DATA BASICS

an introductory text

by

Diane Geraci
Chuck Humphrey
Jim Jacobs

August, 2012

DATA BASICS
an introductory text

Table of Contents

Appendices

Preface

This monograph emerged out of more than ten years of teaching a one-week workshop at the Summer Program of the Inter-university Consortium for Political and Social Research (ICPSR) housed at the University of Michigan. The ICPSR is a major international social science data archive and has provided access to research data since the 1960's. When the three of us were invited in 1990 to teach the introductory workshop on archival data services, we shifted the focus from archiving data, which the ICPSR does very well, to providing local data services on individual campuses, which at the time generally was not being taught anywhere[1]. Instead of training new data archivists, our goal was to prepare librarians, academic staff and computing specialists in the fundamentals of offering local social science data services. Many of our students were their university's liaison with the ICPSR, known as the ICPSR Official Representative, and were responsible for ordering and receiving data from the ICPSR. They attended the workshop seeking insights into how they might better organize data services for their local patrons. Their interests were on the demand side of the data-economy equation. How were they to support the secondary data analysis needs of researchers, students and scholars on their local campuses? What could they do to promote wider use of their universities' membership in the ICPSR?

Each year we have learned from our students which material they have found to be the most useful and we have used this information to modify, and hopefully improve, the next workshop. By listening to the comments and suggestions of our students, we have endeavored to stay close to the issues they see as most relevant. Recently, students have been puzzling over the difference between statistics and data. Many assumed that the two concepts are synonyms and therefore interchangeable. However with the growth of online digital sources for statistics, data and statistics suddenly did not appear so similar. Chapter 1 discusses how these two concepts, while related, are different and how these differences are important to those providing social science data services.

One common administrative approach in organizing data services has been to treat data as a special format library. From the administrative perspective, data services simply did not fit with traditional print materials and furthermore required special skills to support. Consequently, data services were treated in some libraries as a special format. Unfortunately, many special formats end up ghettoized in the library -- that is, pushed away from the mainstream and, as a consequence, underutilized. Indicators of a "format ghetto" include separate catalogues outside the main library catalogue, separate reference services, physical locations tucked away from other services, failing to be integrated with the liaison network between academic staff and subject librarians, isolated

[1] The major exception was a one-day workshop offered at the annual conference of the International Association for Social Science Information Service and Technology (IASSIST).

budgets, etc. Administratively, this approach appears the easiest way to deal with something outside the mainstream; however, the result is often the marginalization of a service. One of the challenges for data services consequently is to mainstream data while still providing the special service support required in data. Chapter 2 discusses the challenges of mainstreaming data within the context of the history of providing data services.

Data libraries build collections and provide services in a larger context of the production and distribution of data. In recent years changes in technology, scholarly communication, and the larger marketplace have affected both production and distribution of data. In chapter 3 we discuss the *data economy* and how it affects data libraries.

The next group of chapters (4-6) addresses issues arising in the search for data. Chapter 4 presents strategies for finding social science data. Clearly, transferable search skills exist between bibliographic tools in the print world and tools for data. However, data do present special bibliographic issues. Data are physically organized in files that belong to a study. This creates a hierarchy of descriptive information to catalogue. Elements that describe a study, such as title, the names of principal investigators, names of funding agencies, the number of files associated with the study, and other descriptors, present one level of the hierarchy. The other level of description covers the content within files. Commonly referred to as a data dictionary, the detail within a file includes variables, the coding scheme used, labeling, the treatment of missing values, etc. Whether to search at the study or variable level becomes part of the strategy when searching for data.

Two other aspects of finding data include the vocabulary of data and reading data documentation. The use of language in the social sciences can create ambiguous concepts that result in difficulties in locating relevant material. This is equally true of social science information in print. The added complexity with social science data is the infusion of concepts from social science research methodology. For example, knowledge about the unit of analysis is critical in locating an appropriate data source. An understanding of social units, the measure of time within a study, and the meaning of space also shape a search. The vocabulary of data is reviewed in Chapter 5 and a glossary is included in Appendix C.

Data documentation contains a great deal of technical information about a study and the contents of files. Although a new standard for data documentation (the Data Documentation Initiative or "DDI") now exists, much documentation is still in non-standard formats. Consequently, the quality of data documentation varies greatly. Assessing the strengths and weaknesses of data documentation is an important skill in providing data services. One needs to develop interpretative skills in working with data documentation and Chapter 6 reviews this issue.

In chapters 7 and 8 we examine the mission and goals of data libraries. We believe that most data libraries share a common mission and common goals. These chapters outline the issues that these raise and set the context within which a data library can define its own objectives, choose the services it will offer, and create its collection policy.

While data libraries have a broad mission and goals in common, all data libraries are not the same. Each data library will have to set its own objectives in the context, not just of the broader mission and goals, but also in its own particular context of users, institutional goals and infrastructure, and resources. These will lead the data library to its own individual choice of what level of service it can provide and the extent and nature of its collections. We set the scene for these choices in chapter 9 by introducing the basics of data preservation. The OAIS standard puts digital preservation squarely in the center of the lifecycle of information and an understanding of the role of preservation is essential for those planning services for data.

Setting objectives begins with the development of a service plan, which we discuss in chapter 10. The process of creating the service plan allows each data library to discover and express its unique needs and services. In chapters 11-13 we examine this concept of "levels of service" which allows each library to identify the mix of services and collections it will and will not provide by matching its resources to the needs of its users in the most effective way.

One of the realities of data service is that special skills are needed to support digital materials. The level of skills needed is directly dependent upon the level of service that is desired. For many smaller, teaching institutions, the number of staff is few and the demands on them are great. Data service in this environment may consist primarily of an identification and acquisition function. Even at this service level, special skills are needed. In a large research institution, extensive reference services may be required and these demand additional data skills. One of the key factors involved in choosing an appropriate level of service is the mix of skills needed by staff. Chapters 11-13 specify the kinds of skills needed for each level of service for reference, collection building, and computing.

In chapter 11 we look at collection choices and some specific objectives open to data libraries. While collection choices were once fairly linear, with one decision leading logically to the next, the variety of options open to the data librarian and the variety of restrictions and limitations imposed on data access now create a matrix of choices. The data library need not have a single, inflexible collection policy, but can use different strategies to deal with different conditions. We also enumerate specific level-of-service options for data collections.

Chapter 12 deals with reference services for data -- a topic that ranges over everything from "ready reference" and more traditional kinds of library reference service to more specialized data, computing, and statistical consulting. We define what these are and give examples of their implementation.

Delivery of data has to involve computing, but the choices of who provides computing resources, what kind of services are made available, and the extent of services is something that each individual data library can define for itself. Chapter 13 examines the kinds of choices, the skills needed, and the possible levels of service open to the data library. One of the reasons that a “statistical” inquiry may be inappropriate for data services, especially when the answer can be obtained from a print source, is that the nature of social science data requires some form of computer processing to transform the data into “statistics” or useful research information. Whether the desired information is descriptive, comparative or used in modeling, the patron will need to process the data using some analytic software, most likely a major statistical system such as SPSS or SAS or Stata. The need to process data can influence the level of reference service provided as well as the methods for providing local access to files. For example, a data service may provide extraction services for patrons where subsets of variables and cases are prepared by reference staff. Alternatively, web-based services may be used to provide online extraction services permitting patrons to perform data extractions on their own.

Finally, chapters 14 through 19 present "strategies" -- very specific techniques used to actually accomplish the objectives defined by levels of service choices. This includes strategies for collection development for data (Chapter 14), strategies related to acquisitions of data (chapter 15), reference strategies (chapter 16), strategies that enable users to access data collections (chapter 17). Also covered are strategies for developing and maintaining a level of computing services appropriate to support the intended level of data service (Chapter 18), and strategies for promoting data among the variety of patrons on a local campus (Chapter 19). The dependence of data services on technology ensures the need to prepare a plan for the future.

Data services are influenced by technology, which changes rapidly, and by the methods of production and distribution of data, which continue to evolve in new ways. This creates a need for continual professional development. Chapter 20 provides a list of organizations and e-mail lists that focus on data or statistics or related issues.

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

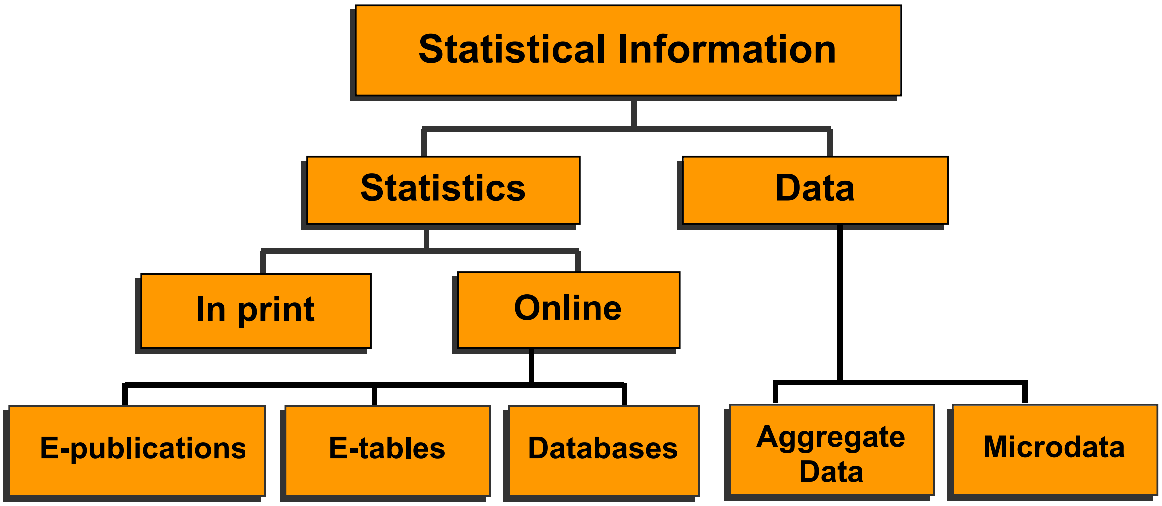
Statistics also have formal roots in an academic field of study with a supporting body of knowledge. Built upon theories of probability and inference, statistical reasoning supports the making of broad generalizations from a smaller number of specific observations. Mathematical models employing stochastic 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. There is a growing literature that addresses preservation of databases, but this field is still in its infancy. 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

[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.

strike, etc. Each element of the unit of observation in this study would be a specific work disruption.

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

[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.

observation, and the number of variables, which are the attributes observed about each case. A detailed description of the number and type of variables 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.

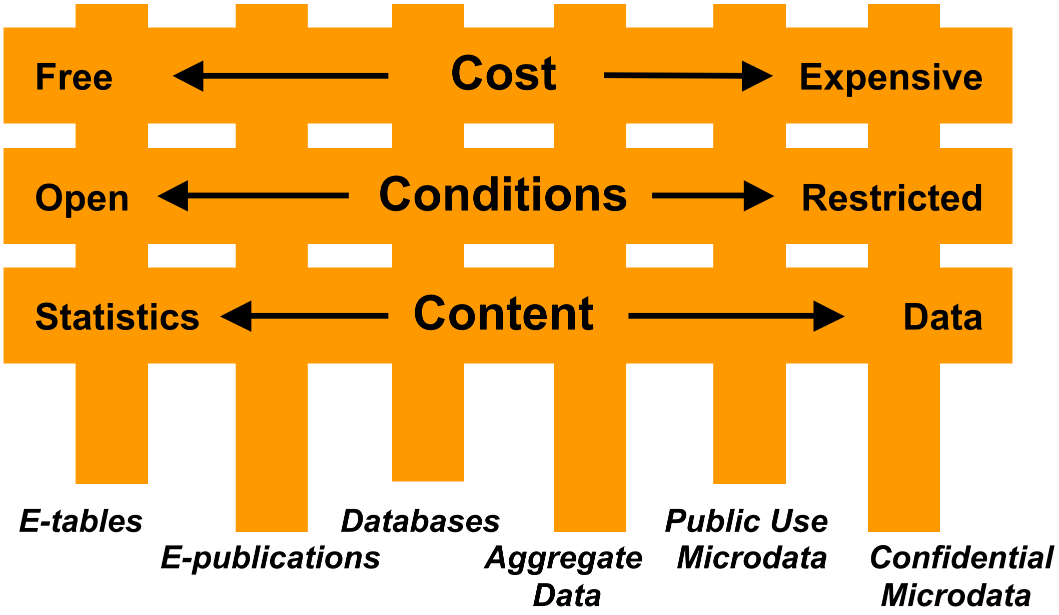
With growing concerns over identity theft, practices permitting access to confidential data have undergone increased scrutiny by governments and the public. This new reality has complicated access to these data for legitimate research uses. New access protocols are being developed by some national statistical 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.

Data Basics

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.

Introducing Data Services

This chapter provides a brief sketch about how data services have developed over the past two decades and comments on some of the directions that data services seem to be headed.

The Variety of Data Services for Social Science Research

Data services supporting quantitative social research have existed in a variety of organizational settings. Libraries, archives, social science research institutes and computer centers have all been homes for these services. Where a data service ends up located depends on a combination of factors, including the fit of the services with the host organization's mission, the research and instructional needs of the host’s clientele and the availability of budget and staff resources. Historical circumstances also play an important role. For example, a prominent researcher may have been influential in starting a data service in an academic department. Initially buoyed by this person’s presence, the data service may continue in its founding location even long after the researcher has left the university simply because the service started there.

This rich variety of organizational settings is exemplified by some examples of long-standing data services in North America. These include Cornell’s Institute for Social and Economic Research; the Odum Institute for Research in Social Science at the University of North Carolina-Chapel Hill; the Center for Electronic Records at the National Archives and Records Administration of the United States; the Data Library and Program Library Service at the University of Wisconsin-Madison; the Rand Corporation Data Library; the Data Library at the University of British Columbia; the Social Science Data Archives at UCLA’s Institute for Social Science Research; UC Data at Berkeley’s Survey Research Center and the Inter-university Consortium for Political and Social Research hosted by the Institution of Social Research at the University of Michigan. These are all well-established services that have been in operation for twenty-five years or much longer. The fact that many of these data services are affiliated with a social science research institute is not a coincidence since these institutions were typical homes of data services in the early years.

In Western Europe, data services for quantitative social science were housed mostly in central, national institutions, such as national data archives, that provided services to individual researchers in their country. Examples include the UK Data Archive (UKDA) located at the University of Essex; the Norwegian Social Science Data Services at the University of Bergen; the Zentralarchiv at the University of Köln; the Data Archiving and Networked Services in the Netherlands, which houses the Steinmetz Archive; and the Danish Data Archives. While the UKDA is the primary and largest data archive in the United Kingdom, regional university data services have existed for many years in Edinburgh and Manchester supporting specialized data interests. More recently, the Economic and Social Data Service (ESDS) was established in 2003 as a distributed service based on collaboration among four institutions: the UKDA, the Institute for Social and Economic Research at the University of Essex, the Manchester Information and Associated Services (MIMAS) and the Cathy Marsh Centre for Census and Survey Research (CCSR). The new model of coordinating distributed services across multiple institutions has emerged as an intentional strategy of the Economic and Social Research Council and the Joint Information Systems Committee in the UK. While the London School of Economics, Oxford University and the University of Edinburgh have long-standing, reputable data services, the institutional arrangements in the UK have traditionally been national services.

Elsewhere in the world, support for numeric social science data has similar organizational affiliations as those in the United States and Western Europe. During the 1990s, several data services emerged at national institutions. For example, the South African Data Archive was established in 1993. Also, countries in Eastern Europe and in Asia introduced national data archives and made their presence known internationally among data service organizations. Finish Data Archive was started in 1999. In other places, data services underwent transformation. For example, the Social Science Data Archive (SSDA) at Australian National University started in 1981 in the Research School of Social Sciences. More recently, the SSDA was incorporated as part of a centre supported by a major endowment and renamed the Australian Social Science Data Archive (ASSDA).

