# 11

# How to Complete Your Dissertation Using Online Data Access and Collection

In the first edition of this book, we extolled the virtues of the personal computer as a time-saving tool for data analysis, writing, literature searching, and organization but suggested that it might be possible for someone to complete a dissertation without using a personal computer. In the second edition, we suggested that no one attempt to complete a dissertation without a personal computer, a host of software, and Internet access. In addition, we described the various software options available for literature review and bibliographic management, as well as quantitative and qualitative analyses, and discussed the need for and functions of utilities to safeguard your computer. We have now moved much of this information to tables and tip boxes that contain references and links to relevant software and database sites. Today's students are sophisticated computer users with experience that typically begins in grade school, and the computer has revolutionized not only how they conduct research but also how they conduct their daily lives. Every stage of the research process is now routinely conducted via web-based applications. Web-based survey design and analysis are rapidly replacing the traditional methods of mail, phone, and face-to-face survey administration, and qualitative research is as likely to be derived from online contact with participants as from face-to-face contact. We have reserved this chapter to discuss methods of online data access, analysis, and collection.

To begin, recall the difference between primary and secondary data analysis. *Primary data analysis* refers to the analysis of data collected by the researcher or by the researcher's trained observers or interviewers. *Secondary analysis* draws upon data collected by other researchers, often for other purposes, or data created by nonresearchers outside the specific context of research. For example, data from the U.S. Census; data collected by the Gallup, Roper, or Field Polling organizations; data collected by federally funded research grants; and data from numerous other sources frequently are made available on the Internet for the explicit purpose of allowing researchers access to them. The first section below describes strategies for obtaining access to secondary data via the Internet and offers some advice regarding the use of secondary data.

## The Pros and Cons of Accessing Secondary Data via the Internet

Conducting secondary analyses of data downloaded from a website is becoming commonplace for all researchers, not just students engaged in dissertation research. The typical process involves the following steps:

1. Locate the site containing the desired data.
2. Obtain the necessary passwords, if any.
3. Master the download format or data extraction system.
4. Download the data.
5. Access the downloaded data with statistical software.

Secondary data analysis is not a new idea. What *is* new is the amount of data available and the ease of access to these data. The world's largest archive of computerized social science data is available from the Inter-university Consortium for Political and Social Research (ICPSR), located within the Institute for Social Research at the University of Michigan (www.icpsr.umich.edu). Here you will find information about how to join the ICPSR and download data from its vast database holdings, as well as a list of member institutions, training in quantitative methods to facilitate effective data use, and other data archive material. Hundreds, if not thousands, of dissertations have been completed using data accessed from the ICPSR holdings. Tip Box 11.1 provides information about the ICPSR. The creation of large archives of data such as the ICPSR

resulted in the coining of a new word, *dataverse*. A dataverse is an archive of research studies. In addition to exploring the ICPSR, the interested reader should visit the Harvard Dataverse Network (http://thedata.harvard.edu/dvn/), which hosts over 53,000 studies representing data from all disciplines worldwide, including "the world's largest collection of social science research data."

**Tip Box 11.1**

### The ICPSR Data Archive (Inter-university Consortium for Political and Social Research)

ICPSR maintains and provides access to a vast archive of social science data for research and instruction and offers training in quantitative methods to facilitate effective data use. ICPSR provides a searchable database of its archival holdings as well as direct downloading of data for member institutions. More than 500,000 data files are available from this site.

- Main Site: www.icpsr.umich.edu

ICPSR also cosponsors the following special topics archives:

- Health and Medical Care Archive (HMCA): www.icpsr.umich.edu/HMCA
- National Archive of Computerized Data on Aging (NACDA): www.icpsr.umich.edu/NACDA
- National Archive of Criminal Justice Data (NACJD): www.icpsr.umich.edu/NACJD
- Substance Abuse and Mental Health Data Archive (SAMHDA): www.icpsr.umich.edu/SAMHDA

The holdings of the U.S. Census Bureau can be found at www.census.gov/main/www/access.html. Here you will find a wealth of information, including "Interactive Internet Data Tools" and a link to "Direct File Access" that will allow you to download Census 2000 data sets. We also recommend the IPUMS (Integrated Public Use Microdata Series) at the University of Minnesota Population Center (www.ipums.org). Here you will find a national census database dating from 1850 and an international database containing census data from around the world. Along with this ready availability of preexisting data come both great advantages and

some serious risks. We first describe our perspective on the advantages of accessing secondary data, and then we add some caveats that should be considered when using secondary data.

