DATA BASICS

an introductory text

by

Diane Geraci
Chuck Humphrey
Jim Jacobs

August, 2012

Statistics? Data? What Are We Talking About?

Librarians traditionally collected statistical abstracts, census tabulations, economic indicators, vital statistics, and a wide range of other statistical information. These materials have tended to be compilations of published tables. More recently, however, statistical information is being acquired in ever increasing quantities over the Internet. As a consequence, this type of material is more readily available to users and librarians alike. In addition, many libraries are now providing access to research data as part of their collections and, through licensed Web services, are making this type of information more directly accessible to the researcher at her or his desktop. With easy access to statistics and research data, librarians need a clear understanding about what these resources are and how they are related.

Popular usage of words with technical meanings can cause confusion about what is being discussed. This Chapter makes a distinction between statistics and data, even though both are commonly viewed as “numbers” and are often used interchangeably. We treat statistics and data, however, as separate types of related information requiring different kinds of library service. Someone providing data services in her library will not necessarily be offering statistical services or vice versa. The notion that different kinds of materials require different types of library support is not new. However, understanding statistics and data as different kinds of material may be new to some. One of the primary purposes of this book is to clarify the types of service needed to support data in the library.

The Origins of Statistics: Official and Non-official

Statistics are generated today about nearly every activity on the planet. Never before have we had so much statistical information about the world in which we live. Why is this type of information so abundant? For one thing, statistics have become a form of currency in today’s information society. Through information technology, society has become very proficient in calculating statistics from the vast quantities of data that are collected. As a result, our lives involve daily transactions revolving around some use of statistical information. For example, statistics about the body weight of passengers are important to airline safety. An article in a Canadian newspaper reported that the average weight of passengers has increased over the past decade. “[T]he Canadian government wants to be sure that average body weights, used to calculate total aircraft loading, are up to date. Transport Canada blamed just that kind of miscalculation for a crash last January that killed 10 people.”[1] This statistic on average body weight has life and death consequences for those traveling by air and exemplifies how statistics have become an important part of a daily activity involving many lives.

[1] “Obese passengers are a costly load,” Montreal Gazette: Montreal, November 8, 2004, p. A 20.

Given the ubiquity of statistics in today's world, where are these statistics coming from? One way to address this question is to group statistics into two categories: official statistics and non-official statistics. The roots of official statistics harkens back to the earliest definition of statistics in the **Oxford English Dictionary**. This definition made reference to the activities of collecting, classifying, and discussing numeric facts about nations. Its usage arose during the 18th Century with the beginnings of the modern nation-state. The statistics mills of these nations compiled figures for the government of the day and the administrators of public services. In this sense, Statistics Canada and Statistics New Zealand, which are the national statistics agencies in these countries today, have been aptly named.

Official statistics may be derived from administrative records, such as birth or death certificates, or from national surveys, such as, a labour force survey used to determine employment statistics. The "official" status of these statistics is due to their origin from governmental sources with formal mandates to gather and process statistical information. In many instances, these mandates are enshrined in the laws of a country. For example, some democracies have laws that mandate a regular census to determine the allocation of seats within their legislative assemblies. Legislation that requires the production of specific statistics bestows a special, official status on them.

One might expect official statistics to have a commonly accepted definition among the national agencies responsible for their production. However, an examination of such agencies reveals no single, widely held definition. A 1998 green paper in the United Kingdom describes three ways in which this concept has been used. [2]

> First, [official statistics] may be defined in terms of *people* providing the service (e.g., the Government Statistical Service). Second, it may be defined in terms of *activities* (e.g., collecting data, publishing statistics, providing statistical advice to support policy work). Third, it may be defined in terms of outputs, or products of statistical work (e.g., the published statistics on the labour market, on crime, on health etc). [Chapter 4]

Combining these three perspectives, official statistics can be understood as the outcomes of professionals within government agencies engaged in activities to produce published statistics.

[2] *Statistics: A Matter of Trust* (Cm 3882). Presented to Parliament by the Economic Secretary to the Treasury by Command of Her Majesty, February 1998. (http://www.archive.official-documents.co.uk/document/ons/govstat/report.htm)

Other characteristics of official statistics have been identified by Statistics New Zealand.

- [Official statistics] are essential to central government decision-making.
- They are of high public interest.
- They require long term continuity of the data.
- They provide international comparability or meet international statistical obligations.
- They need to meet public expectations of impartiality and statistical quality.[3]

The motivation behind the U.K. green paper arose over a growing public concern about the manipulation of official statistics for political ends. The government of the day needed to reestablish public trust in the veracity of the statistical information about the government's performance. According to the U.K. Library Association, official statistics need to be "a dispassionate statement of the government's performance."[4] The public needs to find official statistics credible. Failing this, such statistics run the risk of being considered fabrications or as Benjamin Disraeli said, "lies, damn lies and statistics."