In North America, the 1990s were also a time of rapid change in data services. Specifically, university libraries created new data services at a rate that was unprecedented. The emergence of these new operations seemed to come in two waves. The first wave occurred shortly after the formation of the Census Data Consortium in Canada in 1989 and similarly after the release of the 1990 Census in the United States. Libraries were faced with a deluge of machine-readable data files and statistical series, many of which were distributed on CD-ROMs. Government documents and social science librarians were called upon to provide some level of data support. In some cases, the libraries formed partnerships with other units on their campus to offer joint services. In other instances, they boldly took over data service operations that had formerly been provided through computer centers or research laboratories.

The second wave of activity occurred in the mid-1990s. A new subscription service with Statistics Canada providing access to its standard data products was introduced to Canadian universities in 1996. Known as the Data Liberation Initiative (DLI), this service provided the academic library community with an unparalleled volume of Canadian data files. To support access to these products, the number of Canadian university libraries offering data services grew from around a dozen before 1996 to over sixty-six over a four-year span.

Another contributing factor to this second wave was the digital library movement that took academic libraries by storm worldwide in the late 1990s. As library directors crafted their digital library strategies, some realized that numeric data are a natural part of digital information and consequently, saw data services as belonging under the library umbrella. The number of job advertisements in the late 1990s that included a data services component attests to this trend. Furthermore, increased access to government data through the World Wide Web made some level of data services mandatory in libraries just as the infusion of data-loaded CD-ROMs earlier in the 1990s brought pressure on libraries.

In many cases, data services augmented an existing library service while in other cases data services took the lead in shaping a new service. To enable delivery of services, a unit with primary responsibility for data frequently collaborated with complementary services offered by other units. Many libraries made arrangements with their campus' computing center, a research institute, or an academic department to enact their service plan. Partnership models are often an important factor in initiating new data services and in securing its growth over the first few years. Our experience suggests that partnerships, while beneficial during the start-up period, tend not to last over the long run.

The range of services that can be provided in support of social research is extensive. A quality service incorporates aspects from information management, computing, a service philosophy, and knowledge of social science research methodology, and statistical analysis. Other chapters in this book address these in greater detail. The point made here is that data services are composed of a variety of elements that organizationally may be shared among units across a campus or contained within a single unit. The organizational mix of any one data service will reflect the service environment on its campus and the history behind its inception.

At the beginning of the twenty-first century, the direction taken by many universities in North America has been to provide some form of data services. There are still institutions with long-standing data services outside their library but the overall trend has been to incorporate some aspect of data services within the library's support of the social sciences. The next major movement will be mainstreaming data services in the library. This too will likely take on as many different forms as there are libraries.

Mainstreaming Data: Challenges to Libraries

With the large number of academic libraries involved in some aspect of data services, the organizational challenge became one of integrating data into the library's service offerings. Administratively, mainstreaming social science data in a library presents special challenges. Data are clearly format dependent and require specialized skills to support. However, data are not that unlike other content formats in the library. The library has always supported content of value to research and teaching in a variety of formats. Therefore, the library should treat social science data as yet another format containing content essential to research and teaching. Data services belong in the library and fit well with the traditional role of the library as a valuable partner in research and teaching.

The presence of data services in the library also reinforces and extends the role of the library as laboratory – a place to conduct research. In many ways, a data library is very much like a lab in the physical or medical sciences. Instead of Bunsen burners or centrifuges, this facility has computing resources and valuable, high-quality data. The library is not only a place to obtain copies of the wisdom of the ages but also a place that supports the creation of new knowledge through its services, resources, and now through the addition of data services.

Data services, as already noted, fit well with digital library initiatives. Many of the early digital library projects supported the creation of digital images, audio recordings, and e-text content for the humanities. Just as these digital products are viewed as appropriate additions to library collections in the humanities, so numeric data products should be viewed as appropriate additions to library collections in the social sciences.

Once a library accepts data as a content format for inclusion in its collections, it is also important to avoid the curse of creating a special format library for data. Often library managements have found it easy to isolate special format collections. Typical indicators of such isolation include separate catalogues of holdings, separate reference services, physical isolation from the main collection, a budget that is treated independently, and staff that are not involved with other services in the library. The lean times of the 1990s taught us that survival was more likely to happen by not sticking out as an exception or oddity in the wider system.

There are activities that can be mainstreamed in support of data including aspects of acquisition, cataloguing, and reference. For instance, when reference and collection staff understand the nature of social research and have an awareness of the breadth of resources available, they can incorporate data with other materials as part of a complete service. This also helps facilitate access to data files through informed referrals and acquisition. There are other aspects of data services that are difficult to mainstream. For example, working with statistical software to subset data files may not be a reasonable expectation for all reference staff. These unique aspects should become clearer in later chapters.

An additional benefit of mainstreaming data in the library is building stronger links between the library and the wider data community by developing working relationships with key stakeholders, such as data archive staff, government data producers, faculty and other researchers, computing services staff, and statistical consultants.

Data Basics

Even if all aspects of data services are not mainstreamed in the library, they can be mainstreamed on campus. Such campus mainstreaming would include coordinating services among participating units and explicitly defining campus wide responsibilities for data. For those who provide data services from the vantage point of a research institute, computer center, or other non-library facilities, meeting with the reference staff at your campus' library can enrich both units' services. Minimally, this contact will ensure proper referrals between service points and bring to light overlapping services and duplicate acquisitions. More fundamentally, cooperative services may provide a broader range of services and a richer collection of resources for your clientele.

2.6 - Introducing Data Services

Data Basics

Introducing The Data Marketplace

This Chapter provides background about three important aspects of the environment in which social science data services operate. First, like many raw resources, data are a valuable commodity exchanged in their own market. Even though access to data is not usually discussed in these terms, a data economy determines how and who will have access. The principal stakeholders in this special marketplace will be identified and described below. Secondly, all economies are a reflection of some underlying, basic values. In the data economy, data sharing has been a fundamental value that has operated within a tradition of open data exchanges. Proprietary claims to data, however, threaten the value of data sharing and advance commercial interests in data. How will the inherent conflict between these values play out in the future? Third, the tremendous volume of secondary data sources and the wider acceptance of secondary data analysis are contributing to a change in the norms about how social science is conducted. The role of data is being elevated in the overall logic of how research is done.

The Data Economy: From Commonwealth to Commodification

The current marketplace for data is being shaped by a rapidly changing data economy which, from the perspective of production and distribution, determines who gets what data and how. Prior to the mid-1980s, social science data were distributed largely under a commonwealth economy. Social researchers, united by a common interest, participated for the most part in the exchange of major data collections through institutions that charged either for the marginal cost of redistributing the data or for a membership in a cooperative organization.[1] Fundamental to this economy was a principle of openness expressed through the practice of data sharing. Science, the argument stresses, is dependent on openness. Consequently, the practice of open access to data translates into doing good science.[2]

Two trends are reshaping the data economy. First, governments in the 1980s and 1990s increasingly introduced policies of charging fees for data far beyond the marginal cost of making a copy. Second, growth of an e-economy on the Internet has resulted in speculation in the commercial value of data. In this new data economy, commercialization of data is playing a much larger role than in the past. These trends will have a significant impact on the access that researchers will have to data. Moreover, this movement has the potential of changing how social science research will be conducted in the future. When data become too costly, which, for example, was the case with Statistics Canada data beginning in

[1] Clubb, Austin, Geda and Traugott in Fienberg, Martin and Straf (1985), *Sharing Research Data*, discuss the informal exchange of data among some social science researchers and the more formal intermediary institutions developed to facilitate data sharing (pp. 63-66). See also Sieber, Joan [Ed.], *Sharing Social Science Data: Advantages and Challenges* (1991).

[2] A discussion about the communal ownership of scientific results is found in R.A. Merton, *The Sociology of Science: Theoretical and Empirical Investigation* (1973).

the mid-1980s, only the wealthiest of researchers will have access or whole areas of research may be abandoned.[3]

The next few years will involve contention between the more traditional data commonwealth and the newer forces of data commercialization. The complexity of the data marketplace reflects the competing values of the different stakeholders. Every economy is supported by some underlying values.[4] For example, the ‘free market’ concept is instrumental to capitalism while a key principle of communism is ‘state ownership’. One of the fundamental values of the commonwealth data economy is data sharing. The roots of data sharing can be traced to concepts about openness in science and data stewardship. Science requires verifying findings through replication, which is most likely to occur under open access to scientific information. In the social sciences, access to data is also an important component of replication. The notion of data stewardship implies that the researcher is a caretaker of the data and not the owner of the data.[5] As caretakers, researchers are expected to share data with others.

Not all research data, however, are openly shared.[6] Some researchers selectively exchange data with other researchers creating an underground market.[7] This treats data as a commodity for bartering in a usually closed, informal network. There are also data that simply never appear in the data

[3] The experience in Canada during the 1980s and early 1990s was that many Canadian scholars chose to conduct their research based on data from the United States instead of Canada because access to U.S. data was more affordable and consequently more open.

[4] Denise Love, Luis Paita, and William Custer, in “Data Sharing and Dissemination Strategies for Fostering Competition in Health Care,” *Health Services Research*, Vol 36(1), 2001, describe comparative shopping as the value most critical in the health care data marketplace. They say, “Perhaps the cornerstone of a competitive market is a level of information that allows purchasers to compare the price and quality of services across providers. This link between information and competition in health care has dramatically grown in importance as indicated by the proliferation of data sharing and dissemination initiatives in the health care market.” (p. 278)

[5] C.A. Estabrooks and D.M. Romyn “Data sharing in nursing research: Advantages and challenges” *Canadian Journal of Nursing Research*, 27(1), 1995, pp. 77-88.

[6] Conflict over the ethic of data sharing is found in disciplines other than the social sciences, also. This debate is occurring in the life sciences with examples of competing positions on this issue found in “Secretiveness Found Widespread in Life Sciences”, Science Magazine, Vol. 276(5312), April 25, 1997, pp. 523-525, and “The (Political) Science of Salt”, Science Magazine, Vol. 281(5379), August, 14 1998, pp. 898-907. Epidemiologists at the annual meeting of the American College of Epidemiology addressed this in September 2000 (see “Epidemiologists Wary of Opening Up Their Data”, Science Magazine, Vol. 290(5489), October 6, 2000, pp. 28-29). Psychologists working with magnetic resonance images of the brain objected to a requirement by the editor of the Journal of Cognitive Neuroscience to submit their raw data for inclusion in a public database (see “A Ruckus Over Releasing Images of the Human Brain,” Science Magazine, Vol. 289 (5484), September 1, 2000, pp. 1485-1459). This conflict even erupted on the floor of the United States Congress with the introduction of the Shelby amendment, which required that data be available publicly under the Freedom of Information Act (see Science Magazine, Vol. 285(5427), July 23, 1999, pp. 535-536). Daniel Reidpath and Pascale Allotey examined the willingness of the authors of 29 articles in the British Medical Journal to share the data upon which these articles were based. They discovered a general reluctance by researchers to share their data. See “Data Sharing in Medial Research: an empirical investigation”, Bioethics, Vol. 15(2), 2001, pp. 125-134.

[7] Eric Campbell, Joesl Weissman, Nancyanne Causino, David Blumenthal, in “Data withholding in Academic Medicine: characteristics of faculty denied access to research results and biomaterials,” Research Policy, Vol 29, 2000, p. 305. The authors identify several characteristics of informal networks and the role this underground market plays in determining data access. For example, they note that “the more productive and well known a scientist is, the less likely he or she is to be denied access to others’ research data.” They also suggest “a tit-for-tat exists in which faculty who deny others’ requests for data may be more likely to be victims of data withholding themselves.” (p. 304)

market because their collectors hoard them. Those who hoard data value self-interest over the interest of the research community. Finally, there are those who sell or lease access to data. This treats data as a commodity for purchase rather than as a resource for sharing.

These competing values will shape the discourse around three issues that are important to data service providers: guaranteeing equitable access to data; protecting privacy of human subjects; and ensuring the long-term preservation of data. These three issues are discussed throughout this book.

The information culture of a country also affects the data economy by shaping beliefs about public ownership of data collected by the public sector. Not all countries operate with the premise that government-collected data are publicly owned. For example, in Canada, Crown Copyright applies to all data collected by the public sector allowing the government of the day to determine what information is public and what is not. In the U.S., two strong, competing interests on public information policy exist. There are those who believe that when public funds are used in creating data, the data should belong to the public and access should correspondingly reflect this.[8] Others argue that the private sector should have first opportunity to market public data and the government should not compete in this instance.[9] This results in some public-sector data being sold by the private sector and some being available at little or no cost from the government.

The values of the stakeholders in the data economy and the beliefs of the information culture interact to contribute to or impede open access to and sharing of data.

The Data Marketplace

Data services take place within the context of the data economy, which is a sector of the knowledge economy. The competing values represented by data sharing, data bartering, data hoarding and data commodification as well as the changing norms for conducting research all have an impact on the business of providing data services.

The key stakeholders in the data marketplace consist of data producers, suppliers, service providers, and researchers.[10] Six general categories are used to group these players, which include organizations distributing statistics as well as data[11].

[8] Jacobs, Jim and Karrie Peterson. "The Technical is Political" *Of Significance...* 3(1) 2001, p.25-35. Association of Public Data Users.

[9] See, for instance, *The Role of Government in a Digital Age*, by Joseph E Stiglitz, Peter R. Orszag, and Jonathan M.Orszag, Commissioned by the Computer & Communications Industry Association, October 2000.

[10] Denise Love, Luis Paita, and William Custer, op. cit., describe three major stakeholders competing in the health care market: providers, purchasers, and regulators that set the rules of exchange and use of information. For this economy, they identify three models of data sharing and dissemination: (1) collaboration among providers of health care, (2) coalitions of purchasers, and (3) indirect collaboration among providers and

- Data archives

  The primary functions of data archives are to gather, preserve, and provide access to original research data. As a rule of thumb, the focus of collections in data archives tends to be specific in geographic scope and either general or topical in subject. For example, a special archive in the United States exists for data relevant to human development especially as it relates to women's lives. The UK Data Archive is an example of a general archive that preserves original data from quantitative and qualitative research in a wide variety of economic and social subjects. Some data archives are directly affiliated with a research center engaged in the collection of original data (e.g., UC DATA at University of California, Berkeley's Survey Research Center). Consequently, these and other data archives tend to be closely aligned with the producers of data, particularly academic researchers and government agencies responsible for producing and distributing data. In providing access to their collections, data archives may place conditions or restrictions on some of their holdings as mandated by data depositors. In the data marketplace, data archives are a major source for data.

- Data libraries

  The primary function of data libraries is to support established research communities interested in secondary data analysis by providing access to and assistance with data. Data libraries are most apt to be located in university libraries, computing centers, or research institutes. Sometimes departments or units within a campus or among several campuses form partnerships to create a complement of data services. For example the state university system in California formed a federated membership with the ICPSR to provide data services to its twenty-three campuses. Data libraries are major acquirers of data in the data marketplace and provide a brokerage function between the producers and distributors of data and researchers.

- Commercial data vendors

  Commercial data vendors exist to market data. Some of these vendors repackage free or low-cost government data and sell value-added access to these data. However, the niche most dominated by commercial data vendors is access to proprietary business and economic data. The packaging and sale of security data from markets around the world is an example of the activities of this commercial

purchasers imposed by a third party. We see similar responses in the social science data marketplace in shaping access to data. However, we identify more stakeholders in the social science data marketplace resulting in a wider mix of data access models.

[11] See the discussion in Chapter 1 about the distinction between statistics and data.

data sector. Some vendors promote their data products by offering free value-added services on the Internet such as on-line tutorials or tools for searching variables. In the data marketplace, these vendors are data retailers.

- Government statistical agencies
  Government statistical agencies are responsible for gathering data under legislated mandates. While the requirements of these public agencies to release their statistics and data vary substantially from country to country or even across levels of government within a country, government departments are very important sources of data. The Internet presence of these agencies has increased greatly in the past few years, although often they provide only aggregate tabulations rather than data. Some U.S. statistical agencies do, however, offer extraction services for anonymized microdata. For example the U.S. Census Bureau and the U.S Bureau of Labor Statistics provide microdata subsetting services at their Data FERRET web site. These data sites are among the most rapidly developing services on the Internet. In the data marketplace, government statistical agencies are major suppliers.

