2010

Finding Data Sets Online

Steven Ovadia
CUNY La Guardia Community College

Behavioral & Social Sciences Librarian, 29:81–85, 2010

ISSN: 0163-9269 print / 1544-4546 online
DOI: 10.1080/01639260903572342

Routledge
Taylor & Francis Group

Internet Connection

Finding Data Sets Online

STEVEN OVADIA
LaGuardia Community College, Long Island City, New York

Data seems to be an increasingly important part of everyone's life, both within librarianship and the social sciences. Most librarians are able to easily find charts and tables, but sometimes more detailed information is needed. Sometimes custom charts and tables need to be built. This is where data sets come in. These collections of raw data allow users to manipulate and massage variables as they see fit. Where a chart or a table is limited by whichever variables the author chose to include and make available, raw data represents all of the data collected for a given project, allowing end-users to select and manipulate whichever components strike their fancy in whatever manner makes sense for their own research. While one could devote a book to all of the interesting data sets freely available online, here is a sample of some of the more interesting ones.

INTEGRATED PUBLIC USE MICRODATA SERIES (IPUMS-USA)

IPUMS (http://usa.ipums.org/usa) is a product of the Minnesota Population Center (MPC), part of the University of Minnesota. The MPC has quite a few data sets, from health to time use and covering populations both international and national.

One of the more interesting data sets is IPUMS-USA, which contains raw Census data from 1850 to the present as well as American Community Surveys from 2000–2008. The data also includes Puerto Rico Samples and Puerto Rican Community Surveys. What is meant by raw data? Each record within IPUMS represents a person. Because the data is so complex, it needs to be analyzed with statistical software like SAS or SPSS (although IPUMS does offer limited online analysis via the Survey Documentation and Analysis program developed by University of California–Berkeley's Computer-Assisted Survey

Address correspondence to Steven Ovadia, LaGuardia Community College, Library Media Resources Center, 31-10 Thomson Avenue, Long Island City, NY 11101. E-mail: sovadia@lagcc.cuny.edu

Methods Program), so it's not the kind of information you would use to figure out something basic, like the 2000 population of a state. However, it is the type of data you would use to create a longitudinal chart of a state's population from 1850 to the present. And, as one gets more comfortable with the variables, it would also be the data used to track correlations between educational attainment and socioeconomic status over time in a given area.

The IPUMS documentation is extensive, which is good because the process of using and understanding the data can be very complex. But one cannot help but consider the research possibilities that exist within the Census variables captured in IPUMS. The various Census variables are the heart of the IPUMS experience. As you become more familiar with what's available and for which time periods, all of which IPUMS makes quite transparent within its documentation, users begin to understand how to construct data sets that can reveal meaningful results. For instance, if someone wanted to track literacy in the United States, they could use the LIT variable, but only from 1850 to 1930.

IPUMS does more than just provide access to data. The IPUMS team also helps end users make sense of the Census variables, which are susceptible to change over time. For instance, the 2000 Census was the first to allow people to select more than one race variable. This relatively simple-sounding change proved to be a challenge for IPUMS: "The shift from a single-race to a multiple-race census inquiry is a fundamental conceptual change, and we cannot construct a perfectly backward-compatible variable. We can, however, provide researchers with a range of constructed race variables that maximize historical compatibility for specific research applications" (Ruggle et al. 2003).

For users unsure of the possibilities of IPUMS, the site contains an extensive bibliography of academic work built on IPUMS data. The bibliography can be found at http://bibliography.ipums.org and is searchable by keyword and topic.

If you're not a huge fan of Census data, you still might want to explore the Enumeration Forms part of the site, which lists the actual Census questions asked over the past 150 years as well as all of the instructions associated with that particular Census. While many of the questions seem quite upsetting in terms of today's norms, they're a fascinating examination of how the United States government has framed social issues like race and gender over time.

INTEGRATED POSTSECONDARY EDUCATION DATA SYSTEM (IPEDS) DATA CENTER

IPEDS (http://nces.ed.gov/ipeds/datacenter) aggregates survey data collected by the National Center for Education Statistics (NCES), part of the

U.S. Department of Education's Institute of Education Sciences. Like the U.S. Census site (www.census.gov), IPEDS offers a fair amount of pretabulated data. Unlike the Census site, IPEDS also provides access to data sets, allowing researchers to build custom reports based on the IPEDS variables available.

IPEDS data deals purely with higher education within the United States and allows users to focus either on specific institutions or specific institutional characteristics. These variables include items like enrollment, retention, graduation rates, student ages, and faculty and staff information. In terms of longitudinal explorations, the years of data available vary depending on the variable.