Official statistics have a role to play in "open government," a concept that involves the public in monitoring and holding governments accountable for their policies and programs. This approach to governance relies on the public being well informed about the social and economic conditions in their country. The provision of trustworthy official statistics is one way of keeping the public informed.

Furthermore, official statistics can significantly influence decisions within the financial sector thereby raising the importance of producing reliable statistics. For example, in 2004 the Canada Border Services Agency reported incorrect trade volumes between the United States and Canada for the month of November. The error resulted in an apparent 10 percent monthly drop in U.S. imports to Canada, which indicated a higher U.S. trade deficit than expected. This news led many currency traders to abandon the U.S. dollar and boost the Canadian dollar by as much as 1.6 cents. An investigation discovered that Canada Border Services had shut down its computer system to install upgrades. Unfortunately, this maintenance occurred on a day that typically registers high import traffic. Following the upgrade, the computer system was not restarted and the trade data for this day failed to be recorded. This error was detected by analysts in Statistics Canada but miscommunications between the two agencies

[3] Statistics New Zealand. "Top Down Review of the Official Statistics System Phase 2 Recommended Option for the future role of Statistics New Zealand and the Official Statistics System." Statistics New Zealand, December 5, 2003. Page 5. http://www.stats.govt.nz/sitecore/content/statisphere/Home/about-official-statistics/~/media/statisphere/Files/top-down-review-of-oss-p2-dec03.ashx

[4] The Library Association. "Response to the green paper Statistics: a matter of trust," May 1998. http://www.rss.org.uk/uploadedfiles/documentlibrary/505.doc.

failed to resolve the discrepancy before the statistic was released to the public. An article in the press stated, “Statistics Canada is working with [Canada Border Services] to ensure trade numbers will be reliable.”[5]

Statistics Canada acknowledges that no standard definition of official statistics exists among national statistical agencies. The agency does note, however, that there are generally accepted quality factors that constitute a “fitness of use” underlying official statistics. National statistical agencies go through formal processes to create and release official statistics. These processes involve steps that address the “relevance, accuracy, timeliness, accessibility, interpretability and coherence of a statistic.”[6] Precise definitions of concepts and sound methodologies for collecting and producing statistics are critical aspects of these processes. An essential feature of sound official statistics is having these processes well documented and readily available for public examination.

Non-official statistics come from sources outside the realm of governments or public organizations and include entities such as professional bodies, trade associations, interest groups, banks, research institutes and commercial publishers. The fact that these sources have been labeled non-official does not mean that these statistics are a lower quality. Rather, these sources are outside the scrutiny of public oversight characteristic of government-produced statistics.

Non-official statistics do not have the same public mandate as official statistics. Nevertheless, many of the same reasons for generating official statistics apply to producing non-official statistics. We live is a world where almost every aspect of life is measured. The methodologies used to produce non-official statistics are similar, if not identical, to those employed in the creation of official statistics. The producers of non-official statistics also engage professionals with skills in the collection of data and the generation of statistics. The private sector in the industrial world invests substantially in statistics making use of the services of many private businesses that specialize in producing non-official statistics.

Knowing the process through which statistics have been created and the definitions of the concepts that have been measured are important to users of statistics. Furthermore, knowing whether a statistic is official or non-official is helpful in tracking down more detail about its production.

You Can Count on Statistics

In broad terms, statistics can be thought of as numeric facts and figures produced by official and non-official sources. The following discussion looks at three general ways in which numeric facts are used in every day life. First

[5] Dean Beeby, “Customs agency fumbles trade figures: Repeated errors hurt StatsCan credibility,” **Edmonton Journal**, April 10, 2005, p. A5.
[6] Statistics Canada. “Statistics Canada's Quality Assurance Framework,” 2002. http://www.statcan.gc.ca/pub/12-586-x/12-586-x2002001-eng.pdf.

Figure 1.1
Examples of Popular Statistical Facts and Figures

"Go Figure," **Sports Illustrated**, Vol. 91 (7), 1999, p. 25.	
$750,000	Approximate value of endorsement deals Brandi Chastain has signed since the World Cup-winning goal.
$5	Amount a Little League assistant coach in Ashland, Ore., gave his players as a reward for base hits in an all-star game.

"Harper's Index," **Harper's**, Vol. 301 (1803), 2000, p. 11.	
Number of months last spring that a Louisiana town's sewage lines were connected to its fresh water supply.	3
Gallons of bourbon that flowed into the Kentucky River last May during a fire at a Wild Turkey warehouse.	200,000

"Snapshots®," **USA TODAY**, October 20, 2005 http://www.usatoday.com/news/snapshot.htm		
The title most owned by libraries worldwide is the U.S. Census.		
Top book titles in libraries:	U.S. Census	403,252
	Bible	271,534
	Mother Goose	66,543

turning to the realm of entertainment, many baseball aficionados take pride in memorizing player statistics. In his rookie season, Hank Aaron, for example, played in 122 games and had a batting average of .280. Both of these numbers help summarize a part of Mr. Aaron's first year in major league baseball, namely, (a) he was a regular in the line-up and (b) he successfully hit the ball a little better than one out of four trips to the plate. These two baseball statistics, while giving an overview of his inaugural season, lose the rich detail of each of Mr. Aaron's first 468 at-bats as a professional. Detailed information has been lost yet a simplified overall picture has been presented.