- Inter-governmental and non-governmental agencies
  Inter-governmental and non-governmental agencies provide selective statistical and data products that reflect their organization's area of specialization and interests. For example, a particular NGO may be concerned about environmental protection and consequently provide access to data on hazardous waste spills. Frequently, their statistics or data are available for free; however, some charge substantial fees for their data. Examples of IGOs include the International Monetary Fund and the Organization for Economic Cooperation and Development. Examples of NGOs include the International Social Survey Programme and the Center for International Earth Science Information Network. In the data marketplace, these agencies are important suppliers.

- Social Science Researchers
  The social science research community, which includes faculty, affiliated researchers, undergraduates, and graduate students, plays both a producer and consumer role in the data marketplace. Many researchers contribute valuable original data to this marketplace through privately and publicly supported research. Those receiving public assistance through grants in the creation of data are often encouraged, if not required, to deposit a copy of the data with a data archive. In this role, the researcher is an important data producer. Researchers and their students are also significant data consumers. The value of secondary data analysis is gaining wider recognition

> among the social sciences and as a result, greater demand for data sources for conducting secondary analyses is increasing.

The variety of statistical and data products existing in today's data economy is enormous. This exacerbates the blurring between statistics and data mentioned in Chapter 1. For example, many of the commercial data vendors, government statistical agencies and inter-governmental and non-governmental agencies initially distributed their statistical information in print publications. More recently, these organizations have begun distributing their statistics electronically using e-journals, the World Wide Web, and CD-ROMs. Some are now distributing the data products that underlie their statistical publications.

The Social Data Paradigm, or Why Access Still Matters

The dominant research paradigm in the social sciences has its roots in the experimental method, where outcomes of research contribute to an ever-expanding body of knowledge that is captured in a body of literature. In a simplified representation of this method, hypotheses are forged from incomplete knowledge, which surfaces as gaps or inconsistencies in the body of literature. Experiments are designed to test these hypotheses; predictions are made; and data are collected. The results are then assessed against a decision criterion to reject a null hypothesis. Failure to reject the null hypothesis results in a search for a new hypothesis; rejection of the null hypothesis leads to accepting the alternative hypothesis and moves the research outcomes into the body of knowledge. This simplification of the process is shown in Figure 1. Notice that data have a small role in the overall process between prediction and verification.

As secondary analysis becomes an accepted, mainstream research method and as vast collections of research data are preserved and made available to researchers, a new research paradigm has emerged, which we call the Social Data Paradigm. This paradigm is based on both a body of literature and a body of data. The focus of this method is on the construction of models from the body of data and on testing these models with comparable data. In this paradigm (see Figure 2) data play a more significant role than in the experimental paradigm, where data perform an evidentiary function. In the Social Data Paradigm, the body of data can be as important as the body of literature in the formulation of models and in the discovery stage of research. Data are also important in the testing of models and when the data are not from a secondary source, they are contributed to the larger body of data just as the research outcomes are incorporated within the body of knowledge.

Science and policy-analysis are performed using either of these paradigms. This discussion is not about which paradigm is better or more correct. Rather, the point being made is that with greater access to tremendous volumes of secondary data, the norms that dictate the way in which research is conducted in the social sciences are changing. For example, what was labeled as data dredging in the late 1960s and early 1970s closely resembles what is now being called data mining, an emerging method for exploring huge quantities of data. A technique that was viewed as unacceptable in an earlier period has become more accepted because of changes in research norms associated with the Social Data Paradigm. With wider application of the Social Data Paradigm, the demand for data, and consequently data services, is expected to increase.

3.8 - Introducing Data Services

Data Basics

Search Strategies for Social Science Data

The Bibliographic Structure of Social Science Data

The bibliographic structure of social science data collections is organized around two basic levels of description. First, each study is assigned a study-level description. This level, which has recently become known as metadata, tends to provide information about the project under which data have been produced. Typical descriptive elements include a study title, the principal investigator, the data distributor, an edition number, a publication date, and a project abstract.

The other level of description is based upon the content of individual data files within a study that is the variables. This level of description focuses on the substance of each file including the content of variables, the coding scheme used for each variable, the labels assigned to codes, the assignment of missing values, and the physical record layout. These two levels of description -- studies and variables -- determine the context for data searches.

The Study Description

The two general streams from which social science data flow are social research projects and administrative statistical systems. The identity of the data products from either of these two streams tends to begin with a study description. A study title will often serve as the only descriptor for a data collection. Data archives and libraries have long promoted additionally including of the names of the principal investigators, the name of the data distributor, an edition number, and a publication date in study descriptions. The effort to construct comprehensive study descriptions has been driven by the goal to establish the same kind of authority for data that publishers in the print world have for books.

Unfortunately, study titles are not always carefully chosen and may even fail to describe the content of the data. There are instances where titles only consist of a reference to a grant number or to the name of a principal investigator. There are even situations were studies, even though given a proper title, are better known by the principal investigator than the official title (for example, the Parnes data or the Terman data). Despite these pitfalls, titles remain an important search field in study descriptions.

The ICPSR also includes an abstract in each study description and, in some instances, subject headings. The completeness of a study description helps shape the strategies to employ when searching for data.

Variable Descriptions

To date, most on-line catalogues of data collections are built on study-level descriptions. Researchers, however, tend to begin their data searches focusing on the content of variables. For example, a researcher may be interested in describing the demographics of advocates and opponents of gun control in the U.S. While a study may include items about gun control, few projects will focus solely on gun control and, consequently, there are probably not many study titles containing the words: gun control. Good sources for questions about gun control are omnibus surveys or major national or regional polls, but again, the titles for these studies will probably not mention gun control. This limitation presents a large challenge in the search for secondary data sources.

There are a few indices that have been built on banks of questions from surveys. The ICPSR supports a "Social Science Variables Database" (SSVD)[1] that consists of questions from sixty-nine studies that are marked up using DDI.[2] The Institute for Research in Social Science at the University of North Carolina supports an impressive collection of indices of regional and national polls, including the General Social Survey, a significant omnibus survey in the U.S.

Web-based indices of other specialized collections are also surfacing on the Internet. For example, Carleton University provides Internet access to an index for the Canadian Gallup Polls. The University of Alberta offers an index of all of the questions from the Canadian National Election Studies prior to 1992. . Data libraries that use web-based software such as SDA, Nesstar, or Dataverse may create variable-level indexes either for collections for individual studies. More of these kinds of indices will likely appear on the Internet, but the difficulty for the user shifts to locating them and choosing appropriate ones to search and dealing with multiple indices with a variety of interfaces and functionality.

Caution should be taken when using indices of variables on the Internet. Some indices have been created using the actual wording from the original questionnaires. Other indices have been constructed from the short descriptive labels assigned to variables for use with statistical packages. The former index is preferable to the latter since the brevity of variable labels may have omitted important words or phrases or may have used synonyms or summaries in lieu of the original text.

Knowing the Methodology Can Help Narrow a Search

The vast majority of social data files are built from either questionnaires or administrative records. The number of studies originating from direct behavioral observation or simulations is small in comparison. The description of how data were gathered should be explained in a methodology note in the data

[1] http://www.icpsr.umich.edu/icpsrweb/ICPSR/ssvd/index.jsp
[2] Data Documentation Initiative XML format (http://www.ddialliance.org/)

documentation. In one sense, project methodologies roughly divide studies into two general categories, which are analogous to monographs and serials in the print world. First, there are a number of studies based on a one-time data collection methodology. The project's lifeline exists for the duration of gathering the data once. Secondly, there are studies employing a longitudinal method of data collection, which may involve repeated or continuous surveys with different samples or the use of a panel design in which the same sample is interviewed over time. In this situation, the lifeline of the project will entail the production and release of data files for each period of collection.

Knowing a study’s methodology can assist a search by helping match more closely a research question with an appropriate data collection. For example, research that involves examining phenomena over time will require data gathered using a longitudinal methodology.

A study’s methodology can also have an indirect impact on search strategies. One characteristic of on-going data collections is that they tend to have more extensive documentation. The management alone of multiple data collections simply requires greater attention to supporting materials. The impact on searching is that better documented studies have a greater likelihood of surfacing in search results. This in itself is not necessarily cause for alarm, but the person conducting the search should be aware of this potential search-results bias (which is none the less true of any study that is better documented.)

Key Data Sources

Another factor is conducting a successful search for data is an awareness of the key sources of data. The major producers of social data fall in three general categories: agencies within national statistical systems, academic researchers, and a combination of non-governmental agencies and commercial data services. Discovery of the data collections that exist within each of these categories is a major undertaking. All three categories share some common barriers to identifying data as well as to accessing data. Among the chief impediments to access are cost and confidentiality issues, while the barriers to identification tend to be the absence of indices, catalogues and documentation.

National Statistical Systems. The national statistical systems of many countries are a tremendous source of social data. Studies that are part of a continuous data collection program (for example the Current Population Survey in the United States or the General Social Survey conducted by Statistics Canada) monitor significant aspects of society on a grand scale. No individual researcher, for example, is in the position to replicate a national census and yet several countries distribute public use microdata samples of their censuses or have arrangements whereby researchers can submit analysis requests of census microdata.

A shortcoming of national statistical systems, however, is their tendency to collect large volumes of data on a very narrow set of policy interests, narrow both in terms of the content of the variables and of the type of social responses that are measured (that is, behavioral, attitudinal, evaluative, emotive, etc.) For example, the 1985 Canadian General Social Survey focused on health issues of the general population. This survey was designed by a division within Statistics Canada responsible for measuring labor force activity, an area which tends to ask only behavioral questions (for example, how many hours did you work this past week?) Thus, the Canadian survey on health included questions about the number of visits to health care services and other behavioral indicators of general health, including an item about whether a respondent can touch her or his toes. Absent from this survey were any questions about attitudes or feelings toward health care services or providers.[3]

Not only will the policy interests of a governmental agency limit the scope of data, but the public release of its data will require a review process to protect against disclosure. Typically, a government data file will have to undergo a process to remove or modify key variables that might lead to the possible disclosure of a respondent. Thus, geographic areas are expanded, occupational codes are generalized, incomes are grouped, and in some instances variables are omitted.[4] Research interests focusing on a small geographic area or on a rare subpopulation simply will not be addressable with these data.

Many research concerns exist outside and beyond the data collected by national statistical systems. In fact, government agencies are known to collect data outside of their national statistical system to supplement their data needs. These supplemental collections sometimes provide a greater range of questions and content.

Academic Research Projects. Another significant source of data is academic research funded by public and private granting foundations. These funding agencies support major social research projects that play an important role in the creation of data. Some projects have become institutions and are now recognized as part of a national heritage collection. For example, the American National Election studies, the U.S. General Social Survey, and the Panel Study of Income Dynamics -- all heavily funded by the National Science Foundation in the U.S. -- are long-standing projects viewed as resources for a national (and, indeed, international) community of academic researchers. Projects like these have developed extensive data documentation, including detailed user guides

[3] For further discussion of the research limitations of this study see, "The 1985 Canada Social Survey Program: A review," **IASSIST Quarterly**, vol. 3(1), Spring 1989, pp. 3-11.

[4] For example, the public use microdata file for individuals from the 1986 Census of Canada did not include a variable that would identify those with long-term disabilities or handicaps, although the item was asked in the census (question 20b of the 2B Census form.)

and bibliographies of publications based on the analysis of their data.[5] Furthermore, specialized data products and services have been created for some of these major research projects. All three of the studies mentioned above have Internet Web sites with tools for searching documentation and extracting subsets. Another example is the CD-ROM published in conjunction with the American National Election studies.

Many of the public and private foundations provide grants under the stipulation that the researchers who they support will deposit a copy of the data with an established archive and, thus, make the data widely available. The expectation is that data collected by a funded project should become publicly available allowing others to exploit the data. How many studies actually are deposited because of funding requirements is difficult to determine, especially since foundations rarely police such activities. Nevertheless, those foundations stipulating a deposit requirement are promoting an ethic of open access to data.

Beyond the well-endowed academic research projects that tend to support the mainstream interests of social science disciplines, a variety of original data collection occurs through modest institutional support or by researchers financing their own data projects. Knowledge of these data usually becomes public only after a researcher publishes a book or article based on the data. Access to these data by other researchers, however, is dependent upon the willingness of the researcher to share the data.

The proprietary claim by researchers on their data runs headlong into an important norm of science. The expectation that researchers share their data is not based strictly on an altruistic response by the researcher, although this is not to disparage many generous-hearted researchers. Instead, the principle of sharing is rooted in the epistemological necessity of replication in science. In the social sciences, the replication of research findings is a defining characteristic of the science. In this context, the sharing of research data becomes an obligation of the researcher. The norms of science do not always prevail, however, which continues to make data-sharing a major social science issue.

Some journal editors have begun to require authors to submit a copy of the data upon which their articles are based. A few journals in economics and political science are advancing this policy and the ICPSR has begun a special publication-related archive to support files submitted by authors of articles in these journals[6]. In conjunction with this development, IASSIST and some scholarly journals are promoting the proper citation of data within articles. The advantage of data citations in journals is that data titles will eventually appear in the major bibliographic databases, such as the Social Sciences Citation Index.

[5] For example, there is a bibliography of citations of all known works that make use of the American National Election studies. http://www.electionstudies.org/resources/papers/reference_library.htm

[6] http://www.icpsr.umich.edu/icpsrweb/ICPSR/pra/index.jsp

Finally, another academic source for data is the social research unit[7] responsible for collecting data locally at many universities. For example, the University of Alberta has the Population Research Laboratory that has conducted an annual quality of life survey for Edmonton since 1977. Over the years, the collections of some of these university research units have become substantial. Unfortunately, these collections are rarely indexed and discovering them tends to be more serendipitous than by design.

Non-governmental Agencies and the Private Sector. Two other important sources of social data are non-governmental agencies and the private sector, in particular pollsters, marketing firms, and business information vendors. While a special interest group may function as a non-profit entity, many non-governmental agencies finance their operation through the sale of information from their databases. To protect their revenue base, such agencies are reluctant to release their data.

Other non-governmental agencies are willing to sell their data but the prices vary widely, even within agency. For example, the International Monetary Fund sells a series of statistical titles in data format, including International Financial Statistics (IFS), Balance of Payments Statistics (BOP), Government Finance Statistics (GFS), and Direction of Trade Statistics (DOT). The IFS is distributed both on CD-ROM and as a web-based service. For many years, the ICPSR was permitted to distribute BOP, GFS, IFS, and DOT to ICPSR members, But the IMF changed its policy when it started marketing its data more aggressively.

Not only is the price of data a significant factor when dealing with private sector data vendors, but also the conditions by which data are released differ from the other sectors. For example, polling firms are concerned about the confidentiality of the client who initially paid to have the data collected. For example, a company responsible for a major oil spill may have paid for a survey that assesses public opinion about their eco-disaster. Client confidentiality serves as the justification for polling firms not to release data. Ironically, the pollster's client is given the same protection against disclosure as the respondents who provide the data.

Strategies for Locating Social Science Data

With the proliferation of computer-readable data available on a variety of media including CDs, DVDs, research data files found their way into libraries at an increasing rate. With licensed network access to data, and web-based services, libraries often find themselves as both facilitators and service providers. No longer restricted to magnetic tape and mainframe computers, or even to CDs and

Data Basics

[7] When the economic conditions of the 1990s required many of these units to become self-sufficient, the long-term availability of some of these collections was jeopardized, either because fees were attached to the data or because the research unit was closed. This is an example of one of the many ways in which data are at risk.

DVDs, research data files are more directly accessible to the researcher at her or his desktop. With this change, helping library users locate appropriate web-based sources, identify appropriate data files, and combine and restructure data from multiple sources have become important services in libraries.

Finding an appropriate data file is complicated by a number of factors. Knowledge is required about how data are collected, how they are organized, and how they can be used. Furthermore, being able to distinguish between a request for information and a request for data to be used in a quantitative analysis is important. A statistical table may suffice for the informational request, but the original data from which the table was created may be sought for a researcher who wishes to conduct an analysis beyond the published table.

Locating social science data employs strategies similar to those used in government documents and in traditional archives. Both strategies are characterized by in-depth assistance at the research level and reliance on non-traditional methods for identifying sources.