First and foremost, secondary data are likely to be of much better quality than any graduate student could independently collect. Research is expensive, large-scale survey research is extremely expensive, and large-scale longitudinal survey research is both exorbitantly expensive and impossibly time prohibitive for a graduate student. Yet such data are easily accessible from the Internet. For example, the National Longitudinal Survey of Youth 1997 (NLSY97) is part of the National Longitudinal Surveys (NLS) program, a set of surveys sponsored by the U.S. Department of Labor, Bureau of Labor Statistics (BLS). These surveys have gathered information at multiple points in time on the labor market experiences of diverse groups of men and women. Each of the NLS samples consists of several thousand individuals, some of whom have been surveyed over several decades. These surveys are available from the National Archive of Criminal Justice Data, accessed through ICPSR (www.icpsr.umich.edu/NACJD/). Thus, one of the first advantages of using secondary data is that at virtually every point of the research process, the data are of better quality than an individual graduate student could collect. Tip Box 11.2 provides an extensive list of links to data archives and sites that contain links to other data archives.

A second reason to make use of secondary data sources is the fact that the costs of collecting primary data are often greater than the resources of most graduate students, even those with substantial funding. Collecting high-quality quantitative data is both expensive and time-consuming. This is why many doctoral dissertations are based on cross-sectional data with small, nonrandom samples. This severely limits the ability of these studies to make causal inferences, to generalize to a known population, or to achieve sufficient statistical power. Unfortunately, unless these studies are unique in some other way, they are often not publishable in peer-reviewed journals.

Given the above two glowing recommendations for accessing secondary data via the Internet, why wouldn't everyone want to do so? The first reason is that the data may not contain the measures the researcher needs to directly address the primary research questions of interest. This results in either changing the questions to "fit" the available instrumentation or trying to create an instrument from parts of someone else's instruments.

These "derived measures" do not have the history of reliability and validity studies that support an established measure.

The second reason one might not want to use a secondary data source is that working with one can be difficult and frustrating. Despite the fact that most archive sites attempt to make data easy to locate and provide extensive documentation, the downloading process is not necessarily simple. Be prepared to encounter difficulties. Not all archive sites use the same process for downloading or documenting data, and it might be necessary to learn several methods for doing so. In addition to the potential frustration involved in downloading and accessing a database with your own software, you may also need to struggle with inadequate or inaccurate documentation. You must have access to a complete description of the sampling design, accurate and complete codebooks must be available, and when all else fails, you must have a resource to contact for help when you encounter flaws in the available documentation.

A third reason to avoid secondary data is that your dissertation committee may not approve its use. We find that the reasons for this are twofold. First, the dissertation process is designed to be many things. In addition to representing a new contribution to a field of inquiry, the dissertation is a valuable training vehicle for researchers. Your committee may feel that completing a dissertation should involve full participation in all phases of the research process, including phases that are likely to be skipped if you use secondary data. These include developing a sampling design; designing an instrument; collecting, entering, and cleaning data; and building a database. We would hope, incidentally, that students who do pursue secondary data analyses for their dissertation have previously learned these important skills elsewhere in their graduate student programs. Second, some dissertation advisers have had bad experiences with students who have attempted to use secondary data. These derive primarily from the first two reasons for avoiding secondary data cited above: a lack of a match between the questions the student wants to ask and the questions the data can answer and difficulties with downloading and using the data.

We believe that under the right circumstances, using secondary data is a reasonable and acceptable approach; however, the student must do his or her homework first. You cannot assume that the perfect database will be easy to find and waiting for you. You must search databases at multiple sites for the perfect match between your interests and research

questions and the available data. Tip Boxes 11.1 and 11.2 list some of the largest archives of social science data, but no list can be comprehensive because new archives appear on a regular basis. You might find your ideal database on a site that contains only a few unique holdings. For example, in 1921–1922, during the waning of the eugenics movement and its hereditarian interests, psychologist Lewis Terman launched a study to investigate the maintenance of early intellectual superiority over a 10-year period. This objective was soon extended into the adult years for the purpose of determining the life paths of gifted Californians. So far, 13 waves of data collection have been carried out, beginning in 1921–1922 with interviews of parents and the study children and an array of tests and inventories. The first 1922 and 1928 data collections focused on family life and school experience, and they included interviews and questionnaires administered to mothers of children in the study. (At that time, fathers were not thought to be particularly important in child rearing, so fathers were not included among the respondents.) Various life changes within the Terman sample and new leadership from Robert Sears, Lee Cronbach, Pauline Sears, and Albert Hastorf brought fresh attention to issues of aging, work life and retirement, family, and life evaluation, which were explored in follow-ups in 1972, 1977, 1982, 1986, and 1991–1992 (Lewis Terman Study, n.d.). For a more detailed description of this study and access to the Terman data files, see http://dataserv.libs.uga.edu/icpsr/8092/8092.html.

