The Reference Guide to DATA SOURCES

JULIA BAUDER

The Reference Guide to DATA SOURCES

JULIA BAUDER

ala
editions

An imprint of the American Library Association

CHICAGO 2014

Julia Bauder is the social studies and data services librarian at the Grinnell College Libraries in Grinnell, Iowa. Bauder holds a master's degree from the School of Library and Information Science at Wayne State University in Detroit, Michigan. Before becoming a librarian, she spent several years as a freelance writer and editor of reference books.

Printed in the United States of America

18 17 16 15 14   5 4 3 2 1

Extensive effort has gone into ensuring the reliability of the information in this book; however, the publisher makes no warranty, express or implied, with respect to the material contained herein.

ISBNs: 978-0-8389-1227-0 (paper); 978-0-8389-1962-0 (PDF); 978-0-8389-1963-7 (ePub); 978-0-8389-1964-4 (Kindle). For more information on digital formats, visit the ALA Store at alastore.ala.org and select eEditions.

Library of Congress Cataloging-in-Publication Data

Bauder, Julia.
The reference guide to data sources / Julia Bauder.
p. cm.
Includes bibliographical references and index.
ISBN 978-0-8389-1227-0 (alk. paper)
1. Statistics—Computer network resources—Handbooks, manuals, etc. 2. Social sciences—Statistics—Handbooks, manuals, etc. 3. Internet research—Handbooks, manuals, etc. 4. Electronic reference sources—Handbooks, manuals, etc. 5. Internet searching—Handbooks, manuals, etc. 6. World Wide Web—Subject access—Handbooks, manuals, etc. I. Title. II. Title: Guide to data sources.
HA33.5.B38 2014
025.06—dc23   2014006175

Cover design by Alejandra Diaz. Image ©bloomva/Shutterstock, Inc.

Text design in the Chaparral, Gotham, and Bell Gothic typefaces.

∞ This paper meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper).

The Reference Guide to

DATA SOURCES

ALA Editions purchases fund advocacy,
awareness, and accreditation programs
for library professionals worldwide.

Acknowledgments

THANK YOU TO ANNE O'DWYER FOR INTRODUCING ME TO QUANtitative research in the social sciences, and to all of the faculty at Simon's Rock College who worked so hard to help me become a better writer. I would not have been able to write this book without their patient teaching.

Thank you to Alana Joli Abbott for setting me up with my first freelance writing jobs, which started me on the career trajectory that led to this book.

Thank you to all of the data librarians who graciously welcomed me into the profession. This list includes Chuck Humphrey and Jim Jacobs, whose ICPSR class, Providing Social Science Data Services: Strategies for Design and Operation, was an excellent introduction to the field, and my colleagues in the International Association for Social Science Information Services and Technology and the ACRL Numeric and Geospatial Data Services in Academic Libraries Interest Group.

Thank you to the countless people over the years who have said to me, "Hey, have you seen this great data set?" I first became aware of many of the sources listed in this book because an enthusiastic colleague shared them with me, and I appreciate both the tips and the enthusiasm.

Thank you to Grinnell College for generously funding research leaves for junior faculty, including me, and to all of my colleagues in the Grinnell College Libraries who picked up the slack for me while I was off writing. I would never have had the time to write this book without that support.

Thank you to the wonderful editorial team at ALA Editions. They provided excellent suggestions that improved this book immensely.

And, finally, thank you to all of my friends and family who, for almost a year, put up with me saying, "I can't; I have to write!"

Contents

Appendixes

1 Data Reference Basics

QUESTIONS ABOUT STATISTICS—"WHAT IS THE POPULATION OF London?" "How many people are diagnosed with cancer annually?"—have long been a staple at library reference desks. Most librarians are familiar with the print ready-reference sources that were traditionally kept within reach to answer them—the *Statistical Abstract of the United States, The World Almanac and Book of Facts*, and the *CIA World Factbook*, to name just a few—and, now, with the online equivalents of these publications. Not too many years ago, however, "data reference" was a specialized library service available primarily at research libraries. Few librarians outside of those institutions would ever encounter a question like "I need thirty years of time-series data on the production of beef in Texas at the county level," and even fewer librarians would have been comfortable with the intricacies of working with data on reel-to-reel tapes and punchcards.