Search Strategies for Data: The Government Documents Approach.[8] Two common approaches for locating information in traditional reference, government documents, and data services are *known- item searches* and *subject searches*. Known-item searches are fraught with difficulties, whether in government publications or in data services. Publication patterns and distribution of these materials are often mysterious. Standard citation practices have not been practiced and many titles begin or end with the word "National" or "Survey". Acronyms are frequently used and are often ambiguous. Keyword searching is only helpful if there is enough information to go on. The agency or principal investigator responsible for the publication or data may be helpful in narrowing a search, but they may not be known.

In government documents and data services, subject searches often require exploring more catalogues and specialized tools than when looking for a book or a journal article. The use of commonly accepted subject headings with data files is problematic. For example, Library of Congress subject headings simply will not capture the complexity of a large survey containing hundreds of questions on different topics. To find specific topics raised in the individual questions of a survey, the full text of the questions requires searching. The sheer volume of this task precludes most surveys from being searched in this way.

However, three strategies particular to government documents reference, do work well with data files: *agency, statistical, and special techniques searches*. Many data files are collected, produced, or distributed by government agencies. Thus when looking for a data file, common questions to ask are: "What agency

[8] *Using Government Information Sources: print and electronic*, by Jean L. Sears and Marilyn K. Moody. 2nd edition. Phoenix, Arizona: Oryx, 1994. In its second edition, this work outlines strategies for locating government information in the United States although there are many similarities to national governments elsewhere.

collects these data? " or "What survey focuses on health issues and who produces it?" On-line or print reference tools may confirm an agency's publications, although sometimes it is necessary to call the agency directly to acquire or access data. Some government data contain confidential information and require special approval for their use. This may entail a written statement detailing how the data will be used before permission is granted.

The statistical search is commonly used to locate both documents and data files. Where a search in government documents may end with the location of a particular statistical table or census report, the data search may begin by first locating a document containing related information to what is being sought. The data source for this related document then becomes the target of the data search. The original data file often contains more information than is summarized in a report and will allow researchers to analyze the data in a myriad of ways.

Specialized tools exist that can be used to facilitate the statistical search. Indexes to statistics (e.g., *American statistics index*, and *LexisNexis Statistical*) can be useful in finding statistical tables, which, when examined, will lead to a data-source. Citation indexes (e.g., *Social Sciences Citation Index* or *ISI Web of Science Cited Reference Search*) are an imperfect, but sometimes fruitful, way of tracking down research articles that have used known datasets, which can sometimes lead to related datasets. The ICPSR *Bibliography of Data-Related Literature*, which is a searchable database of over 41,000 citations of known published and unpublished works resulting from analyses of data held in the ICPSR archive, can be used to search for articles by topic and then discover the data analyzed to produce those articles.

Special techniques searches are usually multi-faceted. For example, data on particular health issues may be sought, but only as one criterion of the search. In addition, the researcher may be interested specifically in Black Americans. The search for a data file must ascertain both if questions about particular health issues are contained and if the sample of the data file includes the sought after group. Besides special samples, other elements related to the subject content may be important in a multi-faceted search, including the unit of analysis or what is being studied, the methodology employed in doing the research, or the level of measurement used for a particular variable.

Search Strategies for Data: The Traditional Archives Approach. Traditional archives[9] and data archives have many similarities. Both contain primary source materials that are often used in ways not originally intended by their creator. The wonderful aspect of secondary analysis in social science research is that another researcher can use data collected for one purpose to solve a different problem.

[9] A good reader on traditional archives is, *A Modern Archives Reader: Basic Readings on Archival Theory and Practice*, by G. Chalou. Washington, D.C.: National Archives and Records Administration, 1984.

Furthermore, the results of one survey can be compared to or combined with results of another to raise yet new issues.

In addition, both types of archival collections are labor intensive to use, may not be organized down to the level of specificity that the researcher needs, and may benefit greatly from the experience of staff familiar with the data file or a special collection of letters. It is not uncommon for data services staff or archival staff to work with a patron on several or many successive occasions. In fact, often a researcher starts with one set of questions and changes them according to what is available.

Techniques employed in both government documents reference and in traditional archives are useful when conceptualizing how to provide reference services for social science data.

4.10 - Search Strategies for Data

Data Basics

Data-speak: A Search Vocabulary for Data

The Unit of Analysis and Its Importance

The common foundation for all social science data collections is the unit of analysis. This is the object or objects about which data have been gathered and about which generalizations can be made. With social data, three elements of a research design define the unit of analysis. These include the observed social entity or entities, the role that time plays, and the spatial setting. While all three are present in a study, usually one is dominant while the other two are contextual. Before discussing these three attributes in greater detail, the importance of the unit of analysis in the data reference interview is presented.

The focus of the data reference interview should begin on the unit of analysis. A patron will likely start with a question about a subject, for example, “I’m interested in gun control” or "I'm looking for data about gun control". Before beginning a subject search, however, the data librarian should first to establish the unit of analysis with which the patron wishes to work. Continuing the example about gun control, does the patron wish to locate data about general public opinion of gun control; or does she want the opinions of just gun owners; or is the research interest about national policies on gun control and corresponding homicide rates. If the interest is in general public opinion, the unit of analysis is likely to be all adults in a country; however, if the research focus is on a special subpopulation, such as gun owners, the unit of analysis will remain individuals but individuals belonging to a special group. Finally, if the interest is in national policies about gun control, the unit of analysis will be nation states.

There will be occasions when the patron does not have a clear idea about the unit of analysis with which she wants to work. In this situation, the data librarian should help the patron focus on an appropriate unit by asking a few probing questions. “To whom would you like your research generalizations to address? Are you interested in summarizing your results about the general public or about a specific group? Are your interests about change in opinions over time? Do you want your generalizations to be for a specific geographic area, for example, just western Canada?” The discussion below will show how the answers to these questions will clarify the intended unit of analysis.

In other instances, data on the subject will be available for one unit of analysis but not for the unit that the patron desires. In this situation the data librarian should inform the patron that the unit of analysis for which data are available will not permit the generalizations for the unit that she desires; however, data on the subject do exist for another unit. For example, a researcher may want to generalize her analysis to adults living in western Canada but the data are from a national sample that can only be generalized to all of Canada. The researcher will have to decide if she is willing to accept this restriction.

Identifying the Unit of Analysis.

The sampling methodology reported in a study description[1] can be helpful in identifying the unit of analysis. The sampling method explains the selection process of the objects about which data were gathered. A simple sampling technique will quickly identify the primary unit of analysis. For example, the sample may have been of all Canadians 18 years and older who were not institutionalized. The pool from which the sample is drawn is also known as the sample universe or simply the universe. Check the study description to see if it describes the sample universe.

A more complicated sample design may employ an elaborate selection technique that obfuscates the real unit of analysis. For example, a sampling method may begin with a selection of households and then with the selection of an individual in each household. By initial appearances, the conclusion may be that households were the unit of analysis. This sampling technique is used frequently to identify special subpopulations. To locate a sample of individuals with a disability, a large number of households are first selected and then an individual in each household is contacted to see if anyone in the household is disabled. This approach is also used to identify individuals who have been victimized. In both of these examples, the unit of analysis is individuals, not households.

Some studies have multiple units of analysis. For example, a study about the disabled may collect data on the individual or individuals within a household who are disabled and also collect data on all other individuals in the household. A study examining care for the disabled may employ this kind of research design. In this instance, the disabled will serve as one unit of analysis while the caregivers will be the second unit of analysis.

Social Units, Time and Space

In a research design, the social unit, time, and space define the unit of analysis. In most social surveys, the social unit is the dominant element while time and space provide context. For example, a study may be of the adult population in the United States during the month of July in 1997. The unit of analysis will consist of individual adults, while time (July 1997) and space (the United States) are fixed.

Social Units. The individual is a key social unit in both survey and experimental research. The large abundance of opinion poll data is built upon the individual as the social unit. However, the individual is not the only social unit of interest. Families and households are important units for examining support or economic groups in society. While these units are composed of individuals, there are

[1] For a definition of study descriptions, see Chapter 2.

characteristics of families and households that transcend the individual and that are unique to the family or household. For example, total number of siblings is an attribute of the family; total household income is a household variable. Other important social units including ethnic, racial, and immigrant groups. Organizations can also serve as valuable social units. Examples of this type include unions, universities, companies, and political parties. Even nation states or societies fit into this classification

Social units can be combined or grouped within a study. In some instances these units are in an hierarchical relationship. For example, households and individuals together may constitute the social unit of a study. The Public Use Samples from the United States Census are an example of this relationship. In these data files, individuals are nested within households, that is, the data for each individual in a household are organized on separate records that fall below an overall record for the household.

In addition to an hierarchical design, multiple social units can be organized in a network relationship. In this instance, no single social unit is necessarily subservient to another unit but rather is linked to other units through a network. For example, a study may contain data about students, their parents, their teachers and their school system. Linkages among these units exist in a variety of ways. For example, the students may be linked to their parents to form a family unit, or students and teachers may be linked to form a classroom unit.

In addition to hierarchical and network relationships, social units may be organized according to research constructs. For example, the concept of a married couple is important in family studies. In working with data about couples, a researcher might collect data from both spouses and then make comparisons between the variables of each couple. This dyadic structure takes data from individuals and permits the analysis of couples, which is a social construct.

In the large majority of studies where the social unit is the dominant element, time and space are usually constant, that is, they do not vary within the study. However, data may include variables containing information about time and space. For example, the 1992 General Social Survey of Canada collected time-use data about Canadians over the course of a year. The full sample was divided equally into 12 groups and then each group was assigned to a month. Thus, the final data collection includes the month of the year during which each respondent was interviewed which permits aggregate comparisons of time-use across seasons. A researcher studying seasonal affective disorders would find this helpful.

Time. As mentioned above, most social surveys are snap shots of a single point in time. However, when the focus of research is on changes in human behavior, data observed at a single time point are of marginal use. Three general methods exist for investigating change in the social unit (this applies to organizations as well as individuals). First, there are the *repeated or continuous surveys*. These surveys maintain the same content over time, that is, the same questions are always asked, but they are asked of new samples. There is a small probability that the same individual could be chosen in more than one of the samples in a large omnibus survey such as the General Social Survey in the United States, but no way would exist to link that person's responses between surveys. A typical strategy for analyzing change with repeated or continuous surveys is to compare age cohorts. Thus, change is descriptive of the aggregate and not of the individual.

One method that specifically addresses change in individuals employs a *rotated sample*. Like the continuous survey, a battery of questions is asked repeatedly over time. However, this methodology uses the same sample for a fixed number of repeated interviews. This approach is popular in tracking labor market activity. An individual in the sample may be followed for four points in time and then her sample group is dropped from the survey and a new group is introduced. Changes in behaviors and attitudes can thus be explored both at the individual level as well as the aggregate level using this methodology.

A second survey method for examining change in individuals is the use of a *longitudinal design*. Rotated samples tend to reveal short-term change since respondents may only be in a pool for up to one year. The labor force surveys are taxing and to ask respondents to participate for longer than a year is hardly fair. Longitudinal surveys, however, keep the same individuals in the sample for years. For example, the Panel Study of Income Dynamics in the United States has been conducted annually of a core sample since the late 1960's. Terman collected data from the same individuals from the late 1920's into the 1970's. Clearly, these studies permit an examination of lifetime changes and not just short-term changes.

Up to now, time has been discussed in the context of observing change in a social unit. The three survey methods mentioned above all permit some examination of change in the social unit. Time can itself however become the dominant unit of analysis. Data organized in this fashion are typically called a *time series*. In this instance, all observations are ordered by time. A typical example of a time series is stock market data, which may be organized daily and contain high, low, and closing prices. Event data that involve transactions are also usually organized in a time series.

Space. The spatial unit in survey research usually makes reference to the place where the data were gathered. In this context, political boundaries form the typical spatial unit in a study. Surveys take place within a city, a province or state, or a nation. Space is usually a very important element in the sampling design. For example, a stratified sample will typically begin with a sample of geographic units and then proceed with a selection of households. In this application, the spatial unit assists in identifying the sample for the social unit.

A key social data application of the spatial unit is census geography. While census results are often presented using political geo-references (urban centers, provinces or state, or national totals), enumeration is based on a unique geo-coded system. In Canada, these enumeration areas are the smallest reporting unit for census statistics. Consequently, enumeration areas serve as the building blocks for describing larger geographic areas.

Spatial units also play a prominent role in administrative geography. Any service that requires the allocation of resources over space will employ some method of geo-referenced data. This includes fire and police administration, public schools, and health care delivery. Social data are then summarized within these spatial units. For example, data for public school zones will likely include the number of children in each age category from 5 to 18; police service areas will likely include the number of break and entries in neighborhoods, etc. Another commonly used spatial unit is the postal code, which has become a popular marketing unit.

Level of Observation

Microdata represent the lowest level of observation at which data were collected for a particular unit of analysis. If an individual person was the level at which observations were made in creating data, the microdata consist of the records for all individuals in that study. Researchers tend to prefer microdata because such files are at the required level to permit generalizations about the unit of analysis. Microdata files also present the data in its most manipulative form.

Data that have been summarized from microdata and where the originally observed unit of analysis is no longer identifiable are known as *aggregate data*. The summary tables of censuses released by national statistical agencies are an example of aggregate data. The master file of individual census returns, which is protected for reasons of confidentiality, constitute the microdata[2]. The summary tables are aggregations of these individual census returns over various levels of census geography, such as enumeration areas, census tracts, census subdivisions, census metropolitan areas, etc. From the summary tables, no data for any particular individual can be identified[3].

[2] Anonymized versions of census master files have been prepared in Canada, the United Kingdom, and the United States and are known as public use sample or microdata files. These files are one to five percent samples of the master files where steps have been taken to prevent disclosure by reducing the amount of information they contain. Nevertheless, these products provide valuable access to census data at the microdata level.

[3] Of course, if a cell in a summary table had a count of only one or two people, it may be possible given the census geography of a table to construct some individual level data. To prevent this, national statistical agencies tend to assign randomly the counts of zero, five, or ten to table cells in which the original counts are ten or less.

When an analysis is based on aggregate data, the researcher must take care not to generalize to the wrong unit of analysis. With the census summary tables, individuals are no longer the unit of analysis, rather the table's spatial representation of the census geography constitutes the unit of analysis. To misattribute the unit of analysis from aggregate data is to commit the *ecological fallacy*.

Most aggregate data are created by summarizing the social unit of a microdata file over its spatial and time units. Census summary tables are an example where the spatial unit is used to construct the aggregate structure for summarizing individual census returns. A table showing the number of deaths and live births over a fifty-year period for one state may have been constructed from vital statistics records aggregated over year of death and year of birth within a specific state. In this example, a combination of time and spatial representation constructs the aggregation. All aggregate data files should report the microdata from which they were derived.

Reading Data Documentation

What to Look for in Data Documentation

Good data documentation is essential regardless of whether the data are from a social survey, an experiment, or administrative records. Such documentation will contain specific information about both the **context** and **content** of the data files belonging to a study. The context from which the data originated will help establish appropriate uses of the data for secondary data analysis. This type of information will provide details about the unit or units of analysis in a study. Such topics include the identification of a study's population, a description of any special steps required to analyze the data (for example, correcting for the sample design), and an account of the participation or response rates if the data are from a social survey.

Good data documentation will also contain detailed information about the content of a data file. This type of information will describe the source of each variable (for example, did the variable come from a specific question in a survey or was it derived), explain the coding and location of variables in the record layout, identify missing value assignments, and report how derived variables were created. Without this information, a data file may be useless[1].

Preparing good data documentation, unfortunately, is not a highly rewarded activity. Consequently, the quality of documentation varies greatly. Because of this wide variety, knowing what to look for in data documentation is all the more important. Expect high documentation standards; do not be satisfied with incomplete documentation.

