# Identifying, Accessing and Evaluating Data

FINDING AND ACCESSING DATA CAN BE PROBLEMATIC, BUT MANY OF THE SKILLS USED IN TRADITIONAL REFERENCE CAN BE APPLIED TO DATA DISCOVERY.

**BY JENNIFER HUCK, MS, MLS**

Helping patrons find data is notoriously difficult and frustrating. Common difficulties include the lack of a single place to look for data, difficulty finding data specific to a time period or place, and having to manage patron expectations (Kubas and McBurney 2019). Fortunately, many of the skills used in traditional reference can be applied to data discovery.

I am the data librarian at the University of Virginia Library, an academic library at an R1 research institution. I most frequently help students, faculty and staff in the social sciences, commerce, and education, and less frequently in law, health sciences, and hard sciences. I am also fortunate to have the support of liaison librarians dedicated to these schools and departments.

In this article, I will review the techniques I use when responding to data requests. I will address the questions you should answer before you begin your search, where to look for data, potential access issues, and how to evaluate data.

## Get Answers to These Questions First

In my experience, patrons will most often reach out about a topic, but with widely varying levels of specificity. You will save yourself a lot of time and effort if you learn the details upfront, so be sure to ask for clarification on the following topics as soon as you receive a request.

**Unit of analysis.** Let's say a researcher tells you she needs data to study elections. The unit of analysis could be voters, precincts, districts, countries, and so on. A researcher studying a broader unit (e.g., a country) might be satisfied with data using a smaller unit of observation (e.g., districts), but the reverse is unlikely to be true.

**Geographic coverage.** What kind of geographic coverage does the researcher need? For example, does she need data only for a local state or province, or for several specific countries?

**Time period.** How many years' worth of data is she seeking? Does she want only the most recent year available, or as many years' back as possible?

**Data or statistics?** Is the researcher trying to find data to use for her own secondary analysis? For example, a researcher studying retail purchasing behavior may need scanner data from individual retailers. This researcher might then run a regression or some other statistical analysis. Alternatively, a student studying market trends may only need general statistics about best-selling grocery items to create a simple bar chart in a report. Since no secondary analysis will be conducted, only general statistics are needed.

It is helpful to become acquainted with data-related terms. For a further explanation of "data jargon," see

**JENNIFER HUCK** is the data librarian at the University of Virginia Library. Her main data-related responsibilities encompass data discovery and access and maintaining data collections. She is the liaison to the university's public policy school and is embedded into the Research Data Services + Social, Natural, and Engineering Sciences (RDS+SNE) liaison team. She can be reached at jhuck@virginia.edu.

"Chapter 1: Data Reference Basics" in Bauder (2014) and also check out the Inter-university Consortium for Political and Social Research's Glossary of Social Science Terms (ICPSR n.d.).

## Find the Data

Just as you rely on subject expertise in a regular reference interview, you should also build up your data expertise to support data in the library. Start by familiarizing yourself with the terminology of the field. You can do this by reviewing methods textbooks and journals in the discipline. It's also useful to review the kinds of data that researchers at your institution are using. For a useful listing of specific discovery techniques, see Bordelon (2016).

It is critical to familiarize yourself with the key sources of data in your field. The key question to ask yourself is: Who would care enough to collect and disseminate data on this topic?

**Government agencies.** Government agencies are an essential source of data for a variety of disciplines. In the United States, the Bureau of the Census, the Bureau of Labor Statistics, and the National Center for Education Statistics are examples of federal agencies that disseminate a wide variety of data (Office of Management and Budget 2018). At the international level, there are government agencies in other countries as well as intergovernmental organizations such as the United Nations (and its many programs, funds, and specialized agencies), the Organisation for Economic Co-operation and Development (OECD), and the World Bank. At the local level, it is becoming more common for states and localities to share data through open data portals.