A more thorough treatment of secondary analysis, including web-based data downloads, can be found in *Research Methods in the Social Sciences* by Chava Frankfort-Nachmias and David Nachmias (2008, Chapter 13). If you would like to give web-based data collection serious consideration as a primary research approach, we recommend starting with the ICPSR. You might also consider the University of Georgia Libraries list of data archives (http://dataserv.libs.uga.edu/datasite.html) or the University of California–San Diego Intro to Data and Statistics Research site (http://libguides.ucsd.edu/content.php?pid=221125&sid=1835576), which contains numerous links to frequently used statistics tools, databases, and recommendations regarding statistical software. Finally, you might access Data on the Net (http://3stages.org/idata/), a site with links to 156 other sites that have numeric data ready to download; 54 data libraries and data archives from around the world; 53 catalogs and lists of data from data libraries, archives, and vendors; and 20 organizations that sell and distribute data for a fee.

## The Internet as a Primary Data Collection Resource for Quantitative and Qualitative Data

One can use the Internet as a primary data collection resource in a number of ways. First, the Internet can be accessed as a source of archived records and information. For example, you may wish to obtain articles from the *New York Times* that mention a certain key event. For this purpose, the ProQuest database contains the *New York Times* data archive, searchable by topic or keywords. One of our students used this database to search for articles describing deaths during the Vietnam War period (Huston-Warren, 2006). Second, the Internet can be accessed to monitor unobtrusively online communications that occur in chat rooms, web logs (blogs), online dating services, and other sources of online communications. For example, Stephanie MacKay (2012) examined the discourses of femininity circulating on a female skateboarding blog produced by the Skirtboarders (a group of women skateboarders based in Montréal, Canada).

Third, you may wish to design and publish your own data collection instrument on the Web. Only a few years ago this was a major task, best undertaken only by an experienced webpage designer. However, a large number of Internet-based survey design and data-archiving services are now available to assist researchers with constructing Internet-based surveys and obtaining complete databases in return. This type of service reduces much of the drudgery of having to enter data into a database, as well as the costs of mailing, printing, and entering data. In addition, many sites facilitate online design to make the process of creating a survey simple, fast, and intuitive. Tip Box 11.3 provides links to online survey research companies with information, pricing, and services. Note that the selection of an appropriate service should not be based solely on price. We recommend that you explore the WebSM (Web Survey Methodology) site www.websm.org. WebSM is dedicated to the methodological issues of web surveys. It covers the broad area of interaction between modern technologies and the survey data collection process. Here you will find a web survey bibliography containing the most recent publications exploring web survey methodology and links to survey software and companies that provide web survey services. There are many factors to consider when designing a web-based survey that affect both response rate and data quality. Among these are the sponsor of the survey, the topic, time to completion, and the web-based presentation (Casey & Poropat, 2014; Fan & Yan, 2010; Sánchez-Fernández, Muñoz-Leiva, & Montoro-Ríos, 2012).

Of particular importance, according to Fan and Yan, is that any service you select needs to support different browsers.

> Often, the same web questionnaire could be displayed differently to respondents in different computer configurations, different web browsers, different Internet services, and different Internet transmission capabilities. Because of these variations, some respondents may not be able to browse the questionnaires normally, submit their answers successfully, or even quit the surveys eventually. In addition, it is also important that survey software programs support diverse formats such as XLS and SPSS for effective data importation and data exportation. (p. 137)

Our students have had some experience with the use of online services for survey design and data collection. If you do use a website service for collecting and/or analyzing your data, we caution you to vigilantly monitor the process yourself and not rely on the website proprietor to do so. For instance, one of our dissertation students was seeking a sample size of 200 to complete a survey. To reach a broad audience, she contracted with a web-based research company to host her survey so that participants could complete the scales online. To optimize the likelihood of obtaining her required sample size within a reasonable period of time, she offered a $20 fee to any qualified respondent (young women with eating disorders) who completed the survey. One day after she offered the financial inducement, the website proprietor informed her that she had 576 completed surveys and several hundred more in process. Both aghast and delighted, she shut down the site immediately, realizing that she now had a terrific sample of participants but owed them almost $12,000 out of her own pocket! She was, of course, ethically obligated to pay all legitimate respondents who completed the task. Her mistake was in not setting a limit on the sample size beforehand, an understandable but costly error.