Documenting the Unit or Units of Analysis

Knowing the purpose of a study often helps clarify the context in which the data were collected. For example, a social survey may be conducted to evaluate services, to identify a market, to obtain the mood of the public, to direct policy or decision-making, to conduct basic research, to track change, to establish a benchmark from which change can be assessed, or a combination of these reasons. Each of these objectives will push the collection of data in a certain direction. Consequently, an understanding of a survey's purpose can provide important insight into possible secondary uses of its data as well as specific limitations of the data. Good data documentation will include a section describing the purpose of the study.

[1] Some statistical software will save data with descriptive information about the data but these formats are not a replacement for well documented file descriptions. Similarly, statistical command files containing instructions for software to read data files may be distributed instead of detailed variables descriptions. Again, this is not an acceptable substitute for good documentation.

Data documentation should provide information about the **population** of the data collection. Furthermore if the data are from a social survey, the **sampling methodology** should be described. The population, also known as the universe, encompasses everyone who was eligible for inclusion in a study's data collection. In a survey, the sampling frame serves as the list representing the population from which individuals are chosen to participate in a survey.

Knowing the population of a study helps clarify its appropriateness for a secondary research topic. For example, someone studying early adolescence would not find data useful from a survey where the population age was 18 years and older. Using another example, the population of most Statistics Canada surveys excludes individuals who are institutionalized. Thus, researchers interested in studying people receiving institutional care will not find these surveys appropriate and will need to find alternative data sources.[2]

The sampling methodology explains the rules about how those who participated in the survey were chosen from a specific population. If all of the respondents have an equal probability of being selected, there is little concern about the selection process biasing the sample. However, the sampling techniques commonly employed in national samples rarely utilize a simple, random selection method. Instead, the methods often oversample certain individuals who might otherwise be missed in a simple, random selection. By oversampling, the probability of being selected favors some individuals over others, which thus creates a bias. In such situations, a procedure must be introduced to correct for the unequal selection probabilities before generalizations about the population can be made from the data.

An important mission of data documentation is to communicate the steps needed to adjust a sample if oversampling is part of the sampling design. Statistics Canada, which employs complex sampling methods, calculates a **weight variable**, which is included as part of a public use microdata file. Applying this special variable in an analysis adjusts each respondent’s selection probability putting the individuals in the sample on equal footing.

The weight variable included by Statistics Canada often corrects not only for the sampling methodology, but also scales the sample size up to an estimate of the population. Thus, the sample size may be 12,000 while the weighted-size may be 23,000,000. If the sampling weights include a scaling factor, this information should be included in the data documentation. Some researchers will want to work with the sample size in their analyses rather than the population estimate[3].

Other contextual factors also contribute to the applicability of a survey’s use in

[2] It should be noted that the Canadian National Population Health Survey includes samples from separate populations for the non-institutionalized and the institutionalized.

[3] To work with the sample size and still correct for the sampling design, the weight variable must be rescaled.

secondary data analysis. The geographic coverage and detail of spatial levels may be important. For example, the province is the lowest level of spatial analysis one can achieve with most of Statistics Canada public use microdata files.[4] If a researcher is interested in summarizing results for local health authority districts, for example, she or he will not likely find any of the standard public use microdata files of use.

Time is another contextual factor. The public release of microdata files usually takes a year or longer after the survey has been conducted. The preparation of the Canadian Census public use microdata files, for example, takes three years before the data are released. Thus, research interests requiring very current data will not likely be met with Statistics Canada microdata files. Furthermore, the data of interest may not be collected at an interval frequently enough to be useful for some research. For example, topics in the Canadian General Social Survey are on a five-year cycle.

Documenting the Content

Working with data documentation often entails investigating both the original questionnaire and the accompanying public use microdata file. Both of these products should be examined because the public use microdata file will rarely, if ever, contain the full information captured on the questionnaire. Before releasing a microdata file to the public, Statistics Canada and other national statistical agencies takes steps to anonymize the records within a file to minimize the likelihood of individual respondents being personally identified. Three general practices are followed in anonymizing a file. First, the detail of spatial information is reported only for gross levels of geography. Province is often the only level of geography below the national level provided in a public use microdata file. For surveys with very large samples, such as the individual public use microdata file from the Canadian Census, urban areas of 250,000 people or more, that is, Census Metropolitan Areas, may be included. Certainly, no small-area geographic units are available in public use microdata files.

The second method used to anonymize records is to replace detailed response categories with fewer, more general categories. For example, no public use microdata file will contain four-digit occupational codes. Rather, a general level of occupation is reported using approximately twelve categories in total. For example, applying this technique would change the classification of a *cardiologist*, which is very a specific occupational category, to a *professional*, which is very general occupational grouping including many non-medical professionals, also.

The third method used to protect the identity of respondents is not to release the

[4] The reason for this limited level of geography is discussed under the topic of anonymized files in the next section about content.

information in the public use microdata file. These suppressed variables remain, however, in the master file from which the public use file is produced. Most data documentation for microdata files does not include a list of suppressed variables, although the Canadian Survey of Labour and Income Dynamics is an exception. In these instances, the only way of knowing what has been withheld is to compare the items in the original questionnaire with the documentation for the public use file.

A further reason for comparing the contents of the original questionnaire with the data documentation is that statistical agencies often enrich the original data collection by preparing **derived variables**. For example, a series of questions may be asked about the number of people living in a household and their relationship to one another. Statistics Canada may then combine this information into a single variable that reflects the living arrangements of each respondent in the survey. This simplifies the need to locate, for example, single-parent families within a survey.

The questionnaire also provides the detailed script of the context and the sequence in which answers have been elicited. For example, question order may be important to a researcher. The order of the variables in the data file, however, may not have any relationship with the sequence in which the questions were asked in the survey. Furthermore, knowledge about skip patterns[5] might be critical when selecting certain variables. Answers to one question may skip respondents to different sections in the questionnaire. The only way to determine which questions are applicable to a group of respondents may be to follow the skip pattern within the questionnaire. A copy of the original questionnaire is often the only source for finding this type of information.

A **data dictionary** and **record layout** are two additional parts of good data documentation. The data dictionary documents how information has been transcribed from the questionnaire and coded into variables in the data file. Variables, in turn, are organized in a computer file and assigned specific column locations on a record or line. The document that describes these field or column assignments is known as the record layout. Some data documentation integrates the data dictionary and record layout into a single format. This integrated format is commonly called a **codebook**.

[5] A skip pattern in a questionnaire occurs when respondents are not required to answer a series of inapplicable questions. A branch question is usually employed to identify whether a respondent should continue to follow a line of questioning or should instead skip to another section of the questionnaire. For example, before asking a respondent a series of questions about their recent experiences with higher education, a branch question may be asked, "Have you attended a post-secondary institution either part or full-time within the past twelve months?" Those answering "yes" would continue with the education questions; those who answered "no" would be skipped to another section in the questionnaire.

Questions and Variables

Working with social survey data requires knowing how information from a questionnaire has been converted into variables. The simplest situation occurs when one question produces one variable. Since a variable can only hold one value, this one-to-one correspondence between a single question and one variable only exists for questions eliciting a single answer. However, questions that permit multiple answers will require multiple variables to capture all of the information. The best way to understand how variables are created from questions is to look at some examples.

In the 1994 Canadian National Population Health Survey, question HHLD_Q4 asked, *"Is there a pet in this household?"* The response categories are:

Yes
No (Go to HHLD_Q6)

Clearly, only one answer can be given to this single question. Thus, one variable with codes for the two response categories (yes or no) will capture all of the information for this item[6].

The next question (HHLD_Q5) asks, *"What kind of pet?"* and includes the instruction, *"Mark all that apply".* The three provided response categories are:

Dog
Cat
Other (Go to HHLD_Q6)

For this question, more than one response is permitted. A household could contain a dog, a cat, and some fish, in which case all three of the response categories would be checked. One question exists but more than one response is acceptable. Consequently, a convention for assigning variables must be devised that captures all of this information.

One method is to treat each response as a separate variable. There would be a dog variable, a cat variable, and an other-pet variable each indicating if that particular pet is in a household. Another convention would be to summarize the multiple responses into a single response that is assigned to one variable. The latter convention was actually used in the Statistics Canada study. The answers were all grouped into one of two categories:

At least a dog or a Cat
Other only

[6] To be comprehensive, another code to indicate non-response will also be assigned. The complete response set for a variable includes all of the pre-assigned response categories and the set of codes used to designate non-responses.

“At least a dog or a cat” means that either a dog or a cat was checked with the possibility of another pet also. “Other only” means that neither a dog nor a cat was marked but another pet was. The way the response categories are grouped actually results in a loss of information. It separates households with pets that are not dogs or cats, but lumps together those with dogs or cats with other pets.

The above examples show the possible creation of variables in the following ways:

one question	**→**	**one answer**	**→**	**one variable**
one question	**→**	**multiple answers**	**→**	**multiple variables**
one question	**→**	**multiple answers**	**→**	**one variable with a possible loss of information**

Because statistical agencies anonymize records in their public use data files, two further methods are used to deal with information in questionnaires.

one or more questions	**→**	**one or more *derived* variables**
one or more questions	**→**	**no variables**

An example of a ***derived*** variable from the Canadian National Population Health Survey is the five-category income adequacy variable, where the categories are lowest income, lower middle income, middle income, upper middle income, and highest income. This variable was itself created from two other derived variables – one estimating the size of the household and the other reporting total household income.

An example of a question from this study that was suppressed in the public use file is, *“What is respondent’s date of birth?”* The questionnaire captured this information in DD/MM/YY format. However, a variable does not exist for date of birth in the public use file.

Variables and Values

Variables are commonly identified in documentation using an eight-character mnemonic name. This convention arose because statistical packages restrict variable names to eight characters. These names sometimes refer to the original question numbers. For example, LFS_Q1 is the first question in the LFS[7] section of the questionnaire for the Canadian National Population Health Survey. Variables beginning with DV in Statistics Canada studies are very likely to be derived variables.

[7] LFS deals with labor force status.

Other variable names are quite descriptive of their content. SEX is an obvious example. MARSTATG is the variable name for marital status group. NUMBEDRM is the variable containing the number of bedrooms in a dwelling. However, do not confuse WEIGHTKG, which is the respondent's physical weight, with WT6, which is the respondent's sampling weight.

Two general groups of values are assigned to variables. Values either represent categorical descriptions, such as female or male, or they represent some measurement, total sum, or count, such as the number of bedrooms in a dwelling or the physical weight of the respondent. If the values of a variable are categorical classifications, the variable is known as a **categorical variable**. If the values of a variable are a measurement, total sum, or count, the variable is known as an **analytic variable**.

Another way of classifying variables is in terms of the levels of measurement used in the social sciences. Nominal measurement consists of assigning numbers to categories, which is the same as a categorical variable mentioned above. The numbers have no cardinal or inherent meaning. The category "female" can be assigned the value 0, 1, 2 or any other number, while the category "male" can be assigned any number other than the number assigned to "female". The only connections between the numbers assigned to the categories are convention and coding efficiency (for example, it would not be efficient to assign the value 1111111 to Female and 1111112 to Male).

Analytic variables, however, consist of values containing some properties or relations between the numbers and the content of the variable. If a measurement distinguishes greater or lesser levels of a property without an exact measure in units, the variable is said to have an ordinal level of measurement. The numbers assigned will reflect the increase or decrease of a property without indicating exactly how much the property has gone up or down. For example, DVINC595 uses the values one through five to indicate levels of income adequacy.

1	Lowest Income
2	Lower Middle Income
3	Middle Income
4	Upper Middle Income
5	Highest Income

The scale of measurement for this variable clearly indicates that as one moves up the scale, the income adequacy increases. This scale does not provide, however, the precise amount of "income adequacy" obtained by moving from 1 to 2 or from 2 to 3 (in other words, we don't know if the amount of "income adequacy" measured by subtracting 1 from 2 is the same amount as subtracting 2 from 3). We simply know that by moving up the scale, we increase the level of income adequacy.

When the amount of increase, as we move up the scale, is known in specific units, the level of measurement becomes interval. For example, if actual household income is a variable, the unit of measurement would be in dollars (this is also known as a metric). We know the exact amount of the increase of someone who moves from $20,000 to $30,000, namely, an increase of $10,000. If in addition to having equal intervals, a scale also has the property of an absolute zero[8] then the level of measurement is said to be ratio. A ratio scale also allows one to discuss different values as being multiples of other values. Thus, one can say in the instance of income, someone with an income of $40,000 has twice the income of someone with an income of $20,000.

Variables and Frequencies

Good documentation will also include the frequency distribution of categorical variables. Ideally, if a study includes the use of a weight variable, the frequencies should be reported for both unweighted and weighted results. For example, the frequencies for the variable SEX in the Canadian National Population Health Survey shows the unweighted frequencies to be 8,058 males and 9,568 females. The weighted frequencies for males and females is reported as 11,780,335 and 12,168,269, respectively.

The unweighted frequencies can be used to help verify a data collection. The frequencies for the variable SEX in the raw data file can be obtained using statistical software and then compared with the data documentation to confirm that the file has been read correctly and that the documentation matches the file. This is particularly important when working with studies that have multiple data files or when multiple releases of a data file exist. If the frequencies do not match, the version of the data file may be wrong or the match between the documentation and the data may be wrong.

Frequency distributions are also useful for identifying sample sizes of subgroups within a file. For example, a researcher may only wish to work with those in the age range of 12 to 14. In the Canadian National Population Health Survey, the documentation shows a total of 637 cases in this age range. This may not be enough cases for the type of analysis that the researcher has in mind.

[8] An absolute zero means that the content of a variable can be completely absent. We know that zero degrees Celsius is not the absence of temperature. Thus, the Celsius scale is not a ratio scale. However, an income of zero dollars is an absence of income and, consequently, a ratio-level variable.

A Framework for Defining Data Services

Previous chapters defined social science data, identified the role and value of data as a significant resource in society, and reviewed the institutional arrangements for supporting data over the past twenty years. The focus now shifts to a discussion about the specific nature of services for data. This chapter should be useful both for those beginning a service and those who are looking at ways to reformulate or revitalize an existing service.

We can best define data services through an understanding of several interlocking contexts: the mission of the library, principles of the social science data community, the essential place of service in defining the role of the data library, and the context of technological change. We offer a framework that builds on these contexts and that, in conjunction with best practices, can assist in establishing or revitalizing a data service.

The Mission of Data Libraries

What is the purpose of a library today? If you ask a dozen people you will probably get a dozen different answers shaped by their perspectives and needs and uses of libraries. For instance, students often think of the library as a study hall or their office on campus; many reference librarians see the library as a classroom for teaching information literacy skills; some library directors view the library as like a bookstore and a few have even installed coffee shops in their buildings to appear all the more like a Barnes and Noble or an Indigos. Small public libraries sometimes include as part of their mission providing community space, meeting rooms, and even art galleries. As resources in the library increasingly require the management of license agreements with publishers and information vendors, the library has taken on the appearance of a business office. Librarians are even seen by some as information brokers, helping users find information regardless of its location or presence in the local collection.

Each of these perspectives of the library gives us a useful, but limited, understanding of the purpose of libraries. Each helps us see a part, but only a part, of the library without grasping the whole picture, without understanding the essential role of libraries. To get a better picture, we need to understand the role a library plays in its community, in society at large, and in the life-cycle of information.

We believe the role of a library is to select, acquire, organize, and preserve information for its community of users and to provide access to and services for this information. While other organizations may provide some of these functions or may fulfill complementary roles, few if any organizations other than libraries have these functions as their *primary* mission. The defining characteristic of a library is that its primary function is a commitment to all of these activities. We will frequently refer to these functions as simply *collections* and *services*.

In the digital age and when we are dealing with social science data that is, by definition, digital, the connections between the six OAIS functions take on new importance. It is impossible to preserve information without accounting for its access and usability and it is impossible to assure access and usability if the data are not preserved. (See Chapter 9 for more details on Preservation.) Stated another way, the mission of *data libraries*, in particular, is to remove barriers to access to data while ensuring long-term access and usability of data through preservation activities.

Principles of the Social Science Data Community

The social science data community has a decades-long tradition of principles that drive the collection and use of data. These include equitable access to data, the protection of privacy, and the preservation of data, which we described in detail in Chapter 3 ("Introducing the Data Marketplace"). As strong proponents of these principles, we place paramount importance on models of service that preserve and defend them.