**Data repositories.** Another useful tool in your arsenal is knowing which data repositories might host data of interest to your stakeholders. This is especially useful knowledge when you are looking for data created and shared by other researchers. The Inter-university Consortium for Political and Social Research (ICPSR) is a classic example of a data repository, as it is one of the oldest data repositories anywhere. It hosts a very well-curated collection of social science data, and the quality of the metadata is high. Dataverse is another example of a data repository—it hosts academic research data submitted directly by researchers.

Note that datasets found within Dataverse and other research data repositories will vary widely in the quality of their metadata, data dictionaries, and codebooks. A great way to find data repositories is to explore re3data.org, which has detailed information about more than 2,000 repositories. You can browse repositories by subject, content type, or country.

**Non-repository data hubs.** The "Wild West" of data discovery encompasses author websites and sharing platforms such as GitHub and Kaggle. It is considerably harder to find useful data through these websites, because your best search tool is a regular web search. Unlike a data repository, which is designed to preserve data at the end of the research cycle, data on a website or platform could be in active development and could easily disappear in the future.

You can also try Google Dataset Search, which recently came out of beta. The beta version lacked facets or advanced search features, which I found frustrating. Now that it is out of beta, there are minimal filters available by date and format, which is an improvement. It remains difficult to tell what is and, just as importantly, what is not captured in a Google Dataset search.

## Use Your Usual Research Tools and Techniques

Fortunately, you can rely on some traditional library reference tools and techniques. One of the best things you can do when searching for data is to review the research literature with a relevant indexing platform, such as Google Scholar or Web of Science. What do scholars who study a similar topic use for data? Use their articles to track down the data source. Similarly, you can use Worldcat or library catalogs to discover books and articles related to your topic, then search for data from there.

You can also use statistical abstracts or databases, such as ProQuest Statistical Abstracts or Statista, to track down data. These platforms will often point you to the original data source via citations or source notes, and from there you can keep tracking down your dataset.

Print indexes and reference books are extraordinarily useful if you ever need to search for historical data and statistics. (In my experience, anything before the mid-2000s is a good candidate for this treatment.) Become familiar with federal and state government documents and the print indexes and documents that governments produced, as these will be very helpful when looking for historical statistics. Two titles to know are the American Statistics Index (ASI) and Statistical Reference Index (SRI), both published by the Congressional Information Service and generally covering the late 20th century.

You should also search your library catalog or Worldcat for statistics indexes. Conduct a subject search for "statistics" in combination with "serials" or "periodicals." Note: you are more likely to find historical statistics rather than data.

Remember that you can rely on the library community's resources as well. I particularly like to search LibGuides; you can simultaneously search all LibGuides at https://community.libguides.com/. I also like to search the data pages at university libraries, such as Princeton Data and Statistical Services (https://dss.princeton.edu/) and Duke Data Sources (https://library.duke.edu/data/sources). These are especially helpful for discovering licensed data that my library does not already receive through subscription.

## When in Doubt, Reach Out to Real People

Sometimes the above techniques still will not deliver the desired results. If you think certain data exist and you are simply having trouble finding or accessing them, it is best to reach out directly to

government agencies, archives, special libraries, and/or researchers.

I especially value my communities where we help each other find data. I particularly recommend the International Association for Social Science Information Services and Technology (IASSIST), which essentially is an association of social science data librarians. Their listserv is especially helpful when you have reached the limits of your data discovery skills but are not sure you have exhausted all possibilities. Finding a community of librarians who provide similar discovery services is especially important if you have few or no colleagues at your library.

## Manage Access Issues

One little-discussed but very significant concern about supporting data in the library has nothing to do with discovery at all. Finding the right data can be challenging, to be sure, but real problems can arise regarding access. I frequently work with patrons who know exactly what data they need, but they are having trouble using it for a variety of reasons. Access problems include the following:

**Data are restricted.** Some datasets are restricted, often because they include sensitive or personally identifiable information. If researchers are lucky, they may simply have to apply for access, which will take time. If they are unlucky, the data will not be available for use.