Based on our experience with the above and similar problems encountered by our students, we offer the following recommendations:

1. Research the available online survey companies carefully. Services and pricing vary widely. The more expensive sites may offer no more than less expensive ones, or they may offer more than you need.
2. Once your survey instrument is online, check every word of every question and all response choice categories against your original instrument. Errors and omissions are common, and it is your responsibility to locate them or suffer the consequences.
3. Do a trial run with any online survey or experiment because once it is connected to the Internet, small errors can have large ramifications. If possible,

obtain a small "trail" data set to make sure that the download includes everything. Before going "live," have a few friends and your committee complete the online survey and make recommendations.

4. Use a "stop order" to avoid being overwhelmed with data unless collecting additional data is free. After all, there are advantages to having larger sample sizes.
5. Include numerous checks to make certain that those who complete your survey are actually part of your desired population. For example, if you only want to study males, then include instructions at the beginning indicating that only males are eligible to complete the survey and then also ask for respondents' gender later in the survey.
6. Any incentive, monetary or otherwise, is likely to result in multiple responses from the same respondent or responses from ineligible respondents. Discuss options for preventing this potential problem with the service provider and insert design elements in your survey designed to detect cheaters.
7. Remember that a proposal calling for a given number of responses typically does not take into consideration ineligible respondents or incomplete data. Plan to eliminate at least 20% of your cases due to incomplete data or unusable responses.
8. If possible, restrict access to only those truly eligible, or highly likely to be eligible, to complete your survey. Do this by providing an access code or password so that only those you have contacted and determined to be eligible in advance can complete the survey.

There are some important questions to ask before deciding to collect data through the Internet. First, are you comfortable enough with your computer skills to work through the inevitable difficulties that will arise? Second, do you have a clear idea of the population to be sampled and how you can reach this sample through your computer? Third, is your dissertation committee accepting and supportive of your proposed data collection strategy? Given positive answers to these questions, you may wish to consider data collection via the Internet.

Useful books include Best and Krueger's *Internet Data Collection* (2004); Dillman, Smyth, and Christian's *Internet, Mail, and Mixed-Mode Surveys: The Tailored Design Method* (2009); and Birnbaum's *Introduction to Behavioral Research on the Internet* (2001). The last resource provides helpful guidance for setting up online experiments as well as surveys. More advanced references include Gosling and Johnson's *Advanced Methods for Conducting Online Behavioral Research* (2010), an edited volume with many examples of Internet-based research, including a chapter on collecting data

from social-networking websites and blogs, a chapter on using automated field notes to observe behavior of online participants, and an entire section describing strategies for transporting traditional methodologies to the Web. Finally, the reader may wish to consider the *SAGE Handbook of Online Research Methods* (Fielding, Lee, & Blank, 2008), an edited volume that contains a section titled "Virtual Ethnography," which describes topics such as Internet-based interviewing, online focus groups, and distributed video analysis.

But what about the reliability and validity of web-based surveys, and what about the skepticism of committee members wary of this approach? Can web-based surveys be trusted? These questions have been the focus of considerable research, and answers are beginning to emerge (c.f. Duarte Bonini Campos et al., 2011; van den Berg et al., 2011). Gosling, Vazire, Srivastava, and John (2004) conducted a comparative analysis of a large Internet-based sample ($N$ = 361,703) with 510 published traditional samples to address six preconceptions about Internet questionnaires. Table 11.1 summarizes their findings.

Services that offer web-based surveys are aware of the possibility that some respondents may complete multiple surveys to receive rewards or incentives for participation and that some respondents may not be fully engaged in the survey. Thus, some services are able to monitor who completes the survey and their level of commitment. One of our students, Jennifer Johnston (2013), described the procedures involved in launching her survey on Zoomerang (now merged with SurveyMonkey) as follows:

> A computerized version of the questionnaire was launched on Zoomerang .com. Respondents first encountered a welcome letter, followed by an informed consent page that required their endorsement of a radio button acknowledging comprehension of their eligibility and rights as a participant, before they could begin. Confidentiality was assured by Zoomerang .com and no personally identifying information such as name, home address or state, including URL origination, was released to this researcher. Then the respondents entered the survey itself and if quotas had not yet been filled, they were allowed to continue beyond the initial demographic questions. If quotas were filled, the respondent was informed that participation in the survey was no longer needed. A percent completion indicator was visible on each page of the questionnaire. At the end of the survey, respondents were directed to the debriefing page which again included researcher, faculty chair, and IRB contact information. The questionnaire took about 25 minutes to complete.
>
> Zoom panel members received the customary 50 "zoompoints" for completing the questionnaire. Zoompoints cumulate so that panel