The Service Model

For many years libraries off all kinds expressed their value by measuring the size and scope of their *collections*. Today, many in the library community believe that we can better judge a library by its *services*. Recent studies suggest the value of a library is a combination of financial value and the value of the "impact" of the library.[1]

We believe that, in order to provide services with an impact, a library must be able to select and control a collection of information. A collection without an explicit *service mission* is not a library and a service without a body of information in its control is likely to be little more than an unwanted intermediary. This is true of traditional libraries, but even more true of data libraries because data files require more service than books to be findable and usable and preservable. There are many reasons for this, which we explore in more detail later, but one of the most important reasons that data require services as an integral part of collections is that every data file is different and requires documentation to use. Finding data is not as simple as looking up a subject or author or title in a catalog. Using data is not analogous to picking a book off a shelf and starting to read.

[1] Association of College and Research Libraries, & Oakleaf, M. (2010). *Value of Academic Libraries: A Comprehensive Research Review and Report*. Chicago, IL: Association of College and Research Libraries. Retrieved from http://www.ala.org/ala/mgrps/divs/acrl/issues/value/val_report.pdf

To emphasize the need for services integrated with collections, we approach data library issues from the perspective of a *service model*. The value of a service model is its underlying commitment to the needs of the users of the data collection. With such a perspective one can evaluate each activity of a data library as to whether or not and to what degree it fulfills a needed service goal. In a later chapter, we explore in more detail the development of a service plan to implement the service model.

A Framework for defining levels of service

When designing a new data service or evaluating and revitalizing an existing service, it is necessary to have a general idea of what kind of service you want. Before you get into the details of precisely what services you will offer or even who the primary data users are, it is useful to have a general idea against which you can evaluate the decisions you make in designing or evaluating the new service. We refer to this as a “framework” and include in it the general mission, principles, and service model described above.

In addition, this framework utilizes a typology constructed of three types of innovation and three components of library service. All these elements together provide a context within which one can make choices based on available resources and skills and can scale activities and services appropriately and do all of this while staying focused on principles and purpose. Later we will use the flexibility this affords to examine different *levels* of service that are possible with different resources.

As we will see later in this chapter, this approach also helps you avoid the temptation of letting technology drive service decisions.

For many institutions, establishing data services will be an act of innovation requiring a new look at established policies, procedures, and organizational values. At the same time, the core functions of the library and the principles of equitable access, protection of privacy, and preservation and usability of data must be maintained.

The task of simultaneously planning innovation, incorporating rapidly changing technology into a new service, and protecting traditional values is difficult. Understanding where new services might direct an organization and how new services relate to old ones can help an organization plan for a new service. Knowing how innovation and service components influence each other should therefore bring clarity to the process of defining and choosing which services and what levels of those services to offer. We begin by looking at three kinds of technological change or innovation.

Three Kinds of Technological Change

Those interested in the sociology of change and the effects of technology on the workplace often identify phases or stages of change. Several writers describe three stages of technological change or innovation in similar ways. Richard Lucier, speaking about university libraries, calls these the three stages of technology diffusion: modernization, innovation and transformation.[2]

Clifford Lynch, addressing the issue of library automation, uses these same terms and elaborates by describing a three-phase procession of the effects of information technology on organizations.[3]

1. Modernization -- doing what you are already doing, though more efficiently;
2. Innovation -- experimenting with new capabilities that the technology makes possible;
3. Transformation -- fundamentally altering the nature of the organization through these capabilities.

John Naisbitt, in his book *Megatrends*, also describes three stages of technology as it spreads through society.[4]

1. New technology following the line of least resistance;
2. Technology being used to improve previous technology;
3. Discovering new directions or uses for technology.

The table below summarizes these insights about innovation.

Type of Change	Characterized by…
Modernization	Doing what we've always done, but using technology to do more and to increase efficiency.
Innovation	Doing things we've wanted to do, but could not do without the technology.
Transformation	Doing new things that we didn't imagine until technology made it possible

An example in the context of the data library may help clarify this typology. If data files are seen as containers for information, then adding these containers to library collections is much like adding any other new container-format. This activity varies little from traditional, pre-digital library policies and procedures and

[2] Lucier, Richard. "The University as Library." Follett Lecture Series. University of Leeds, 6th June 1996. http://www.ukoln.ac.uk/services/papers/follett/lucier/paper.html

[3] Lynch, Clifford. "From Automation to Transformation: Forty Years of Libraries and Information." *Educause Review*. January/February 2000: 60-68. Lynch attributes these to Richard West and Peter Lyman.

[4] Naisbitt, J. (1982). *Megatrends*. New York: Warner Books.

thus qualifies as *modernization*, that is, doing something libraries have always done (building collections) but doing more of it by incorporating new formats. Furthermore, adding records about the data files to the library OPAC is again a matter of doing more of the same that has always been done, but with a new format. Similarly, providing users with licensed access to online data services is very much like providing licensed access to bibliographic databases and full-text journals.

New uses of data documentation provide an example of *innovation* within reference services. For example, the introduction of the Data Documentation Initiative (DDI) standard[5] applies developments in text mark-up encoding using XML to data documentation. This new standard, in conjunction with XML delivery over the Internet, can enable searching documentation and access to data at the level of individual variables.[6] This is a technological change that is enabling a search and retrieval functionality that data services have wanted for many years. Similarly, service providers have long wished for an easy way to deliver data to the desktop of users. The level of connectivity provided by the Internet is now creating the ability to do this.

In a small way, *transformation* occurs within this paradigm when new uses of data are discovered. For example, a data librarian with more advanced tools than simple catalogs of studies might help a user discover a new use for a data file that was never previously considered. Furthermore, data files may be organized and stored in ways that facilitate combing data from different sources. A simple example of this is the merging of the storage, catalogs, finding aids, and services for social science and demographic data with spatial data. Services and formats of information that were once treated separately by the library may now become part of more integrated services and use.

A larger *transformation* can occur in organizational culture as well as in practices. For example, offering data services in a library introduces “raw” or primary research materials into the wider collection of the library. In this way, a user of data in the library resembles a physicist working with sub-atomic particles in a cyclotron or a chemist conducting experiments in a laboratory. The data user, working with the raw materials of research, is changing the role of the library to incorporate a laboratory function, giving us a new metaphor for the library to add to those mentioned above. The library is being transformed from a well-organized warehouse of facts and research outcomes to a laboratory where the patron engages raw materials to create new knowledge. The library is also taking on more central role in the management of information across its lifecycle rather than thinking of information as having a life-span after which it is discarded.

The table below summarizes these ideas about technological change and areas of data services.

[5] Data Documentation Initiative, http://www.icpsr.umich.edu/DDI

[6] See Chapter 4, “Search Strategies,” for more on searching at the variable level

Type of innovation	Characterized by…	Some Types of service
Modernization	Doing what we've always done, but using technology to do more and to increase efficiency.	Including numeric data and documentation to library collections and catalogs.
Innovation	Doing things we've wanted to do, but could not do without the technology.	Providing detailed search and retrieval access to data at the variable level.
Transformation	Doing new things that we didn't imagine until technology made it possible	Combining and reusing data to create new datasets and new knowledge. Transforming the library from a warehouse into a laboratory. Managing data across its lifecycle.

Using these terms and an understanding of these types of innovation should make it easier during the data services planning process to clarify types of services, define needed and desired services, and even understand how services that are practical and possible today may mesh with and pave the way for other services in the future.

Three Components of a Data Library

The task of determining what are the appropriate levels of data service should be guided by the defining roles of the library: services and collections. Another way of looking at these roles is to look at library data services through three of its functions or component parts: technology, the service-provider function, and the collection. In the early stages of planning a data service, looking at each of these components in isolation can reveal what appear to be attractive shortcuts and cost savings. Unfortunately, such an approach may produce distractions that fail to account for the complementary nature of these components and could result in data services that do not meet the mission of the library.

One example of such a distraction is the attempt to define data services as a *technological problem* that can be solved simply through technology. A second distraction emphasizes the *service provider* while neglecting the value of a collection, while a third distraction focuses solely on the *collection* – providing "content" but neglecting service.

Technology component. Although data services are dependent on technological tools, determining a service model appropriate for data is more than a technological problem. A sound service model combines technical tools with non-technical resources and skills. Technological “solutions” such as software, disk space, network connectivity, the Internet and the Web, do not constitute the service but are tools that can help provide the service. In choosing levels of service, it is essential to view technology as a tool that helps accomplish a goal, but not as an end in itself. The choice of data services should not be driven by technology.

For example, take the simple idea of "putting data on the web." This is not, by itself, a service, though it might be a method of providing a service. If the goal of the service is "to make data files more easily retrievable," then using the tool (the web) might be one way to achieve that goal. Thus, in this example, "the service" is ease-of-access; "the method" is "the web." Whether this is a good or bad way of providing the service depends on a number of other questions. We will examine this kind of question in more detail in a later chapter. In the context of a service model, though, it is important before investigating those details to have a good understanding of your service goals and to be able to differentiate them from technological tactics of meeting those goals. Tactics should serve goals and goals should serve a mission.

Similarly, depositing data into an Institutional Repository does not, by itself, adequately address collection building, preservation, or service. The IR as a technological solution is only as useful and effective as the policies it implements.

Service provider component. Many libraries are redefining their role in the information lifecycle because many information producers and distributors are switching to a business model of providing contractual access to a service (e.g., licensed access to electronic journals from a secure web site) instead of the old model of selling a reusable copy of information (e.g., a subscription to a hard copy journal). This new library role is that of “service provider” in which the library negotiates contracts and manages online access for authorized users to licensed content. A data service perspective that focuses solely on this service-provider role tends to view library service narrowly in terms of contract negotiation with information aggregators. This model misses the importance of providing access and service to a selected and organized collection of information and instead transfers the locus of selection decisions from the library to vendors, distributors, and producers. Librarians have come to recognize this viewpoint as “the big deal” after an article by Kenneth Frazier in which he outlines the problems caused by reliance of libraries on online aggregators of journals.[7]

[7] Frazier, Kenneth. "The Librarians' Dilemma Contemplating the Costs of the 'Big Deal'." *D-Lib Magazine* March 2001 Volume 7 Number 3. Note that libraries are starting to rethink the Big Deal: Howard, J. (2011, July 17). Libraries Abandon Expensive “Big Deal” Subscription Packages to Multiple Journals. *The Chronicle of Higher Education*. Retrieved from https://chronicle.com/article/Libraries-%20Abandon-Expensive/128220/

For a library to rely on data vendors, distributors, producers, and large data archives is not in itself a bad strategy. In fact, including other data distributors as partners will almost certainly be a part of the mix of strategies employed by most data service plans.

This can present new problems, however, if a library outsources the responsibility of selecting and organizing materials it may be more difficult to provide a selected collection of information that adequately matches the needs of the library's designated community. The library will most likely find it needs to mix careful selection of appropriate vendors and distributors as partners in its collection responsibility with other collection strategies. If the library or its funding agency sees the function as no more than a contract negotiator, the perspective of the library is incomplete.

It may present problems for users as well. When data are available only in vendor "silos," users may find it more difficult to locate the right silo for the data they need. Silos can provide a specific context for information. Libraries have traditionally broken such silos and provided a new context of the information selected for a specific community. This is user-centered rather than vendor- or distributor-centered. Recognizing this, the library that outsources selection and acquisition may find it needs to develop new services to help users find the data they need. Users faced with more than one vendor's online interface may find that they need to learn the intricacies of several systems and may find that the particular services they require are provided for some data and not for others. Recognizing this, the library that outsources data-service may find it needs to develop its own data services to help users deal with multiple, sometimes inadequate, online services.

Collection component. Where too much focus on the service-provider component sees only a need for an intermediary and no need for a collection, a preoccupation with the collection leads to the idea of *disintermediation*.

Disintermediation is a term used by those who believe that society now needs fewer "intermediaries." Examples of this idea include the use of ATM machines instead of human tellers and the possible irrelevance of publishers and stockbrokers when readers can deal directly with writers and individuals can trade directly on the stock market.

In the area of information dissemination, we hear pundits describe how the Internet will allow people to get the information they need without having to use cumbersome intermediaries such as newspapers and magazines (and libraries). The same argument may be used to question the need for data services. Why, after all, is there a need for a "service" if data users can get what they need directly from data vendors and distributors? What need is there for service staff if the collection is easily accessible? Others have dealt with the fallacies of disintermediation,[8] but we wish to address specifically why disintermediation is unlikely to provide better -- or even adequate -- service for data.

One reason is that the very conditions that allow less mediation for some transactions actually increase the need for mediation for others. Attempts to save costs (and disintermediation is often attractive to managers who see it as a cost-saving measure) by limiting staff and relying on automation may actually increase the demand for data. The result is an increased need for staff to help users identify data that meet their needs.

In the data services world, this can be illustrated with the idea of the "data vending machine."[9] The concept is for a service that delivers data as quickly and efficiently as a vending machine dispenses a cold drink. This idea has appeal to data users who know what they want and to service providers who long to help a larger number of users access data very quickly. Technologists will be tempted by the challenge of designing a system that does the "vending" of the files. And, as noted above, the disintermediated approach also has an appeal to those seeking to limit costs by minimizing the number of people who are involved in the delivery of data.

Although technology will continue to make it easier for users to access data, technology cannot make the plethora of data choices as simple as selecting a beverage. Vending machines work best when there are few choices and many people who will be satisfied by one of the available choices. It is certainly true that the data vending machine would adequately fit the needs of some users some of the time, but it is doubtful that it would fit the needs of all users all of the time. Data is not an end product the way books or articles are. Data users are more like bakers than fast food consumers. While some data users sometimes need only one particular file, know what it is, and would be happy if they could just grab it, most users of data discover that identifying an appropriate dataset for a research question *is itself a matter of research* and not just a selection from among a few common choices.

There is another problem with the vending machine approach. In our experience, as data files become easier to locate and acquire, people who have never done quantitative analysis become more interested in using data. These new data users are more likely to need help in locating and acquiring data and are less likely to seek a known item, less likely to be familiar with using data documentation, and will often not be familiar with how to use a data file. A data vending machine that does a good job of dispensing known items works very

[8] See, for instance, Phil Agre, "Information Technology in the Political Process" remarks at the Congressional Seminar on "Technology and Social Change" organized by the Consortium of Social Science Associations, June, 1998; and Sakar, Butler, and Steinfield, "Intermediaries and cybermediaries: a continuing role for mediating players in the electronic marketplace," *Journal of Computer Mediated-Communications* 1(3), 1995.

[9] We believe the first one to use this term was our colleague Laura Guy who, for many years was the data librarian at the Data Program and Library Service at the University of Wisconsin.

poorly for this kind of user. The data vending machine approach may be a solution for one limited kind of use (i.e., may solve one particular service goal for one category of user), but it does not solve all service goals for all users.

In the planning stages of a data service, planners may look at the possibility of replacing a human-intermediated service with an automated, disintermediated service. But, as seen in the example of the data vending machine, it may be that such an attempt may be seen more accurately not as replacing a service but as *adding a new service*, and one that may *generate a need for other services* as it attracts new users.

A second problem of disintermediation and focus on collection without adequate attention to service is the problem of quantity. The mere availability of vast quantities of information does not make the job of the user easier. While the proliferation of information sources and the tools that make it easy for users to acquire data files quickly and efficiently without an intermediary are welcome changes, the proliferation of information sources can create new problems for users. Ease-of-access does not necessarily equate to ease-of-use.

We are not the first to call attention to this problem, which is becoming increasingly obvious to users of the World Wide Web. Phillip D. Long, senior strategist for the Academic Computing Enterprise at MIT, wrote,

> *The beauty of the Internet is in the quantity of data that can be found on it. The bane of the Internet is that the vast majority isn't want you want.*[10]

And psychologist Barry Schwartz has written an entire book on the problem of too many choices, particularly for consumers, with the revealing title *The Paradox of Choice: Why More Is Less*.[11]

When a user is faced with multiple sources and multiple formats of what may appear to be the same data, the user's task has not been simplified, but made more complex. This is precisely where the value of a library providing selection and services can make it easier for a user to efficiently locate and obtain the best data for a research question. Providing selection without service or access without selection does not help the user. Providing both is the defining function of the library. We believe that intermediaries will be more necessary than ever as direct access translates increasingly to more choices making professional assistance that much more important. We will revisit this issue in a later chapter when we examine levels of reference service.