Two changes—the rise of the Internet, which has made disseminating data much less complicated, and the spread of statistical software packages into the undergraduate curriculum—have blurred the line between statistics reference and data reference. As more data and more user-friendly tools to work with data have become available, interest in finding and using quantitative data has grown. Yet data-related reference questions are often some of the most daunting questions for general reference librarians. My hope is that this guide makes those questions a little less daunting.

FOR WHOM IS THIS GUIDE INTENDED?

Although I hope that all reference librarians find this guide helpful, its primary audience is librarians at public libraries, high school libraries, and the majority of academic libraries that do not employ staff dedicated to data reference. Librarians working with undergraduates at large research institutions may also find that the guide allows them to answer some basic data reference questions without having to refer students to a dedicated data services librarian.

Because the intended audience is primarily librarians at institutions that have not made a major financial commitment to data services, this guide focuses on freely available, online sources for data. In many subject areas, much of the most frequently used data can be found online for free if one knows where to look. On the occasions when a commonly requested type of data is not freely available from any online source, subscription databases or print series containing the data may be mentioned. Librarians whose institutions subscribe to many statistical databases may find Lynda M. Kellam's guide *Numeric Data Services and Sources for the General Reference Librarian* (Chandos, 2011), which focuses more on subscription databases for data reference, to be helpful for their situation.

This guide focuses on data available through English-language interfaces, although it occasionally makes reference to data available in foreign-language interfaces. It does not include certain specialized scientific data sources, such as gene sequence databases or databases of chemical structures; nor does it include qualitative data sources. Instead, the focus is on quantitative social science data, broadly defined—that is, data that is primarily useful in the context of social science disciplines such as economics and sociology, as well as scientific data that frequently is or can be deployed in the context of government policy making, such as data on public health, climate change, or natural disasters.

In the following thematic chapters 2–26, the discussion is divided into three categories: major sources for U.S. data, major sources for international data, and minor sources. The sites listed as major sources are large databases, typically from the major international or federal government agencies with responsibility for a given area; they are the most likely resources for answering most common reference questions in their areas.

Minor sources were selectively chosen from the dozens of smaller data sources in each area because they fill a gap in the data that is available from the major sources or because they present a selection of data from the major sources in a more user-friendly interface.

DATA JARGON

To communicate with patrons who need data, and to get the most out of the rest of this book, it helps to be aware of certain types of data jargon. As with any kind of reference question, before undertaking a search for data it is important to be sure you understand exactly what your patrons want. In the case of data that means understanding not only exactly the topic on which they need data but also the characteristics of the data they need: whether they need *statistics* or *data*, what their desired *universe* and *unit of analysis* are, whether they want *time series* or *cross-sectional* data, and more. The italicized terms, and other related concepts, are defined below.

Data versus Statistics

The terms *data* and *statistics* are often used as if they mean the same thing (sometimes even in this book), but in fact there is an important distinction between them. Data is raw input for some sort of statistical analysis. A list of all of the traffic accidents in New Jersey in 2010, with information about the drivers (e.g., age, blood alcohol content, whether they were using a cell phone at the time of the accident) and the accident (e.g., time of day, weather, number of cars involved) would be data. Unless you are interested in information about a specific accident (say, if you are a lawyer representing one of the drivers), this list is not likely to be terribly informative by itself. To be able to say anything about road safety in New Jersey generally, you would need statistics. Statistics, in this context, are the results of a statistical analysis of the data. *Statistical analysis* does not have to mean some sort of complicated multivariate regression. In many cases, it is simply an average, a percentage, or a frequency. For example, the percentage of accidents that occur during snowstorms, or the frequency of accidents involving teenage drivers, are examples of statistics that could be generated from this data.