[Sidebar: 1] *Statistics are frequently used to condense a large amount of information into a few numbers and in this context, provide a concise, descriptive summary.*

Sports statistics are generally popular. Daily newspapers report box scores containing numeric summaries of all kinds of sporting events. A weekly feature in **Sports Illustrated**, called "Go Figure," provides a variety of numeric facts related to sports (for example, see Figure 1.1). The reader occasionally is left wondering if the publisher is suggesting relationships among some of these statistics. For instance, in the August 22, 1999 issue, two facts regarding monetary awards for athletic achievement were reported. One fact dealt with the value of the commercial endorsements received by the woman who scored the winning goal in the World Cup of women's soccer. The other fact was about a Little League all-star game in which each player on one team was given five dollars per hit. An implicit association between these two numeric facts points to the omnipresence of money in sport. Regardless of age and more recently gender, money is offered as an enticement for achievement. The implied equations are that athletic achievement equals money or that money drives athletic achievement.

Outside of sports, Harper's Index™ is a regular column of numeric facts that, while not necessarily logically related, are strung together to make an amusing statement about today's world. For example in the August 2000 issue, two numeric facts were listed that dealt with water quality in separate parts of the U.S. One community in Louisiana had it sewage lines connected to its fresh water supply for three months. In a separate incident, 200,000 gallons of bourbon spilled into the Kentucky River as a result of a warehouse fire. Implicit concerns about what people are drinking or not drinking seem to be associated with these two facts. Depending where you live, bourbon and water may not be a wise drink!

When relationships between phenomena are being explicitly examined – unlike the previous two examples – statistics can be used to show that as one thing changes by a certain amount or percent, another thing increases or decreases correspondingly. In the same issue of Harper's Index cited above, one of the numeric facts reported gonorrhea rates among teens and young adults decreased nine percent when the beer tax is raised by 20 cents. As the price of beer went up, the rate of gonorrhea went down. Here the numbers have been used to indicate a link between the two phenomena. While many things in life seem to be correlated or somehow connected, these associations are not necessarily causal. Consequently, some numeric relationships, while intriguing, have no substantive basis.

[Sidebar: 2] *Another common use of statistics is to summarize relationships or associations among things*.

A commonly heard expression is, "I don't want to become just another statistic." This expression typically arises in relation to highway fatalities or divorce or spells of unemployment. "Becoming a statistic" is also colloquially used in reference to a loss of individuality where being a number is seen as being marginal in the existence of a larger crowd, such as being just one of millions. In these examples, a statistic serves as a measuring stick against which comparisons are made: I don't want to be among the annual highway death count or I want to be seen as different from all of the others. Using a different example, a statistic such as life expectancy at birth indicates a measure of projected longevity. In this case, the statistic serves as a baseline for a baby's expected lifespan.

[Sidebar: 3] *Statistics are frequently used as signposts or yardsticks against which things are compared.*

One technique used when making comparisons between groups of different sizes is to adjust the statistic being used to a norm or equal standard. In an issue of **Sports Illustrated** following the Summer Olympics in 2000, the comparison of the number of medals won by Australia to those by the United States was adjusted to the population sizes of these two countries. Controlling for overall population size, Australia won 3.03 medals per million population compared to only 0.352 medals per million Americans.[7] The number of medals won by these two countries was transformed to a comparable rate based on a million people in the population*. To make fair comparisons between groups of different sizes, the statistic is often normalized or transformed to a standard baseline.*