In *The Social Life of Information*, John Seely Brown and Paul Duguid note three distinctions between knowledge and information. First, knowledge usually entails

[10] Phillip D. Long, "Infectious Adoption" *Campus Technology* (June 29, 2004) http://campustechnology.com/articles/2004/06/infectious-adoption.aspx

[11] Schwartz, Barry. *The Paradox of Choice: Why More Is Less.* New York : ECCO, 2004

a knower while information can stand alone. Second, knowledge is "hard to pick up and transfer," while information is self contained and can be stored in a database, written down, passed around, etc. Third, knowledge is something we digest rather than hold and it entails understanding. This leads them to several conclusions, including this one:

> *Circulating human knowledge … is not simply a matter of search and retrieval, as some views of knowledge management might have us believe*.[12]

Data files are the information that can be stored and passed around, but data service providers possess knowledge that enables them to help data users locate and acquire the data they need. Information by itself, without knowledgeable staff, is a collection without service.

Conclusions

The mission of the library, the principles of the social science data community, the essential place of service in defining the role of the data library, and the context of technological change provide a framework for designing a successful data service that addresses the needs of users and avoids the pitfalls of simplistic planning. Goals of a service must be driven by user needs of a specific user community. Preservation and Access are not mutually exclusive choices, but complementary functions. Collections and Services are not isolated options but functions that depend on each other. In the next chapters, we will examine how to build a service plan and choose levels of service that match available resources without compromising this essential framework.

[12] *The social life of information* by John Seely Brown and Paul Duguid, Boston : Harvard Business School Press, c2000.

7.12 A Framework for Data Libraries

Data Basics

Roles of data libraries

The previous chapter defined the mission of data libraries as removing barriers to access while ensuring long-term access to and usability of data. To accomplish this mission, data libraries must assume certain roles, pursue specific goals, and define their user communities and set the levels of service their resources can support to meet their goals. This chapter introduces many of the underlying issues that relate to data collections. Many of these issues constrain or otherwise affect what services a data library can provide. An understanding of these underlying issues will allow us, in the following chapters, to better explore the options available when choosing appropriate levels of service.

The Essential Role of a Library

As mentioned in chapter 7, the roles a library must assume in order to fulfill its mission are the traditional ones of selecting, acquiring, and preserving information, providing organization to the collection so that items in the collection can be located, and providing access to and service for the information. Briefly:

> *The role of the library is to Select, Acquire, Organize, and Preserve information, and to provide Access to and Services for that information.*

Although some librarians question these roles in the digital world, we strongly believe, as noted above, that these are the activities that define a library. Any organization that does less than this may provide a useful service, but it is not a library. If an organization fulfills all these roles as its primary mission (not as a mission that is secondary to some other primary mission -- such as making money for stockholders), what would we call it but "a library"?

In the traditional library of paper and ink, books and journals, these roles were well understood and connected to each other in a linear way, each one leading logically to the next in a well defined way. To "select" a book implied acquiring it. Acquiring a book assumed creating organization through cataloging and shelving. Though many libraries have taken a rather passive approach to preservation, the need for active preservation programs has become increasingly apparent and a key part of most large libraries. Although "access" does not mean exactly the same thing in all libraries (some have "open stacks" and others do not, for instance), all traditional libraries allow users to use the materials in their collections in some way. And, libraries have a variety of services (e.g., reference, circulation and interlibrary loan, photocopying, microform readers, catalogs and indexes, user instruction, etc.) that help users locate and use their collections.

But in the digital library and, as we will see, even more so in the data library, these roles overlap, are less well defined, and can be accomplished in many different ways. Rather than each activity leading logically to the next, a variety of choices interrelate in a complex grid of options and opportunities. Each choice creates or limits the available choices for other activities. Thus, "selecting" something does not necessarily imply "acquiring" anything, and even "preservation" can be accomplished without acquiring anything. Yet, while the parts may seem disconnected in the digital age, when we look closer we find that they are more closely intertwined than ever. As Paul Conway, head of the Preservation Department of Yale University Library, pointed out, there is an inherent interconnectedness of access and preservation.

> *In the digital world, the concept of access is transformed from a convenient byproduct of the preservation process to its central motif.*[1]

.

Collections

Collections are so closely tied to services in the data library environment that any discussion of the one must necessarily involve the other. Librarians familiar with collection development may notice parallels between the collection issues for data and those for electronic journals and licensed bibliographic databases, but with some new and interesting variations.

In some areas of collection development, libraries have had to choose between opposing options such as “access vs. ownership”, “just-in-time vs. just-in-case” and “license vs. own.” In many cases the choice must be based on economic affordability rather than service or collection preferences. While data collections face some of the same choices, the principles of sharing social science data sometimes give data librarians additional options and more flexibility. Additionally, “access” to data is a more complex issue than access to, for instance, a journal article or book. Collection decisions for data are therefore quite entwined with service and computing issues.

Stewardship

At its simplest level, collections of data are about stewardship in the sense of caring for, managing, and preserving data for long-term access and use. A data library can acquire and preserve data files, work with the data community to establish partnerships for stewardship, or rely on a membership organization to provide stewardship. In most cases, a data library will use a mix of all these strategies. In Chapter 11, we will outline in detail different levels of collection services that can enable data libraries with very different resources to accomplish stewardship without compromising their data library mission.

Licensing

Licensing access to data is not a new issue at all. Much social science data has always been available through license rather than purchase.

[1] Conway, Paul. "Preservation in the Digital World" Council on Library and Information Resources, Pub62 (March 1996). http://www.clir.org/pubs/reports/conway2/

Many licensing agreements for data allow direct access to raw data. Such agreements allow the data library (or in some cases the individual user) to actually acquire copies of data files, but impose legal restrictions on use of those files. An example is the standard ICPSR terms of use agreement.[2]

Other agreements involve licensing access to a data service rather than to raw data files. We discuss collection issues related to such licensed services below, under "Access vs. Ownership."

The conditions of license agreements vary widely, but include limitation of use of the data to members of the licensing entity, confidentiality clauses that protect the privacy of survey respondents, and retention of ownership of the data by the producer or distributor. Some licenses to academic institutions restrict use of the data to academic, non-profit use. Many licenses request notification of the data producer of publications based on the data.

While some data used by social scientists are proprietary and have very strict licensing conditions, many important and widely used data sets are either freely available or available with limited licensing restrictions.

One license model that provides more benefits than restrictions to data libraries is the data archive membership model. ICPSR uses membership fees to support their archival activities. Members pay fees that grant them access to the contents of the data archive; the archive uses the membership fees to provide archival preservation for the data. The restrictions imposed by membership (e.g., members agree not to re-distribute data to non-members) are balanced by the benefits of membership. These benefits include participation in the selection and retention policies and governance of ICPSR; access to a large collection of data; access to services such as *MyData*; and assurance of the long-term preservation of the data. Members gain all of these benefits for a predictable cost – the membership fee.

Access vs. Ownership

The model for traditional library collections is for the library to purchase, acquire, and store materials physically. While this model is still predominant for books, libraries are increasingly relying on other models for electronic journals, bibliographic databases, and other digital resources. Rather than physically acquire a copy of an electronic journal or a database, a library may license access to the information which remains physically stored elsewhere and owned and controlled by the vendor or publisher. This model challenges the concept of "collection" since the library does not own, control, or even possess a copy of the material. These two different models are often described by the phrase "access

[2] http://icpsr-support.blogspot.com/2009/01/what-are-icpsrs-terms-of-use.html

vs. ownership." The collection issues involved in this apparent dichotomy are a bit more complicated than the simple description implies, however.

Licensing, access, ownership, and possession of a copy of information are not the same, though they often interrelate and affect each other. Some licensing agreements, for instance, allow for the physical transfer of a complete data file, while others allow only for access to the data through a service. Some, but not all, government data that are available through data services on the web have no licensing restrictions. Some government data files are available for downloading (and thus possession and ownership) from the government and also available from commercial vendors through service agreements (without possession or ownership).

As noted above, some data are available through licensing access to a data service rather than to raw data files. Examples include Thomson *DataStream*[3] (which provides client/server software for access to financial data) and the International Monetary Fund's web-based access to *International Financial Statistics*.[4] In these cases, the data library never acquires the raw data files. "Access" is often limited by the functionality of the service itself. For example, the complete dataset may not be available directly; the number of data points, observations, or cases may be limited in each transaction; there may be additional fees for actually downloading data; data may be available in aggregate form or in a format for printing rather than for analyzing with statistical software. While online data service vendors can provide a valuable service when they enhance access to data, the data library should not confuse ease-of-use with stewardship. Commercial data service vendors usually contract for short-term access to data, not for stewardship of data. The data library does has no *control* over the content, management, preservation, or access to the data in most licensing agreements. Neither does the data library have any control over the functional mechanisms (discovery, access, delivery, API, data formats) of a commercial service.

Government agencies that provide online data services have much in common with commercial data services. They emphasize ease of access or ease of use, but some do not provide access to raw data and they do not guarantee stewardship of data. While governments are becoming aware of the need to preserve digital information, we are not convinced that they will do so; there are simply too many things working against their being able to do so. Cost is not the least of these. Governments without sufficient funds to provide basic government functions can hardly be expected to place a high budgetary priority on a low-visibility service that will only serve future generations of researchers. In addition, the private sector is quite vocal in demanding that governments not do those

[3] http://online.thomsonreuters.com/datastream/
[4] http://www.imf.org

things that the private sector might like to do.[5] This makes data stewardship a politically charged issue. The combination of high cost, relatively low demand, and political vulnerability makes us believe that governments will not succeed in providing long-term access to data even if many people inside and outside government believe it should.

Are local collections necessary?

In looking at the issues involved in building a collection of data, one question stands out. Now that data are so abundantly available on the Internet, do data libraries need to collect data? In other words, do data libraries need to have data files in their physical possession or control? Or, in library terminology, does *access* to a virtual collection obviate the need for *ownership* of a local collection?

The temptation is to say that access is sufficient. There are indeed circumstances in which access is a useful strategy – at least for the short term. Technological advances make it increasingly possible for data libraries to share data through online services, thus reducing the need to store data locally. Such developments increase the options available to data libraries, but do not necessarily replace the need for local collections. There are still significant reasons for data libraries to maintain collections of data in addition to relying on remote data sources.

First, when there is no local copy of a dataset, the local data library has no control over the accessibility of the data. Access can be interrupted without notice for short or long periods.

Second, with a local copy, a data library can ensure which edition or version of a dataset it is making available. Datasets on the web are sometimes amended, truncated, updated and otherwise altered without notification or record. While there are some situations (e.g., when data users want the most up-to-date version of a rapidly changing database) where this is an advantage, there are others where it can cause misunderstanding, inaccurate analysis, and worse.

Third, one way of assuring long-term access and usability of data is for a data library to obtain a copy and provide the preservation and documentation necessary.

Users do not to care whether the data they want are stored locally or remotely *as long as they can reliably and quickly get the data they want when they need them*. The data librarian can address this issue by asking if, without a local collection, the data in question will be available when users need them. As the

[5] See, for instance, Stiglitz, Joseph E., Orszag, Peter R., and Orszag, Jonathan M. The Role of Government in a Digital Age. Commissioned by the Computer and Communications Industry Association. Washington, DC: Computer and Communications Industry Association October 2000. http://unpan1.un.org/intradoc/groups/public/documents/apcity/unpan002055.pdf

data library develops a service plan and makes decisions about the levels of service that it will provide, it should address what its role will be in ensuring long term access and usability for the data its users need. There are different ways of doing this and different roles may apply to different data sets and different user communities.

If data are available through an institutional membership in a consortium such as ICPSR, one is confident that the data will be available on a continuous basis, especially when membership in the consortium includes a say in the access policies of the organization. This is similar to the control and guarantees built in to the JSTOR scholarly journal archive.[6] In the case of the ICPSR, the mission of the organization is to archive data and to make data available to members[7].

Reliance on a commercial vendor or a government agency to provide users with access to data can be a great deal riskier, however. For one thing, your institution likely will have little if any voice in ensuring long-term access to the data. Governments and some commercial vendors often favor the most recent data and delete older, historical data. Data may also disappear if the vendor goes out of business, changes its mission, or is bought by another company that no longer supports a particular data product. Government agencies also change missions, alter budgets, and may even face elimination. Official government archives that have an explicit mission to collect data provide good assurance of long-term preservation, but they often lack the explicit mission or budget to ensure adequate access. Agency budgets and missions have been challenged in many countries and in most jurisdictions after the financial collapse of 2008. Political changes have enabled government-minimalists and libertarians who seek to reduce the role of government. Some public-access technologists and private-sector fundamentalists have been pushing for a minimal role for government distribution of data.[8] These issues are not new.[9] All these factors can endanger long-term access to data, and one reason to obtain copies of data is to ensure that decisions about the retention and weeding of data in your collection and the speed of access to data by your users are under your control. (There are situations in which a copy of a data product cannot be obtained for a local collection because of licensing or distribution restrictions; see chapter 15 on Acquisition Strategies for more details.)

A fourth reason to maintain a local collection of data is to support specific applications of the data. A copy of the data may be obtained in order to improve

[6] “Challenges and Opportunities Presented by Archiving in the Electronic Era,” Kevin M. Guthrie. Paper presented at the JSTOR Participants' Meeting, January 16, 2000. http://dx.doi.org/10.1353/pla.2001.0017

[7] ICPSR Mission Statement. http://www.icpsr.umich.edu/icpsrweb/ICPSR/org/mission.jsp

[8] Robinson, D., Yu, H., Zeller, W. P., & Felten, E. W. (2008, May 28). Government Data and the Invisible Hand. *SSRN*. Retrieved August 30, 2008, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1138083

[9] Jim Jacobs, Karrie Peterson. "The Technical is Political," *Of Significance...* 3(1) 2001, p.25-35. Association of Public Data Users. http://3stages.org/jj/w/apdu.html

access for local users. For instance, a large data file available on the Internet may be customized for local users by creating a subset that is in heavy demand. Creating a subset of census data for the geographic areas of key interest to your clientele is an example of this type of value-added service. Making data available locally in a preferred format might be an appropriate service for some data libraries. For example, a data library might download a data file, convert it once to the locally preferred statistical software format, and make the converted file available locally in order to save users time and effort. A data library might also obtain a copy of a locally-popular datafile, create DDI metadata for it, and load it all into a locally hosted web-service using SDA or Nesstar or Dataverse. (Whether or not these activities reflect an appropriate level of service is another matter. We examine levels of service in the following chapters.)

Another example: some Internet statistical services offer nicely formatted tables of aggregate statistics but do not make raw data available for downloading. The data library might obtain a copy of the raw data file from another source so that local users would be able to drill down into the data and do data-mining and other analyses.

For the reasons outlined here, many data libraries offer a mix of locally stored files and access to remote data and services.

Service vs. Collection

It might seem that one can't have a collection without providing service and can't provide a service without having a collection. While this is somewhat true in the paper-and-ink world, it is becoming increasingly possible to divorce these roles and provide one without the other. We do maintain that the two go together and that the best service is based on a collection in control of the library. Nevertheless, there are other options.

As an example, consider again Thomson *DataStream*. A data library could subscribe to this commercial data service and obtain access for the library's users a vast amount of financial data and a service for users to locate, identify, download, and use the data. But the fees that the library pays for this service do not add anything to the library's collection. This particular service does not offer any data to the subscriber when the contract for the service ends. Choosing such a service is definitely choosing to provide access to and service for these data without adding to a long-term collection. This is not necessarily a bad strategy. To use an analogy to a completely different realm, this is like subscribing to cable TV rather than buying videos. Each has its advantages and disadvantages. The point is simply that it is possible to provide a service without a collection.

It is also possible to have a collection and provide little or no service for the collection. This is much more common, even in traditional libraries. Consider the literally thousands of federal government CD-ROMs in many U.S. Federal

Depository Libraries -- many sitting in drawers without any service available for their use except, perhaps, a catalog record.[10]

The data librarian will often be faced with service vs. collection options. Understanding the implications of these options for the mission of the data library is the key to choosing the best level of service.