**Table 11.1** Six Preconceptions About Internet Methods

| Preconception | Finding |
|---|---|
| 1. Internet samples are not demographically diverse (e.g., Krantz & Dalal, 2000). | *Mixed.* Internet samples are *more* diverse than traditional samples in many domains (e.g., gender), though they are not completely representative of the population. |
| 2. Internet samples are maladjusted, socially isolated, or depressed (e.g., Kraut et al., 1998). | *Myth.* Internet users do not differ from nonusers on markers of adjustment and depression. |
| 3. Internet data do not generalize across presentation formats (e.g., Azar, 2000). | *Myth.* Internet findings have been replicated across two presentation formats of the Big Five Inventory. |
| 4. Internet participants are unmotivated (e.g., Buchanan, 2000). | *Myth.* Internet methods provide means for motivating participants (e.g., feedback). |
| 5. Internet data are compromised by anonymity of participants (e.g., Skitka & Sargis, 2006). | *Fact.* However, Internet researchers can take steps to eliminate repeat responders. |
| 6. Internet-based findings differ from those obtained with other methods (e.g., Krantz & Dalal, 2000). | *Myth?* Evidence so far suggests that Internet-based findings are consistent with findings based on traditional methods (e.g., on self-esteem, personality), but more data are needed. |

*Source:* Gosling, Vazire, Srivastava, & John, 2004, pp. 93–104.

members can purchase music, books, electronics, cookware, and so on. Fifty Zoompoints is equivalent to about $1.00 toward a purchase. This level of compensation generally motivates panel members to complete surveys in their free time, but is not significant enough to create coercion. The Zoom panel is managed and maintained by Market Research Tools, the research firm that owns Zoomerang.com. They maintain and monitor the "health," accuracy, and usefulness of the panel using some traditional panel management techniques, such as tracking representativeness, panelist frequency of activity, and boosting recruitment when panel members leave, but also their own patented technological solution to panel fraud, employing algorithms which detect whether each respondent is unique, purportedly real, and engaged—called "true sample technology." Furthermore, although

> the panel is predominantly recruited for market/product research, the members are never solicited or marketed to buy products or services (MarketTools, February 2009). (p. 54)

The Internet can also support the implementation of true experimental designs. For example, Sánchez-Fernández et al. (2012) examined factors that affect response rates to web-based surveys by systematically varying the frequency of email reminders, personalizing emails sent to potential respondents, and using an incentive to create a balanced $2 \times 2 \times 2$ factorial design. Thus, with a little creative thinking, a researcher can utilize the Internet not only to conduct a cross-sectional survey but also to implement longitudinal studies and perform experimental and quasi-experimental research.

Finally, we wish to mention a more recent technology known as *crowdsourcing*. Crowdsourcing refers to "obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community" (www.wikipedia.com). Amazon Mechanical Turk (MTurk, www.mturk.com) allows individuals, businesses, or any other group that needs a task performed over the Internet to solicit "workers" to perform that task for a set fee. In MTurk, the tasks are referred to as Human Intelligence Tasks, or HITs. If you wish to have your questionnaire completed, you first enroll as a "Requestor" and subsequently ask "Workers" to do so for a set fee. Through the use of MTurk, the requestor can specify the exact qualifications that a respondent must fulfill before accepting the HIT and may accept or reject the result. You can set the price you are willing to pay for each completed HIT, a penny or a dollar, or more, but you will be required to pay 10% of the price of successfully completed HITs to Amazon. The use of MTurk as a source of participants in social science research has not gone unnoticed, and some unique and interesting research has been conducted, including into the characteristics of the samples obtained through this method. Buhrmester, Kwang, and Gosling (2011) concluded that

> (a) MTurk participants are slightly more demographically diverse than are standard Internet samples and are significantly more diverse than typical American college samples; (b) participation is affected by compensation rate and task length, but participants can still be recruited rapidly and inexpensively; (c) realistic compensation rates do not affect data quality, and (d) the data obtained are at least as reliable as those obtained via traditional methods. (p. 3)

One of our students, Lauren White (2013), utilized MTurk to explore video game play and verbal reasoning. By assigning participants randomly to hyperlinks directing participants to a randomly assigned video game, she was able to develop a mixed factorial design with verbal reasoning accuracy as the dependent measure. All survey materials were stored at SurveyMonkey, with a hyperlink to the survey given at MTurk. White described her experience as having both benefits and drawbacks.