The Discipline of Statistics

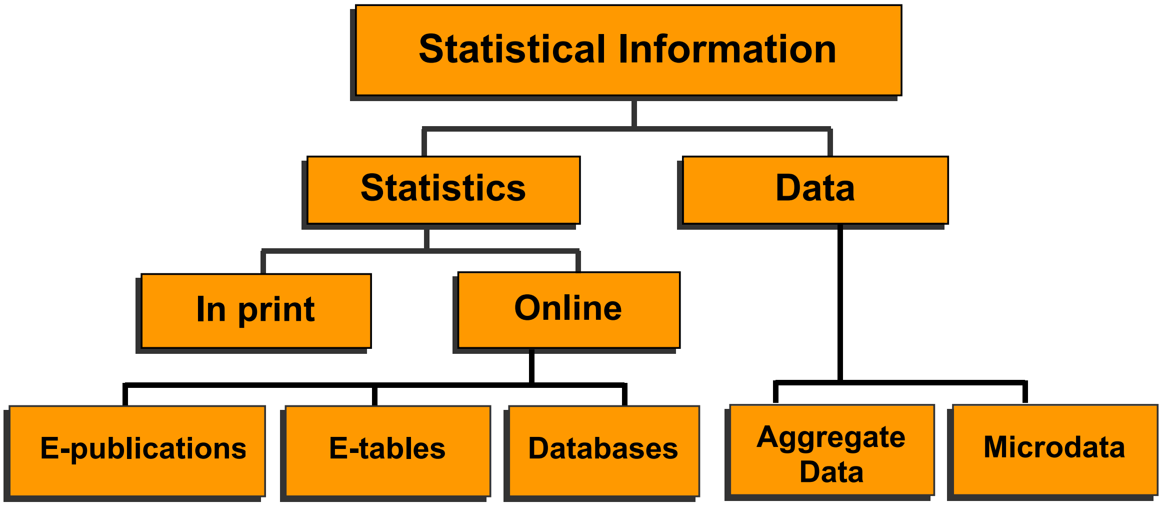
Statistics also have formal roots in an academic field of study with a supporting body of knowledge. Built upon theories of probability and inference, statistical reasoning supports the making of broad generalizations from a smaller number of specific observations. Mathematical models employing stochastic or error components are instrumental in making such generalizations. These theories of statistics are used to investigate all kinds of phenomena, including human and natural events.

There is a more technical meaning of the word, statistics, in the discipline of statistics. Specifically, a sample estimate of a population parameter in a model is called a statistic. In a city, one can speak of the average age of all people living within its boundaries. This average for the city is known as the population parameter for age. A sample of people living in the city may be drawn from which the average age for the population is estimated. This average age determined from a sample is technically known as a statistic. This is important to be able to distinguish between a purely technical use of the term, statistics, and its more popular usages.

Many of the professionals who work in the agencies that produce official and non-official statistics have received training in formal statistics. This type of training may occur in the discipline of mathematical statistics or in any of a number of fields in which applied statistics are taught, such as sociology, psychology, economics, or epidemiology. Out of the theoretical and applied study of statistics comes new knowledge about the methodologies and techniques used to produce statistics.

[7] "Go Figure," **Sports Illustrated**, Vol. 93 (14), 2000, p. 24.

Chart 1
Categories of Statistical Information



The significance of theoretical and applied statistics in our treatment of statistics is that they constitute the underpinnings used in the production of official and non-official statistics.

Categories of Statistical Information

Questions involving statistics tend to be common and plentiful at almost any library reference desk. Many of these inquiries are for simple numeric facts and in anticipation of this, most reference services are stocked with a variety of yearbooks, almanacs and government publications to assist in finding an answer. More recently however, statistics have become available online in electronic publications, tables and databases.

In the past, librarians tended to separate statistics from data based on the medium of the resource. If the statistical information was available in electronic format, the item was handled as data. If, on the other hand, the source was in print, the item was treated as statistics. When the reference desk only carried print sources for quick reference, only statistics were found there. Now that sources for statistics are more likely to be electronic than in print, the medium of the source has lost its usefulness in differentiating statistics from data.

Chart 1 shows a framework for statistical information that encompasses statistics and data, which we find more appropriate than categorizing materials simply on the basis of format. In this framework, statistics are processed information representing someone's view or analysis derived from a data source, model or simulation. Statistics are ready for intake and, as such, are organized in displays and presentation layouts, most commonly in the form of tables and graphs. Data, on the other hand, represent the raw information stored in computer files from which statistics are created. These files have been prepared in a specific data structure suitable for processing. Data in this context are not readable by the human eye in the same way that statistics are but rather are organized for computer use. Data also require additional, critical information – known as metadata – to be understood. Without metadata, data cannot be processed in a meaningful way.

Statistics, Formats and Access

The format in which statistics are disseminated has an impact on access to these materials. As mentioned earlier, statistical yearbooks and abstracts in print are common items on quick or ready reference shelves to assist the librarian. The primary tools for finding statistics in print have been library online public access catalogues (OPACs) and the product lists of data producers.

In addition, online bibliographic databases make it possible to search for articles in which research findings are reported. For example, Medline™ covers over 4,600 journals, many of which publish articles containing statistics from research outcomes. With the advent of full-text databases, this information can be increasingly retrieved online as well. Some vendors have created searchable databases that index tables within publications and articles. Statistical Insight™ (formerly Statistical Universe) and Tablebase™ are examples of full-text databases that permit searching at the individual table level rather than the publication level.