Certain pieces of information can be treated as either statistics or data, depending on what the user wants to do with that information. Take, for example, the unemployment rate of the United States in November 1980: 7.5 percent, according to the Bureau of Labor Statistics.[1] This number is a statistic—the product of statistical analysis of the data gathered from individuals in the labor force by the Current Population Survey—and, for a history student writing a paper about how economic conditions affected the 1980 presidential race, it might be all that he needs. However, an economics student who wants to test the hypothesis that changes in the price of oil affect the unemployment rate in the United States might treat that same number as a data point: one of many monthly unemployment rates that she will use to run regressions, thereby generating other statistics.

Subtypes of Data: Microdata and Aggregate Data

The term *microdata* is used to refer specifically to the kind of data that is, unequivocally, data rather than statistics: raw observations, survey responses, and the like that are not the product of any kind of statistical analysis or summary. *Microdata* often refers to data about individual people. A spreadsheet where each row contains a single person's responses to the questions on a survey is an example of microdata. The traffic accident data mentioned above would also be considered microdata, as would information about individual stores collected as part of the Economic Census or daily rainfall totals for a specific location as reported by the National Weather Service.

The converse of microdata is *aggregate data*—data produced by some sort of statistical procedure, such as averaging or, in the most basic and perhaps the most common example, simply adding up the number of cases. Monthly unemployment rates are an example of what might be referred to as aggregate data, as are election results by precinct and data about retail establishments by county. If the description of a data table ends with "by state," "by gender," or something similar, you are almost certainly dealing with aggregate data.

Public-Use Data versus Restricted Data

In this book, I try to focus on freely available data, where "freely" means two different things: that no monetary payments are required to access the data, and that there are no onerous restrictions on who may have access to the data or the conditions under which they may use the data. Nevertheless, some of the resources mentioned in this book contain both *public-use data* (data with either no restrictions at all on access or with minimal registration requirements) and *restricted data* (data for which access is conditional and requires an approval process). Accessing restricted data may be as simple as filling out a short online form and waiting for approval, or it may be as complicated as completing an extensive certification process and traveling to a designated data enclave (a secure facility with equipment and policies designed to prevent the unauthorized sharing of confidential information) to use the data.

Why is some data public-use and other data restricted? The section "What Data Is Not Disseminated?" (p. 10) explains how concerns about privacy and confidentiality can lead to restrictions on data access.

Surveys, Censuses, and Administrative Data

Surveys are one of the most common methods of gathering data in the social sciences. In a survey, data is gathered from a *sample* (a small subset) of the population (sometimes called the *universe*), and that data is then used to make estimates about the entire population. For example, a typical public opinion survey might do telephone interviews with one thousand people (the sample) randomly selected from all adults over 18 years of age who reside in the United States (the universe). These thousand people's responses would then be used to estimate how the entire adult population of the United States feels about, say, the president's job performance. This contrasts with a *census*, which is often conducted like a survey in that people are asked to answer questions over the telephone or to fill out a form, but, instead of contacting a sample of the population, the goal with a census is to contact every single person (or, in the case of the Economic Census, the Census of Agriculture, or the Census of Jails, every single institution) in the population. It also contrasts with *administrative data*—in which official records (say, birth certificates, tax returns, or customs declaration forms) are used to gather data, rather than asking people or institutions directly to provide information about themselves.

Cross-Sectional, Longitudinal, and Time-Series Data

Many studies gather data at only a single point in time: a survey is written, people respond to it over a few days or weeks, the data is analyzed, and then the study is complete. This type of data is relatively cheap and easy to gather, but it is difficult or impossible to use it to examine changes over time. These types of studies are called *cross-sectional* studies or surveys. Because so many studies are cross-sectional, the types of data that are collected over time are more challenging to find. However, because they can be used to do more sophisticated types of analyses, they are particularly valuable to researchers.

The two main types of data that have been collected over time are *longitudinal data* (sometimes called *panel data*), which follows the same individuals (the "panel") for months, years, or sometimes decades; and *time-series data*—any data collected at relatively regular intervals over an extended period of time. The major macroeconomic indicators—such as gross domestic product and the monthly unemployment rate—are time-series data, since they have been reported monthly or annually for many decades. Daily stock prices for a group of companies are also time-series data. Certain ongoing surveys and public opinion polls intentionally ask the same question in the same way over many years, which creates a time series of public opinion on certain topics. Longitudinal data is relatively rare, although several longitudinal data sets are mentioned in chapter 20. Time-series data on a variety of topics is readily available.