> The most prominent benefit was the rapid response rate gathered to reach the target *N* for each study. Data collection lasted for approximately 10 days for each study, with almost 200 complete survey materials for each study; conducting this type of data collection in person would have taken much longer to complete. (p. 79)

She added:

> Participants were compensated $0.10 for completed test materials. The survey materials were estimated to take approximately 30 minutes to complete. AMT estimated average wages for duration of participation; based on this, $0.10 compensation for a 30-minute survey was a low cost. This may have deterred quality participants from completing the survey materials and instead caused participation fatigue and reduced motivation. Furthermore, towards the end of the survey (i.e., the post-test verbal reasoning questions), there was an increase in randomly guessed responses. Study 1 contained 6AFC, with 16% indicating random guessing; study 2 contained 5AFC, with 20% indicating random guessing. (p. 79)

White (personal communication) offers the following suggestions regarding the use of Amazon Mechanical Turk, some of which represent advice all users of Internet surveys should consider:

1. Keep survey materials as clear and easy to manage as possible. The harder the materials, the more complex the tasks for AMT users. If tasks are too hard or unclear, you'll get poor results or missing data from your survey.
2. Monitor the rate at which the survey is completed. Most responses start coming in within several hours of the survey being posted. If you notice there are not many responses, repost the survey or change some of the qualification criteria.
3. When setting up an AMT survey, AMT will provide an estimate of how much you should be paying people for completing your survey materials. Try to pay as much as possible, but not too much.
4. AMT has great FAQs and guidelines on its site—very helpful info there too!

## Big Data: What Is It, and Does It Make Sense as a Source of Dissertation Data?

Our discussion of the availability of secondary databases and Internet-based research would be incomplete without some mention of "big data." The Internet makes it possible to collect, organize, and share large amounts of information in a way that has never been possible before. According to Cukier and Mayer-Schoenberger (2013),

> As recently as the year 2000, only one-quarter of all the world's stored information was digital. The rest was preserved on paper, film, and other analog media. But because the amount of digital data expands so quickly—doubling around every three years—that situation was swiftly inverted. Today, less than two percent of all stored information is nondigital. (p. 28)

This trend, referred to as "datafication" by Cukier and Mayer-Schoenberger, has created new opportunities for research in both the natural and social sciences, but it has also raised some interesting questions about how researchers approach their craft. One of these is to challenge the notion of "sampling." Our undergraduate research courses taught us that we sample because it is either impossible or excessively expensive to use the entire population. Our undergraduate statistics courses also taught us to use null hypothesis significance testing to enable "scientific" generalization from samples to populations. However, with big data we are much less constrained by the limits of small samples and the statistics of generalization. For example, most of the U.S. Census can be downloaded by anyone with a computer large enough to store it. Moreover, in databases with millions of cases, virtually any relationship will be statistically significant. New statistical approaches may make more sense when the limits of size and cost no longer exert constraints on our ability to collect, store, organize, and analyze data.

Data mining, the science and art of discovering patterns and extracting relationships hidden in big data, as opposed to testing causal hypotheses, has recently become a popular topic (c.f. Han, Kamber, & Pei, 2011; Tan, Steinbach, & Kuman, 2013). Of particular interest is Russell's (2013) *Mining the Social Web: Data Mining Facebook, Twitter, LinkedIn, Google+, GitHub, and More*. Of course, such "blind empiricism" has not been without its critics, and in the absence of a sound theoretical framework, blindly searching a large database for potential patterns may be unsuitable for most dissertations. One critic, Huba (2013), stated:

> Big data this, big data that. Wow. At the end we will have better ways to sell underwear, automobiles, and "next day" pills (although in the latter case politics and religion might actually trump Amazon and Google). Blind empiricism. Every time you click a key on the Internet it goes into some big database.
>
> "Little data"—lovingly crafted to test theories and collected and analyzed with great care by highly trained professionals—has built our theories of personality, social interactions, the cosmos, and the behavioral economics of buying or saving. (paras. 1–2)

The data and how they are used depend on the goals, ingenuity, and skill of the researcher. Thus, we encourage you to consider all options for data sources as you pursue your research interests.

## Internet-Based Data Analysis

A more recent advance in the use of online data archives involves bypassing the downloading process altogether and analyzing data while online. A number of data archive sites now offer this option. For example, at both www.icpsr.umich.edu/icpsrweb/ and http://sda.berkeley.edu, the user may choose to either select and download many types of data or conduct a wide range of data management and statistical procedures while online. These procedures include recoding and computing variables, cross-tabulation, mean comparisons, and multiple regression analysis.