Many print sources containing statistics have been converted to electronic publications, which brings us to the first category below online statistics in Chart 1. These titles may be continuations of serials in print or copies of publications disseminated in both print and electronic format. They are likely to be found in library OPACs, where the record describing the item may contain an online link directly to the e-publication. These resources are also often indexed by online search engines, which expand the discovery possibilities. E-publications of statistical titles are typically distributed in PDF format. Newer releases of the Adobe Reader™, which display the PDF format, contain tools that permit highlighting and copying columns from tables in e-publications. While not the most efficient way of transferring statistics to analysis software, this approach is superior to keying in the statistics by hand, which is the option for statistics in print. The Adobe Reader's search tool also facilitates locating statistics in this kind of document.

E-tables constitute the second category of online statistics. Unlike tables organized in a book-like structure, which is the case with e-publications, e-tables are displayed on Web pages. Discovery of these statistics is largely dependent on Internet search engines, although e-tables are often organized using subject headings on the Web site of their producer. They tend to be static HTML representations of tables, although some offer a drop-down list to regenerate the table to display other dimensions. Certain producers of e-tables offer options for downloading these statistics in formats compatible with other software or in common interchange formats. For example, one interchange format commonly used is the comma separated value or CSV format, which is processed by a variety of analysis systems, including Excel and statistical packages such as SPSS and SAS.

The third category of online statistics consists of statistical databases accessible through the Web. One can usually search for specific statistics within this type of database, although the initial discovery of the database may depend on other information sources. This database approach often employs the use of Web forms to describe the view of the table that a person desires. A query of the database is generated from the choices made on the Web form. In return, the database delivers a table of statistics in HTML or in a variety of other displays or interchange formats. A database of statistics shares similar properties with aggregate data, which is described in more detail below. Suffice it to say, such databases organize their statistics according to time or geography. Separate options for specifying a time period and a geographic reference for statistics are typical of the Web forms used in conjunction with this kind of database.

As previously mentioned, statistics are processed data that have been organized for display in tables or graphs. Consequently, access to statistics has been shaped by the medium on which they have been organized for dissemination. The tools for discovering and locating statistics similarly are dependent on format and moving from print to electronic format has increased the potential for the discovery of and access to statistics. The progression of online formats from e-publications to e-tables to databases moves access from replicas of statistical reports in print, to electronic versions of print tables, to databases of statistics open to numerous possibilities of on-demand displays.

Two paradoxes arise from this improved electronic environment and the resulting improved access to statistics. First, being able to create statistical displays on demand has not been accompanied with a parallel improvement in the metadata to cite these sources. Can someone other than the person reporting a statistic actually find and retrieve it? This is less of an issue with the more static e-publications and e-tables, which can usually be located through a reference to a URL. Statistics from a database, however, are more challenging to cite because of the dynamic ways in which they are delivered. When some databases are updated, existing statistics are changed or deleted. Not only does this create a problem in retrieving previously cited statistics but it also raises the question of preservation for long-term access. This brings us to the second paradox: an apparently inverse relationship between convenience of dissemination and preservation standards. The more convenient it becomes to disseminate statistics through databases on the Internet, the less attention is given to the standard by which this information needs to be organized for the purposes of preservation. Database delivery of statistics typically requires storing them in the proprietary format of a commercial software system that often fails to support a recognized preservation format. The issue of preservation is one to which we will return in subsequent chapters.

Understanding Data

Like statistics, the term *data* has a variety of common-language uses that can obfuscate understanding. For example, the image of a popular Star Trek character is conjured up in the minds of some when they hear *data*. In this context, Data is a fictional entity; the discussion below focuses on factual data.

Typically, *data* refer to vast quantities of information. For example, the press will report today's data on trading activities of various stock markets. The high volume of ticker values from these stock markets flowing across the television screen or printed in the newspaper portrays this meaning of data. Consistent with this are references to data associated with instrument readings where a tremendous amount of information is quickly generated. Examples of this include the steady stream of images flowing from weather satellites or the abundance of seismic recordings captured during oil and gas exploration. Following the Concorde disaster in July 2000, some in the press referred to the black box containing the aircraft's instrument readings leading up to the crash as the “data box.” The usage was in reference to the continuous instrument recordings of critical components on the aircraft. *The large volume of raw information produced by processes, such as buying and selling securities on an exchange or voting in an election, or by instruments, such as EKG monitors or spectrographs, are commonly identified as data. Furthermore, this raw information requires subsequent processing to be of practical analytic value.*

The concept of data also has important technical meanings. In computer science, data are the binary values stored in memory (RAM) and loaded sequentially into registers of the central processing unit (CPU), either to perform an operation or to be manipulated. Everything in a computer is encoded in binary, including the operating system and application software, which are files containing instructions to drive the operations of the CPU as well as the information to be manipulated by the CPU. Computer programs written in a higher-level language, that is, a level other than the CPU's binary instruction set, are converted to binary instructions either through a compiler, which translates code en masse into the CPU's instruction set, or through an interpreter, which does the translation on the fly. Computer scientists regard the output of a compiler, which is typically stored in a file, as binary data. Similarly, the binary output of interpreted code is also called data.