Unit of Analysis and Unit of Observation

In the context of data reference, the distinction between *unit of analysis* and *unit of observation* is subtle but important. The unit of analysis is the type or level of thing that the patron wishes to study. For example, in an education study, plausible units of analysis could be students, teachers, schools, school districts, states, or countries. The unit of observation, on the other hand, is the type or level of thing about which the original researchers gathered data. Different units of analysis often, but not always, require different units of observation in the data sets. For example, a researcher who wants to compare average test scores in different school districts (using school districts as the unit of analysis) would want different data than a researcher who wants to study how students' socioeconomic backgrounds affect their test scores (using students as the unit of analysis), even though both might approach the reference desk asking for "data about students' test scores." The patron who wants to compare test scores across school districts may be able to use data with individual students as the unit of observation, as long as the data file contains a school district for each student; the patron could use that data to calculate average test scores by school district. The opposite is not true: the patron who wants to use students as the unit of analysis is not likely to be satisfied by data with school districts as the unit of observation.

North American Industry Classification System Codes

North American Industry Classification System (NAICS) codes (www.census.gov/eos/www/naics/) are used by the governments of the United States, Canada, and Mexico as well as some private data sources, to classify businesses and workers into industries for statistical purposes. These codes, which range in length from two digits (designating broad sectors) to six digits (designating specific industries) are arranged hierarchically, so that adding additional digits to the end of a broader code allows one to designate a more narrowly defined industry within the sector. For example, by adding digits to 62, "Health Care and Social Assistance," one can move to 622, "Hospitals," and then to 6222 "Psychiatric and Substance Abuse Hospitals." For many industries, the most specific code is actually a five-digit code (or occasionally even a four-digit code), and the five-digit codes are the most specific codes that can reliably be used for comparisons with data disseminated by Mexico and Canada. In the cases where a five-digit code is the most specific, some data interfaces offer an identically labeled six-digit code with a "0" on the end as an option. The examples below illustrate the system, using the 2012 version of NAICS:

44–45	Retail Trade
445	Food and Beverage Stores

4452	Specialty Food Stores
44529	Other Specialty Food Stores
445291	Baked Goods Stores
51	Information
515	Broadcasting (except Internet)
5151	Radio and Television Broadcasting
51512	Television Broadcasting
515120	Television Broadcasting

Different sets of classification codes are used by the United Nations and other organizations, and also for international trade data. Although the codes differ, the basic concept of adding digits to numbers to designate more specific industries or products is incorporated into all of the systems. These systems include the Standard International Trade Classification (SITC), a product classification (e.g., "beverages," "textile yarn") that is managed by the United Nations; International Standard Industrial Classification (ISIC), an industry classification (e.g., "manufacture of beverages," "spinning, weaving and finishing of textiles") that is also managed by the United Nations; the Harmonized System (HS), a product classification that is managed by the World Customs Organization; and Schedule B, product classification managed by the U.S. Census Bureau that is based on the HS. These codes are primarily encountered in the sources listed in chapter 16.

WHO GATHERS AND DISSEMINATES DATA?

Data collection and dissemination are expensive, time-consuming enterprises. Developing and testing a survey, hiring and training interviewers, paying postage or telephone charges, following up with people who do not respond, cleaning and analyzing the data: the costs add up quickly. Plus, in many areas it is not a profitable activity for the organizations that undertake it; much of the time, the decision to gather and share data is driven by academic or administrative motives rather than profit. These two factors of data gathering and data sharing—high costs and often low pecuniary rewards—mean that in many areas governments and intergovernmental organizations with large budgets and no need to show a profit are the best or only source of data. This is especially true for data about small geographic areas, such as provinces and counties. Thousands of times more respondents are needed to be able to calculate accurate estimates for each of the three thousand-plus counties in the United States than to calculate an accurate estimate for the entire United States, for example, which makes it substantially more expensive to conduct a survey capable of producing county-level data than one intended to produce only national data.