A particularly interesting approach to online data analysis has been developed by the American Institutes for Research (AIR). AIR Lighthouse (http://lighthouse.air.org/timss/) empowers users to ask their own questions of complex data sets without specialized research or statistical skills. Users can create custom-run tables, graphs, and other statistical presentations over the Internet. Lighthouse integrates multiple complex surveys, assessments, and other data collections and captures the knowledge of expert statistical analysts. This knowledge database is stored along with the data themselves. To the user, the system seems to "know the data" and to choose the right analytic procedures. This knowledge enables the system to hide the technical details and sophisticated statistical procedures from the user, who sees only perfectly tailored answers to his or her queries. Anyone interested in online data analysis should explore this creative approach to statistical analysis. Of course, the "wisdom" inherent in the Lighthouse system does not alleviate you of the responsibility of being familiar with the statistical methods that were employed in analyzing your data.

## Tip Box 11.2

### Data Archives and Libraries

| Name of Site | URL | Comments |
|---|---|---|
| Center for Demography and Ecology (University of Wisconsin–Madison) | www.ssc.wisc.edu/cde/ | This research cooperative boasts "one of the country's finest collections of machine-readable data files in demography." From the home page, search "Signature Themes" and then "Data." |
| Cornell Institute for Social and Economic Research (Cornell University) | www.ciser.cornell.edu/ASPs/datasource.asp?CATEGORY=2 | Direct access to selected data sets available on the Internet and links to dozens of similar sites. |
| Center for International Earth Science Information Network (CIESIN) | www.ciesin.org | Data on world population, environment, health, and geography. Includes several interactive systems to search for data. See "Data & Information" link. |
| Consortium of European Social Science Data Archives | www.cessda.org | Search for data contained in archives around the world. |
| Data & Information Services Center (University of Wisconsin–Madison) | www.disc.wisc.edu | A collection of social science and cross-disciplinary data files. Longitudinal surveys, macroeconomic indicators, election studies, population studies, socialization patterns, poverty measures, labor force participation, public opinion polls, education and health data, and census data comprise the scope of the collection. |
| Harvard Dataverse Network | http://thedata.harvard.edu/dvn/ | A repository for research data where researchers can "Share, Cite, Reuse, Archive Research Data." Searchable via a wide variety of keywords and topics. |
| The Odum Institute (University of North Carolina) | www.irss.unc.edu/odum/home2.jsp | Public opinion data from the Louis Harris polls, Carolina and Southern Focus Polls, and |

| Name of Site | URL | Comments |
|---|---|---|
| | | the National Network of State Polls. Includes a searchable database to retrieve questions and frequencies. Selected data files also available for downloading. Click the "Data Archive" link. |
| International Social Survey Programme (ISSP) | www.issp.org | Cross-national collaboration on social science surveys in 34 countries. See "Archive and Data." |
| National Archives and Records Administration—Center for Electronic Records | www.archives.gov/research/ | Information regarding electronic records, including numeric data files, generated by U.S. government agencies and available for purchase through the National Archives and Records Administration. See "Access to Archival Databases." |
| National Data Archive of Child Abuse and Neglect (NDACAN) | www.ndacan.cornell.edu | Information regarding NDACAN, including its mission, publications, and available data sets. |
| Roper Center for Public Opinion Research | www.ropercenter.uconn.edu | An extensive archive of opinion polls, including Gallup polls and many others. See "Data Access." |
| UK Data Archive (University of Essex) | www.data-archive.ac.uk | Curator of the largest collection of digital data in the social sciences and humanities in the United Kingdom. Archives over 7,000 data sets in the social sciences. |
| Data on the Net (University of California–San Diego) | http://libguides.ucsd.edu/content.php?pid=221125&sid=1835576 | Many links to sites with data, along with lengthy descriptions of what can be found at each site. This directory can be searched by keyword. Make sure to examine "Multidisciplinary Data Finders." |

*(Continued)*

(Continued)

| Name of Site | URL | Comments |
|---|---|---|
| Specific Sites for Economic, Social, and Political Data | | |
| Gallup | www.gallup.com | Public opinion data from Gallup, including some tables and statistics and articles from the company's newsletter and other reports. |
| General Social Survey | www.icpsr.umich.edu/GSS/ and http://sda.berkeley.edu | Information regarding the biennial personal interview survey conducted by the National Opinion Research Center (NORC). Includes a search engine to search the codebook for relevant variables and an extract utility to select data. Numerous other databases available at both sites. Extensive online data analysis options. |
| Federal Election Commission | www.fec.gov | Offers downloadable data on campaign financing. See the Campaign Finance Disclosure Portal. |
| The World Bank | http://data.worldbank.org | Household surveys for numerous countries. Access conditions vary by country. Click the "DATA" link and browse data by country, topic, or economic indicators. |
| American National Election Studies (ANES) | www.electionstudies.org | ANES conducts national surveys of the American electorate. The time-series of studies now spans 5 decades. This site provides information on the mission and procedures of the ANES and other documentation. |
| Panel Study of Income Dynamics | http://psidonline.isr.umich.edu | Information regarding the Panel Study of Income Dynamics, a longitudinal study of American families ongoing since 1968. Covers topics such as employment, income, wealth, housing, and health. |