A more general definition of data in computing is anything that is stored in a file. This might be a compiled program or a document, such as the file containing a draft of this chapter. Think of the typical application software comprising a popular office suite in today's computing environment. There are word processing files, database files, presentation files, spreadsheet files, image files, sound files, help files and much more. The contents of all of these different types of files are commonly called data. *The concept of data in information technology has technical uses that are both general and detailed. Generally, data consist of the raw contents of files, which are usually prepared for some type of processing. More specifically, the concept of data is used in computer science to identify the raw binary values being operated or manipulated in the CPU and stored in memory.*

Introducing Social Science Data

The concept of **social science data** derives its meaning both from information technology and social research methodology. Social science data in this context are the digital resources out of which social and economic statistics are produced. The data do not spontaneously spring into existence but are produced from an intentional research methodology.[8] A variety of methods exist to collect data systematically and consistently, which are essential attributes of sound methodologies.

In social surveys, information is usually collected from people about their opinions, behaviors, experiences, attitudes, and personal characteristics. For example, a poll of 2,552 adults was conducted during the Canadian federal election in 2000 following the nationally televised debates in French and English among party leaders.[9] Each person in this sample was asked, “In your opinion, who won this debate?” In this instance, an individual adult in the sample represents one member of the unit of observation, that is, the object about which data are observed and collected. Combining the answer to this question and all other questions in the poll for every respondent provides the raw material that when organized in a specific data structure becomes the data of this survey.

Individuals do not always constitute the unit of observation in social science research. For example, a labour economist studying dispute resolution might focus on strikes or labour disputes as the unit of observation. While strikes involve people, the object studied in this hypothetical case is a labour disruption in the workplace. The information about each strike might include the number of workers involved, the duration of the work action, the issues of the dispute, the industry in which the dispute occurred, whether the courts intervened during the strike, etc. Each element of the unit of observation in this study would be a specific work disruption.

[8] Administrative record management, while not commonly viewed as a research methodology, usually consists of systematic and consistent information collection practices. As a result, most administrative records can be shaped into a social science research design after the fact.

[9] This example is from the Globe/CTV/Ipsos-Reid poll reported in *The Globe and Mail*, November 13, 2000, p. A8.

Another example where people are not the unit of observation is seen annually in the fall issue of Maclean's magazine in which Canadian universities are ranked according to a survey of post-secondary institutions. This survey collects information from each university about the grades of incoming undergraduates, research grants revenue, money spent on scholarships and bursaries, size of the library's collection, among other factors. All of this information is combined to rank universities. In the Maclean's survey, institutions are the unit of observation.

The way in which information in the Maclean's survey and in all other surveys is organized for statistical processing is fundamental to social science data. The structure of social science data is built upon the concept of a unit of observation, which one may think of as the backbone of this data structure. Everything else is built on this backbone. *A defining characteristic of social science data is its structure, which is determined by its unit of observation. Data are the raw information collected about each individual member of the unit of observation organized in a specific structure, while statistics summarize properties or relationships about the unit of observation.*[10]

Social science data are stored in computer files using a physical format dependent upon the statistical software being used and, as previously mentioned, a logical structure determined for the unit of observation. This distinction helps differentiate social science data files from other files said to contain data. While a social science data file may be opened in a word processor, the organization of the information in the file should leave little doubt that the contents are data to be processed by statistical software. Some ambiguity may exist whether the information has been prepared for a spreadsheet or database package. Structural differences do exist, however, between spreadsheet data and data for statistical software. *The physical organization of social science data in computer files is dependent upon a logical structure based on the unit of observation. Furthermore, the contents of a file organized in this manner should usually be recognizable when displayed.*

Social science data must be processed to be of practical use. Statistical software accomplishes this by reading the data from the file in which it has been stored and then analyzing it through a variety of different statistical procedures. The logical structure of a social science data file has specific properties, including the number of cases, that is, the number of individual members of the unit of observation, and the number of variables, which are the attributes observed about each case. A detailed description of the number and type of variables

[10] The general rule is that data are collected and stored at the level of the unit of observation. Summaries of these become statistics. Some research designs consist of multiple levels at which a unit may be observed or layers in which one or more units of analysis exist. In these instances, summarizing the data from a lower level to a higher level will result in new data. Whether these new data are seen as statistics or data will depend upon the intended use of the summarization. Some may use these as data for further analysis. Others may treat the summaries as statistics.

must be communicated to the statistical software using the command language of the package.

The physical computer files in which these data are stored also have properties, such as the length of the longest and shortest lines in the file and the number of lines in the file. These physical file properties have a direct relationship to the content of the data and must be fully explained in accompanying data documentation for the data to be understood.

Data, Formats, and Access

Data are grouped into two categories (see Chart 1): aggregate data and microdata.

