David Cuillier and Charles N. Davis (2011). It's a great how-to guide that's written for journalism students and working journalists, but the lessons can be applied by people in any field.

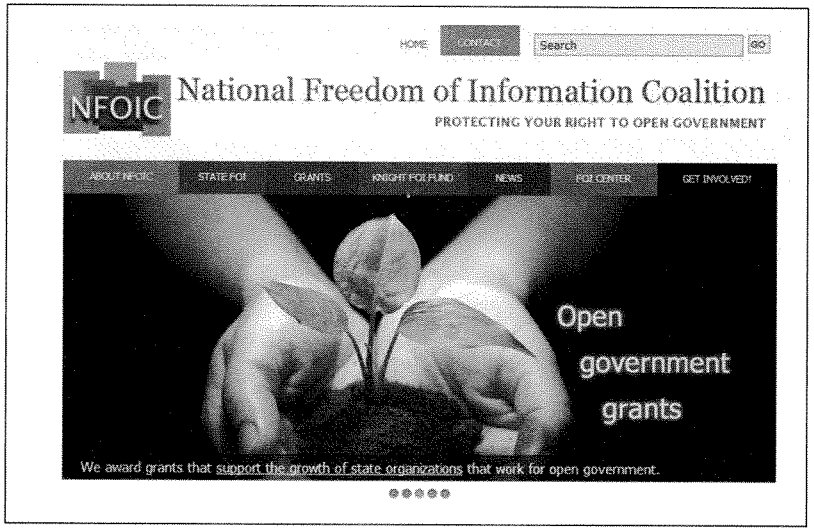
Another great resource is the Open Government Guide from the Reporters Committee for Freedom of the Press, which is a nonprofit organization that assists journalists.

The OGG provides detailed information about FOIA, federal privacy and state open-records laws. The guide includes specific information about how the laws treat data.

In some states, dedicated agencies can sometimes assist in data requests. In Connecticut anyone who's denied a request for data or other records under the state Freedom of Information Act can appeal that decision to the Freedom of Information Commission. The FOI Commission holds hearings about the complaints and decides whether the agencies have violated the law. If the agency has violated the law, the commission may order the agency to produce the data (Connecticut Office of Governmental Accountability, n.d.).

In Texas, the attorney general's office gets involved in some data requests. When an agency denies information, the agency is supposed to ask for a decision from the AG's office. Both the requester and the agency denying the information are then allowed to provide input to the AG's office (Attorney General of Texas: Greg Abbott, n.d.).

In addition to these official government agencies, you can get help from nongovernmental organizations. A good starting point for identifying such groups in your state is the National Freedom of Information Coalition, which is based at the University of Missouri. The NFOIC (www.nfoic.org) represents the interests of state-level open government groups and, on its website, provides links to them. So, if someone in Florida is looking for help with challenges getting data from a state or local agency, he or she can go to the NFOIC site, follow a link to the Florida First Amendment Foundation and seek help from that organization.



Source: National Freedom of Information Coalition. Retrieved from www.nfoic.org.

Note: National Freedom of Information Coalition website. The NFOIC represents the interests of state-level open government groups.

As we've seen, it can be tough to identify and obtain data that are held offline, as government agencies do not readily inform us about the data they keep. Fortunately, we can use the clues—such as records retention schedules and search forms—to uncover databases. Then, we can ask for the data informally, or make a formal open-records request.

Our next step in working with the data is testing them, so we know their flaws and limits. Only after that can we analyze and visualize them.

ON YOUR OWN

Find your state's open-records law. Cite it and write a few paragraphs describing the process for requesting data. Include details about any appeals or possible intervention by government officials in other agencies.

Identify three possible offline databases in your state by using the American Library Association's State Agency Databases wiki. Cite the URLs and write a summary of what columns of data the databases might contain.

Identify your state or local open-government group using the NFOIC membership list. Record the URL for the group's website and contact information. Note whether the group provides any assistance to the public.

Find a federal GAO or state audit report that mentions a database. Cite the report and write a brief summary of the database and how the auditors used it.

Find an instance in which FERPA privacy concerns by a public agency have resulted in student journalists being denied data or other information. Summarize the details. Do you believe this denial was legitimate? Why or why not?

Write a mock open-records request letter asking for public employee data from your city for the past three years. Request the data as an Excel file, to be provided via email.

Find your state in the Reporters Committee for Freedom of the Press Open Government Guide. How does your state's law treat data? Does the law say computerized data are public records? What does the law say in terms of your rights to request data in a particular format? Cite the law and write a summary.

SECTION III

EVALUATING AND CLEANING DATA