| Name of Site | URL | Comments |
|---|---|---|
| Uniform Crime Reports (University of Virginia) | http://fisher.lib.virginia.edu/collections/stats/crime/ | Interactive system for retrieving county-level crime and arrest data. |
| United Kingdom Election Results | www.election.demon.co.uk | Provides links to election results from British parliamentary elections since 1983. |
| U.S. Department of Housing and Urban Development | www.huduser.org | Data pertaining to housing needs, market conditions, and community development. |
| Statistics about … (University of Minnesota) | http://govpubs.lib.umn.edu/stat.phtml | Links to selected statistical tables, publications, and indicators arranged by subject. Go to "Statistics," then "Key Databases." |
| USDA Economics, Statistics, and Market Information System | http://usda.mannlib.cornell.edu | Publications and data sets about agriculture available from the statistical units of the USDA: Economic Research Service, National Agricultural Statistics Service, and the World Agricultural Outlook Board. |
| Government Statistical Agencies | | |
| Bureau of the Census: www.census.gov<br>Bureau of Economic Analysis: www.bea.gov<br>Bureau of Justice Statistics: www.ojp.usdoj.gov/bjs<br>Bureau of Labor Statistics: www.bls.gov<br>Bureau of Transportation Statistics: www.bts.gov<br>Economic Research Service: www.ers.usda.gov<br>Energy Information Administration: www.eia.doe.gov<br>FedStats: www.fedstats.gov<br>Centers for Medicare and Medicaid Services: www.cms.gov<br>National Center for Education Statistics: www.ed.gov/NCES<br>National Center for Health Statistics: www.cdc.gov/nchs<br>National Science Foundation, Division of Science Resources Studies: www.nsf.gov/sbe/srs/stats.htm<br>Statistics Canada: www.statcan.ca<br>Statistics Division of the United Nations Department for Economic and Social Information and Policy Analysis: www.un.org/Depts/unsd | | |

## Tip Box 11.3

### Online Services for Survey Design and Data Collection

| Company Name/ Product | Features | Pricing | Service Limitations/ Comments |
|---|---|---|---|
| CreateSurvey www.createsurvey.com | Standard features; educational discount | A personal account is $15 for one month and $199 for one year. Allows up to 10 different surveys with unlimited questions and up to 1,000 respondents per month. | Survey housed on company server for a set amount of time. Definitely worth serious consideration. |
| FormSite www.formsite.com | Weekly survey traffic report; multiple language support | $19.95 up to $99.95 per month depending on desired number of responses, amount of storage, and number of items. Free 14-day trial. | Survey housed on company server for a set amount of time; limited (but large) number of responses per month. |
| HostedSurvey www.hostedsurvey.com | Standard features; educational discount | Charge is prepaid per number of responses; first 50 responses are free, then around $0.45 each up to 1,000. | Survey housed on company server for 18 months from purchase. |
| SuperSurvey www.supersurvey.com | Standard features | Wide variety of pricing options, single survey, monthly and yearly. | A single survey, up to 1,000 respondents, is $19.95 and remains active for 1 year. |
| SurveyMonkey www.surveymonkey.com | Standard features; unlimited surveys | $17 per month for Select plan. Unlimited questions and responses. | World's most popular online survey tool. Has merged with Zoomerang. |

| Company Name/ Product | Features | Pricing | Service Limitations/ Comments |
|---|---|---|---|
| SurveyProf www.surveyprof.com | Free site | Free for students. | We have no experience with this site. |
| FluidSurveys http://fluidsurveys.com | Standard features | A free starter for small surveys and a pro package for $17 per month with unlimited responses. | Mobile applications through smartphones and tablets. |
| Google docs www.google.com/google-d-s/createforms.html | Free site | Free, but requires download of Google Drive. | Thousands of templates to choose from, but very limited in terms of flexibility with no skip patterns. Would recommend only for very short, simple surveys. Output to Excel spreadsheet. |

*Note.* We examined a large number of sites offering Internet-based web hosting of surveys. We excluded many of these because they did not seem designed for students seeking to conduct a single web-based survey. We do not "rank" these services or believe that the above list is comprehensive. We do believe the list is more than adequate to meet the needs of nearly everyone, and we suggest that you explore the sites and options to make an informed decision.