Aggregate data are composed of statistics organized in a social science data structure. These statistics are often stored in a database, which sometimes is the same database providing access to online statistics. They are closely related, distinguished by the structure in which the statistics have been retrieved from the database. As mentioned above, the unit of observation is the backbone underlying the organization of social science data. If the statistics are retrieved and organized using a specific unit of observation, collectively they constitute aggregate data.

Aggregate data are organized using one or a combination of three units of observation. A unit of time is one of these structuring factors. In this case, statistics are arranged along a timeline and are commonly referred to as a time-series. This type of aggregate data is particularly useful in identifying trends or changes over time and models representing the performance of the economy are often built using time-series data. Because a high volume of economic and financial statistics is organized this way, the business sector is a primary producer and user of these data. Consequently, access to time-series databases often entails purchasing these data from commercial vendors.

Spatial or geographic units make up another observational factor around which aggregate data are organized. With the advent of Geographic Information Systems (GIS), the demand for statistics organized according to geographic units grew tremendously. Typical spatial units include the variety of Census geographies that are used to capture and disseminate Census statistics. Geographic areas associated with the delivery of services are also popular. These include ZIP or Postal Codes, health regions, school districts, and a wide variety of other public service boundaries for police, fire and transit.

“Small area statistics” is a special category of spatial aggregate data. These data files consist of statistics for small geographic areas, such as neighbourhoods. The creation of this special class of aggregate data is governed by an inverse relationship: the smaller the geographic area for which statistics are desired, the larger the overall data source required to derive them. Smaller areas require larger samples to ensure enough cases exist to produce accurate estimates for each geographic area. Therefore, these data are usually calculated from a population or manufacturing census or an administrative database with enough cases to create accurate small-area summaries.

The third factor structuring aggregate data is social content. Also known as “cross-classified” tables, these files are composed of statistics constructed around the categories of social-content variables. Examples include the cause of death detailed in codes of the *International Classification of Diseases*, Tenth Revision (ICD10), the *National Incident-Based Reporting System* (NIBRS) crime categories used in the *Uniform Crime Reports*, and the *Carnegie Classification of Institutions of Higher Education*™ framework for recognizing and describing institutional diversity in U.S. higher education. Cross-classified tables are typically found in health, education and justice where the origin of much of these data is from administrative databases.

Access to aggregate data is typically through database retrieval, although some e-tables can be reshaped to display statistics where the rows represent time, geography or categories of specific social content. Many producers offer online Web retrieval of their aggregate data.

Microdata, the second data category in Chart 1, contain information collected directly from a specific unit of observation. Previous examples made reference to voting-aged citizens, labour disruptions, or post-secondary institutions as units of observation. The information collected about each member of the observed unit consists of characteristics or attributes of these entities. Given the richness of detail held in microdata, they have extensive research value well beyond their initial purpose. Consequently, microdata files are important objects around which mediated services must be provided. It is the nature of data services associated with microdata that is a primary focus of this book.

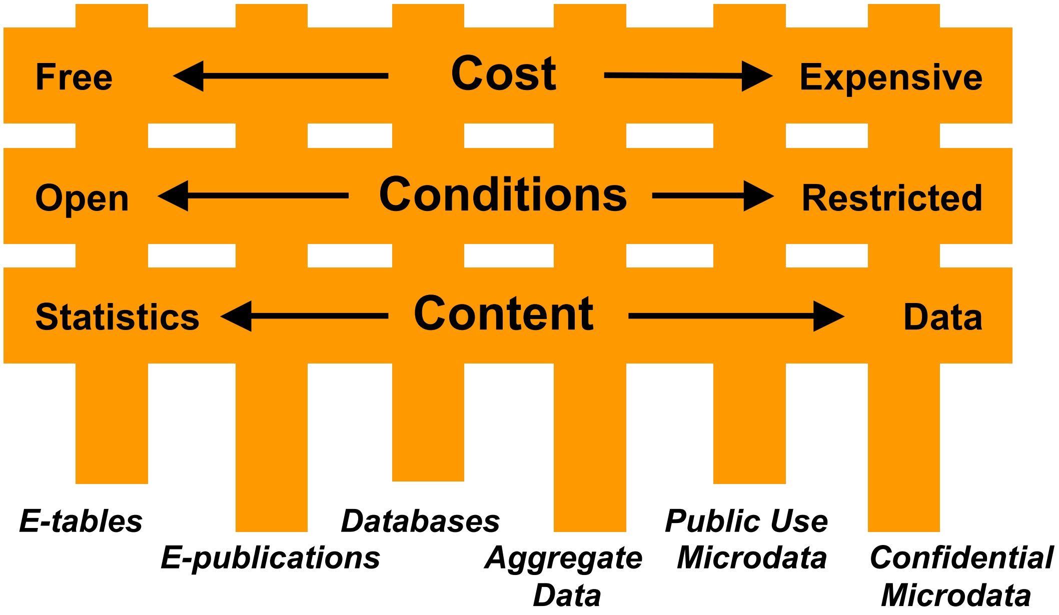
The level of detail contained in microdata files raises concern over the privacy of the individuals whose characteristics constitute the data. As a result, microdata are prepared as either confidential or public-use files. The contents of the confidential files contain enough information to result in the discovery of the identity of members in the unit of observation. National statistical agencies go to great lengths to protect the identity of those from whom they collect information and are often governed by laws stipulating the conditions under which access to these data is allowed. Similarly, academic research ethics boards demand specific practices be followed to protect human subjects whose data are confidential.