The data can be very helpful for institutional queries, like which public two-year schools have the highest graduation rates. But the data can also be helpful for social science researchers trying to pull gender and race factors out of education data. The IPEDS data, and the NCES Web site in general, can be challenging to navigate. There is a lot of information, but documentation often does not seem to be readily available. However, it is important to note there is an IPEDS data help telephone line. The author has used the help line in the past and found the people answering questions to be both knowledgeable and helpful.

It is also worth noting the NCES is also the home of Library Statistics Program (http://nces.ed.gov/surveys/libraries/aca_data.asp). This includes the results of the Academic Libraries Survey (ALS) as public-use data sets. This data is similar to the IPEDS data but deals only with academic libraries. It allows users to extract budget data, collection information, and library-specific faculty and staffing numbers.

This data can seem a bit limited at first glance, but there are some very interesting results that can come out of it. A good example of this is Weiner's (2009) study of how library's contribute to the reputation of a university. The study took advantage of several data sets, including ALS and IPEDS.

However, IPEDS is not without some controversy, as some community college leaders have said IPEDS data does not accurately measure the work of community colleges (Bradley 2009). As with any data not collected by the researcher using it, the questions asked, as well as *how* they are asked, need to be evaluated to ensure the data set accurately captures what the researcher is trying to quantify.

DATA.GOV

IPEDS and IPUMS are two very specific data collections that will probably be used for very targeted research. Data.gov (http://www.data.gov) takes a different approach to data sets, serving as a clearinghouse/catalog for United States government generated data collections. It's a relatively new project, having launched in May 2008.

Conceptually, the idea behind Data.gov is sometimes referred to as Government 2.0, "the notion that technology cannot only make it easier to get the government's information but can be turned back onto the government itself to make the bureaucracy more responsive to its constituents, a faster and smarter beast" (Homans 2009). Researchers will not necessarily not have an advocacy role in mind when they search Data.gov, but they may be able to find some useful information, although at this point in Data.gov's development, this is not a given.

Data.gov divides data into three main categories: raw data, tools, and geodata. Raw data indicates data in a format that can be manipulated by the end user. Specifically, the formats include XML, Text/CSV, KML/KMZ, Feeds, XLS, and ESRI Shapefile. The tools area indicates links to sites that have some kind of data mining/extraction functionality. Geodata indicates federal geospatial data sets. Data.gov also links to state and local data collections, although as of this writing, only four states and the District of Columbia were represented.

Searching Data.gov can be a bit of a mixed bag, since the scope of content is so wide. One can find active mines and mineral plants in the United States just as easily as one can find Department of Labor statistics on injuries, illnesses, and fatalities. The incredibly wide range of material available lends itself to browsing via one of the site's 26 categories. The data can also be searched by the agency responsible for creating it.

As it stands now, Data.gov is more useful in theory than in practice. While there is some interesting, important data to be found, the data does not seem to as well maintained or curated as what you'll find via a tool like IPUMS, which focuses considerable energy on just one data collection. Still, one cannot help but be excited by the potential of something like Data.gov as it becomes more of a fully developed product. Imagine having easy access to raw government data and imagine the convenience of having it all in one place across agencies. Also, with all of this data housed in one place, it seems much easier to serendipitously discover other helpful data sets.

CONCLUSION

These online data sets can provide researchers with possibilities and opportunities. Reviewing the data and variables available, one cannot help but think of all of the research questions that might be answered by using these collections. Ready-made raw data collections like these also make it much easier for time-strapped researchers. Instead of going out to collect data on a topic, a lot of times an existing data set might already have completed the work.

These collections are not great ready-reference tools, but for anyone interested in these topics and looking to repurpose data for their own research

needs, the convenience and flexibility of these data sets can be a tremendous time saver.

REFERENCES

Bradley, Paul. 2009. Updating rules and measures: Leaders push for new approach to measuring performance. *Community College Week 21*(15): 6–7.

Homans, Charles. 2009. The geekdom of crowds: The Obama administration experiments with data-driven democracy. *Washington Monthly 41*(7/8): 13–17.

Ruggle, Steven, Matthew Sobek, Miriam L. King, Carolyn Liebler, and Catherine A. Fitch. 2003. IPUMS redesign. *Historical Methods 36*(1): 9–19.

Weiner, Sharon. 2009. The contribution of the library to the reputation of a university. *The Journal of Academic Librarianship 35*(1): 3–13.