**Data must be licensed and are expensive.** Unlike journals or booksellers, many data vendors are unaccustomed to working with libraries. You may sometimes need to explain how and why your library needs to license materials, and that process can be time-intensive. Licensed data products also tend to be very expensive, often pricing them out of reach. If your library successfully licenses the data, you will need to find a place to securely store and distribute the datasets. For more information on libraries licensing data, see Hogenboom and Hayslett (2017) and Hogenboom et al. (2011).

**A researcher has an idea about what she wants to use, but is not clear how to actually work with the data.** For example, I sometimes work with researchers who realize they need to use the census for their research, but are not at all certain about the limits of the census and how it should be treated.

**Data come in a format with which the researcher has no experience.** Sometimes data are available but only in unusual, deprecated, or proprietary formats (truer of historical data) or emerging formats such as APIs (truer of contemporary data). The researcher may not be comfortable with such formats, thus adding a layer of inaccessibility. An even simpler example is the researcher who has very little statistical training and needs help with basic file formats and statistical analysis.

In all these cases, it is apparent that supporting data in the library is not only about discovery—researchers may come to the library knowing exactly what data they need. Libraries can be useful to these patrons by helping them understand and solve their access issues.

## Evaluate the Data

Once you find and access your data, you will need to evaluate it. Fortunately, many of the typical ways librarians evaluate information apply to data and statistics as well. Tests like CRAAP still hold. Questions unique to evaluating data and statistics relate to collection methods and documentation such as codebooks (Carleton College Library n.d.; North Dakota State University Libraries n.d.; University of Washington University Libraries n.d.). Being able to evaluate collection methods and documentation requires a deeper understanding of the discipline you serve and will be different for every discipline.

## Conclusion

Even seasoned data librarians encounter difficulties and roadblocks when conducting a data reference search (Kubas and McBurney 2019). It is essential to have some subject expertise and a baseline understanding about how data is used in the discipline you serve. Finding a community of librarians who can help you build your skills will be beneficial. Although it can be challenging to provide data discovery services, your expertise will grow with practice. **SLA**

### REFERENCES


Bauder, J. 2014. *The Reference Guide to Data Sources*. ALA Editions. Chicago, Ill.: American Library Association.

Bordelon, B. 2016. Data Reference: Strategies for Subject Librarians. In L. Kellam and K. Thompson (Eds.), *Databrarianship: The Academic Data Librarian in Theory and Practice*. Chicago, Ill.: Association of College and Research Libraries.

Carleton College Library. n.d. Data, Datasets, and Statistical Resources: Factors to Consider when Evaluating Statistics. Webpage. Northfield, Minn.: Carleton College.

Hogenboom, K., and M. Hayslett. 2017. Pioneers in the Wild West: Managing Data Collections. *Portal: Libraries and the Academy*, 17(2): 295-319.

Hogenboom, K., T. Teper, and L. Wiley. 2011. Collecting Small Data. *Research Library Issues: A Bimonthly Report from ARL, CNI, and SPARC*, 276, 12-19.

Inter-university Consortium for Political and Social Research. n.d. Find & Analyze Data: Glossary of Social Science Terms. Webpage. Ann Arbor, Mich.: The Regents of the University of Michigan.

Kubas, A., and J. McBurney. 2019. Frustrations and roadblocks in data reference librarianship. *IASSIST Quarterly*, 43(1): 1-18.

North Dakota State University Libraries. n.d. Finding Data and Statistics: Evaluating and Finding Data and Statistics. Webpage. Fargo, N.D.: North Dakota State University.

Office of Management and Budget. 2018. *Statistical Programs of the United States Government*. Washington, D.C.: Executive Office of the President.

University of Washington University Libraries. n.d. Savvy Info Consumers: Data & Statistics. Webpage. Seattle, Wash.: The University of Washington.