With growing concerns over identity theft, practices permitting access to confidential data have undergone increased scrutiny by governments and the public. This new reality has complicated access to these data for legitimate research uses. New access protocols are being developed by some national statistical agencies to provide strictly controlled research access to confidential data. For example, both the Bureau of the Census in the United States and Statistics Canada support data enclaves in which selected researchers are allowed to analyze confidential microdata in a tightly controlled computing facility. All research output must pass a disclosure analysis by an employee of the agency before it can be removed from the facility.

Another approach in providing access to microdata while protecting confidentiality involves creating public-use files. These products are generated from confidential files by undergoing transformations to anonymise the data. This method employs practices that guard against disclosure by making the data "safe". Producers use a variety of strategies to minimize the likelihood of disclosure. All personal identification information is removed, such as phone numbers and names. Only gross levels of geography are included, for example, state or province. Variables with a large number of detailed categories are collapsed into a few general categories. For example, occupations may be captured using hundreds of 7-digit classification codes but for the public-use file, occupation may be collapsed to a dozen general categories. The upper values of variables that represent rare occurrences may be capped. For example, the number of people living in a household may be capped at 4 or more. Sometimes a variable is deemed to be too sensitive to be released and the entire variable is suppressed in the public-use file. In other instances a case will be judged to be too unique and will therefore be suppressed. Most public-use microdata files undergo a strict review process by the data producer before the data are released for dissemination.

Access to microdata is contingent on techniques that safeguard against disclosure by not creating too great a barrier to work with the data or by not diluting the data beyond meaningful research value. The paradox facing producers of microdata is whether to make the data safe by reducing its content or to make the research outcomes of confidential data safe through some form of disclosure analysis. The former broadens public access to the microdata but reduces the scope of research that can be performed. The latter expands the range of possible research but narrows the field of researchers who have access to the microdata as a result of the overhead entailed with disclosure analysis. The decisions of data producers on how to address the issue of microdata confidentiality have a major impact on both those who would use the data and those providing data services. The importance of this issue is revisited throughout this book.

Chart 2
Continuum of Access



Continuum of Access

For those providing data services, creating access to statistical information is a primary mission. While the classification of statistical information shown in Chart 1 is a useful way of thinking about the variety of materials that exist, understanding access to these resources requires another tool. Three factors have a direct impact on access to statistics and data. Together they form a continuum along which producers disseminate statistical information.

Cost is always a limiting factor in providing access to resources. Is the information free, inexpensive, expensive or prohibitively expensive? The conditions under which access is granted constitute another factor. Is the use of the statistical information open without restrictions? Or are there conditions that tightly govern the use of the materials? A third factor is the amount of processing that is required to work with the information. Statistics are processed data and present a predetermined view of the data. To be of analytic use, data, on the other hand, require processing. This third factor identifies whether access is desired for statistics or data.

These three factors constitute the continuum of access. On one end, access consists of free statistics that are openly available without any restrictions. The other end of the continuum represents very costly data that are highly restricted and therefore have very limited access.

Applying this model in conjunction with the statistical information framework, different channels appear through which statistical information is disseminated. When it comes to official statistics, the Open Data movement in many countries has led to the dissemination of free, open statistics. For example, beginning in February 2012, Statistics Canada opened access to the millions of time-series aggregate data in its CANSIM database on the Internet without fee. Prior to this, Statistics Canada charged $3.00 per time-series.

Moving along the continuum away from statistics and toward data, public-use microdata files appear. They may be free or involve charges for access and almost always carry some conditions of use. For example, the Inter-university Consortium for Political and Social Research (ICPSR), which houses a large data archive, distributes some microdata files only to individuals at member-paying institutions. Other files have been deposited with the ICPSR through government-funded sponsors and are available to the public without charge. Progressing from public-use to confidential files, access becomes tightly restricted and the cost of the service to support this access tends to be quite high.

The typology of statistical information and the continuum of access together provide a helpful way of thinking about how to identify and locate this information. The categories of statistical information help in finding an appropriate product, while the continuum of access points toward the channel or channels through which the statistical information is disseminated.

Understanding how data and statistics are related and yet different is essential before exploring the variety of approaches in providing social science data services. The rest of this book focuses on data and data services, with a particular emphasis on microdata.