Access in the Digital Age: The Data Object

In the traditional library, "access" essentially meant physical access to volumes of books and similar physical objects. Today, we often speak of "the digital object" when speaking of the contents of digital libraries. Such "objects" often have analogs in the paper-and-ink world: book-objects, journal-article-objects, image-objects, sound-recording-objects. In many cases (at least at the "modernization" stage of technical change as described in the previous chapter), the object is well understood and well defined. Although digital libraries may advance (in the innovation and transformation stages of technological change) to offering "pieces" of objects (e.g., a chapter from a book, an illustration or table from an article, a movement from a symphony), the relationship of these parts to the whole is still well defined and understood. There is even a terminology for describing digital objects made up of pieces that relate to each other in a defined way: "complex digital objects." In most cases, the parts of a complex digital object obtain meaning from the context of the other digital objects (e.g., “Chapter 2” follows “Chapter 1”).

There are similarities and differences between more traditional digital objects and "data objects." Data files are highly structured and very well defined by codebooks and other documentation and metadata. In that way they are similar to other complex digital objects. On the other hand, while it is probably less often that a user wants, for instance, an illustration from an article without the whole article or a chapter without the whole book, it is often the case that a data user wants a subset of a data file: a few variables, a defined set of cases. There are two reasons for this.

First, data files are the raw materials produced by primary research. They are intended for analysis. This is in stark contrast to most materials that libraries have traditionally collected. Books, journal articles, and so forth are the results of analysis rather than the materials for analysis. The very nature of data and books means that they will be used differently.

Second, while the parts of a book (e.g., chapters) and the parts of journal articles (e.g., paragraphs, illustrations, tables) obtain their meaning in the context of the other parts, a subset of a data file *obtains its meaning from metadata* that

[10] Hernandez, John and Tom Byrnes. "CD-ROM Analysis Projects" Spring 2004 Depository Library Council Meeting, St. Louis, MO April 21, 2004

describe the data file. Thus, most traditional library objects are intended for use as a whole while data objects are intended for use with metadata.

These differences between more traditional digital objects and data objects have implications that we will explore when examining alternative levels of service and acquisition strategies.

Summary

In this chapter we have outlined the collection issues that affect both data collection decisions and service decisions. In the next chapter we will examine the ways that a service plan can address these issues allowing a data library to have an appropriate service and an appropriate collection that both meets the needs of users and fulfills the mission of the data library.

8.10 Roles of Data Libraries

Data Basics

Preservation Basics for Providing Data Services

This chapter provides an overview of the essential functions and terminology associated with the preservation of research data. Digital preservation is a very large and often complex topic and the preservation of data is a sub-topic that contains many unique issues. The contexts in which data occur are volatile and quickly change over time, exacerbating efforts to preserve data. Issues arising include clarifying content (e.g., the quantity and quality of data to be preserved), evolving technologies (e.g., software, storage, networking, infrastructures), competing stakeholders (e.g., publishers, researchers, governments, universities, the public, future users), satisfying regulatory requirements (e.g., data management plans), dealing with economic factors, and more. In a brief chapter, we cannot begin to cover these in any depth and, indeed, any discussion of these kinds of issues are sensitive to today's data environment and rapid changes in technology. So, in this chapter we focus on the functional issues of data preservation and describe some basic, essential, unchanging issues and introduce terminology that should inform any discussion of preservation. To do this we summarize lessons we can learn from the Open Archival Information System (OAIS) standard. We follow that with a brief overview of how repositories can be certified as trustworthy with the *Audit and Certification of Trustworthy Digital Repositories* standard.

Introducing OAIS

The *Reference Model for an Open Archival Information System* (OAIS)[1] was developed to provide a common set of terms and concepts to help us understand and compare organizations that preserve information for access and use. It has been adopted as an ISO standard[2] and is, effectively, *the* standard for digital archives.

What OAIS is -- and is not

OAIS is all about *de*scription, not *pre*scription. OAIS is a "recommended practice" document that describes a "reference model" for organizations with a mandate to preserve information for the long term. A "reference model" is a very specific kind of standard that describes concepts, frameworks, and terminology to help practitioners understand, describe, and compare systems.

[1] Council of the Consultative Committee for Space Data Systems. *Reference Model for an Open Archival Information System (OAIS)*. Magenta Book, issue 2. Washington, D.C.: Consultative Committee for Space Data Systems, 2012. CCSDS Publications 650.0-M-2. http://public.ccsds.org/publications/archive/650x0m2.pdf

[2] International Organization for Standardization. *Open archival information system -- Reference model* (ISO 14721:2003). [141 pp] http://www.iso.org/iso/catalogue_detail.htm?csnumber=24683

OAIS does not prescribe any particular methods of designing, implementing, or managing an archive. The reference model does not specify a design or an implementation. Furthermore, it does not require specific software, hardware, file formats, databases schemas, or metadata standards. Nor does it demand any technological solutions. Although its focus is on digital information, it is so generalizable that it is applicable to non-digital information as well.

OAIS does provide a common terminology, a description of essential functional responsibilities, and an "information model" for preservation. Together, these provide us with a standard that we can use to discuss, design, describe, manage, and evaluate digital repositories of all sizes and kinds.

A library can be an OAIS archive

OAIS is generalizable. Since it is not prescriptive, all kinds of libraries and archives can use OAIS and it is equally applicable both to "libraries" and to "archives" that have long-term preservation of information as part of their mission. It is relevant to large and small institutions and to public, academic, special, and school libraries.

One key purpose of OAIS is to disambiguate terms that are often used by different communities to mean different things. The word "archive" in the title refers to any organization that accepts the responsibilities described in OAIS, including organizations that had not previously thought of themselves as performing an archival function. It is intended to be applicable to all disciplines and organizations that do, or expect to, preserve and provide information in digital form. The OAIS model is:

> "…applicable to organizations with the responsibility of making information available for the Long Term. This includes organizations with other responsibilities, such as processing and distribution in response to programmatic needs. This model is also of interest to those organizations and individuals who create information that may need Long Term Preservation and those that may need to acquire information from such Archives." [1.2][3]

Essentially, OAIS gives everyone who is concerned about long-term preservation of information a functional model, an information model, and a vocabulary for conceiving, discussing, designing, managing, and evaluating preservation activities.

We are not alone

When begin to design a new data service or evaluate an existing service, we have to think of the context within which we operate. It is not possible to design a data service without taking into account the policies, practices, and technologies of

[3] References to OAIS are to section numbers in the "Magenta" edition of 2012.

producers, distributors, software vendors, other data services, other archives, and many other data and technology stakeholders. The realm of possibilities comes from all of these factors. It is a Good Thing that we are not alone in developing data services because we have many partners with whom we can work to develop services beyond our ability to do so alone. But all of these contexts also constrain our options, limiting what we can do and how we can do it with laws, regulations, contracts, technologies, and other barriers.

As we design services for data in our own institutions, OAIS helps us think about our own roles in long-term preservation, the roles of the communities we serve, and the roles of the larger communities within which we operate. We must take into account how all these intersect and interact with each other. The OAIS models and terminology help us focus on the lifecycle of information production, preservation, and use. It enables us to do a better, more complete job of designing our services to respond better to the needs of our user communities.

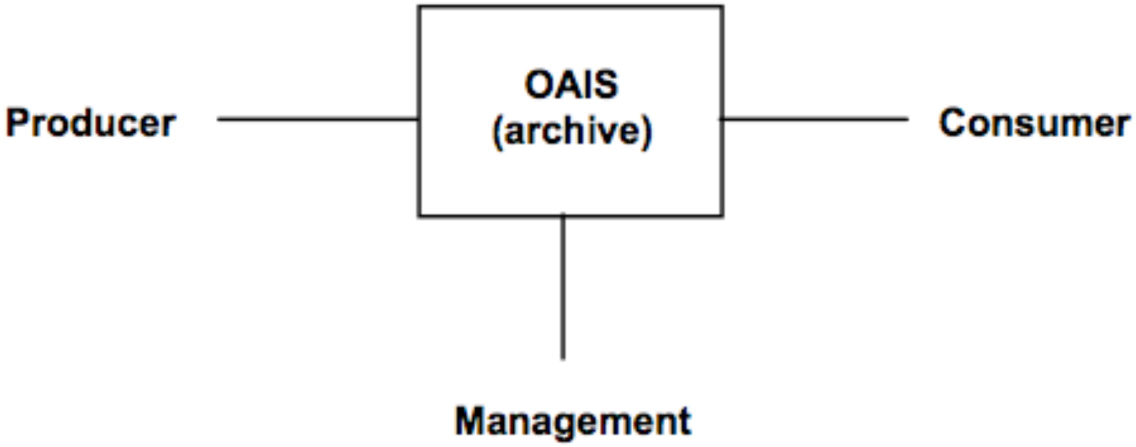
We will examine various options for levels of services and strategies that we can use in designing our data services in later chapters of this book. Our decisions can and should be informed by OAIS.

OAIS does not give answers -- it prompts an archive to ask questions

By focusing on functionality, OAIS helps us design archives that work. It prompts us to ask questions, such as, Who are we doing this for? What information will we preserve? How will we ensure preservation and understandability? Different organizations will answer these questions differently; and their answers will help define their service.

OAIS is actually quite simple!

While OAIS contains a lot of details, its strength is found in its essential simplicity. If you give any two or three experienced data archivists thirty minutes to develop a functional model, they will likely come up with the basics of OAIS. These basics highlight roles within the information lifecycle.



OAIS differentiates the roles of producer and archive. It says, for example, that, although producers may sometimes wish to assume the role of archive for the information it produces, some long-term preservation activities may conflict with the goals of producers, which may focus primarily on rapid production and dissemination.

OAIS notes the need for archive management and preservation planning.

OAIS shows that the archive must take into account the “consumer” – its Designated Community.

OAIS does have a lot of detail and lots of terminology and diagrams and richness but, at heart, it describes very basic, simple, easily understood concepts. This is what makes it applicable to many different archives. By first defining roles, it can then focus on the functions of those roles and the interactions between them.

The Heart and Soul of Preservation: The Designated community

The Concept of the “Designated Community” is fundamental to OAIS. It is mentioned over 75 times and is in every section of the book. OAIS is all about choosing a designated community and meeting the needs of that community.

OAIS defines a Designated Community as “An identified group of potential Consumers who should be able to understand a particular set of information.” [1.7]

Essentially OAIS says that the first question an archive must answer is “for whom am I doing this?” The answer to this question leads to other questions: What information will be selected for the archive? How will the archive ensure that the information is understandable by the community?

A Designated Community can be very small and specialized or very large and general. An archive can have more than one Designated Community and a

Designated Community can be composed of multiple user communities. Many archives will have other archives as one of their Designated Communities.

OAIS is concerned with the Long Term

OAIS archives are concerned about the preservation of information “indefinitely” and “permanently.”

OAIS defines the “Long Term” as “A period of time long enough for there to be concern about the impacts of changing technologies, including support for new media and data formats, and of a changing Designated Community, on the information being held in an OAIS. This period extends into the indefinite future.” [1.7]

OAIS is concerned with the permanence of the information, not the institution. It does not assume that any given institution will be permanent, but does assume that steps will be taken to assure the permanence of the information even if an institution fails. This means that an institution can design OAIS archive to be temporary as long as it is part of a chain of custody that ensures the long-term preservation of information. There are already examples of “staging repositories” whose designated community is another archive that will take over longer term preservation.[4]

In order to preserve digital information for the long term, it is necessary to have some functional criteria to establish how we can know if the information will be usable in the future. This means that OAIS is concerned with more than ‘bit storage.’ It is concerned with guaranteeing access and understandability of the bits preserved. [2]

This means that the information preserved must be:

- Not just preserved, but discoverable.[2.2.2]
- Not just discoverable, but deliverable. [2.3.3]
- Not just deliverable as bits, but readable. [2.2.1]
- Not just readable, but understandable. [2.2.1]
- Not just understandable, but usable. [4.1.1.5]

OAIS does not tell you how to make your information usable but it does give you the context you need to determine if it is. The key here is that the archive role is different from the producer and consumer roles. The role of the archive is one of addressing the functional needs of the information and the community. It is *not* a

[4] Steinhart, G., Dietrich, D., & Green, A. (2009). Establishing Trust in a Chain of Preservation: The TRAC Checklist Applied to a Data Staging Repository (DataStaR). *D-Lib*, *15*(9/10). http://www.dlib.org/dlib/september09/steinhart/09steinhart.html

passive role. It is *not* a role of just accepting what the producer gives without question. Nor is it a role of just delivering any old information package to data users.

Conforming to OAIS.

Although OAIS does not prescribe solutions, it does specify two criteria that an archive must meet to "conform" to OAIS [1.4]. An archive is required to:

1. Conform to the OAIS Information Model
2. Fulfill six OAIS Responsibilities

The Information model = Content + metadata!

It takes almost forty pages to completely describe the OAIS Information Model in detail [2.2 and 4.2], but, in a nutshell, the model boils down to "Content plus metadata." This model will be familiar both to librarians (who are familiar with many kinds of metadata -- particularly descriptive metadata and rights-management metadata -- that point to information packaged as books and journals and journal articles and digital objects and so forth), and to data librarians (who are familiar with data files plus data documentation).

It is beyond the scope of this chapter to describe the OAIS Information Model in any detail. Instead, we present here a quick overview of a few essential OAIS Information Model concepts.

Metadata. OAIS rarely uses the term "metadata." Instead, it describes different kinds of information that an archive must preserve. Some of those kinds of information (descriptive information, rights information, administrative information) are often called metadata in other contexts. OAIS defines four broad categories of this kind of information:

- *Preservation Description Information* (PDI). This includes Reference, Context, Provenance, Fixity, and Access Rights information [4.2.1.4.2] and is essential for managing long-term preservation.

- *Packaging Information* is the information that enables bits to be read from a specific medium [4.2.1.4.3].

- *Descriptive Information* is used for discovery and access [4.2.1.4.4]. This is the same kind of information that is typically recorded in an OPAC or in Dublin Core format as dcelements in DDI 3.

- *Representation Information* is used to convert bit sequences into meaningful information [4.2.1.3]. This is the sort of information that we find in a traditional data "codebook" or, today, in DDI files.

Content. OAIS describes “Content Information” as “the original target of preservation” and defines it as a data object (i.e., bits) plus its associated “Representation Information.”

Information Packages. The OAIS information model describes “packages” of information that include the content and metadata as described above. There are, however, different kinds of packages and the content of the packages for the same information content may vary. The three packages are:

- ***SIP.*** *Submission Information Package.* The package that is sent (submitted) from a Producer to an OAIS. Its form and detailed content are typically negotiated between the Producer and the OAIS. Most SIPs will have some Content Information and some PDI. [2.2.3]

- ***AIP.*** *Archival Information Package*. When an OAIS “ingests” content, one or more SIPs are transformed into one or more AIPs for preservation. The AIP has a complete set of PDI for the associated Content Information. [2.2.3]

- ***DIP***. *Dissemination Information Package*. In response to a request, the OAIS provides all or a part of an AIP to a Consumer in the form of a DIP. The DIP may also include collections of AIPs, and it may or may not have complete PDI.

There is a great deal of detail in OAIS about the Information Model, but even the few simple ideas described above should give us a better insight into some of the everyday decisions that a data service must make.

For example, OAIS does not specify how we accept, or store, or deliver information packages -- it only specifies that they fulfill their functions. OAIS specifies the *function* of these different packages. It makes us think about the functionality first and format or media or file-type only as a means to an end of acceptable functionality.

The Information Model allows us to understand that the archive need not “package” information for preservation or for delivery to the user in the same way as it was received from the producer. Indeed, in many cases we may find that function may dictate that all three packages will be different and optimized for their particular function (production, preservation, use).

It also gives us a common terminology that will help us discuss digital library issues (e.g., acquisition, preservation, service, delivery) with different communities within the library. Information Technology professionals, database administrators, subject bibliographers, reference service providers, business office managers, and library administrators need not have a detailed understanding of social science data documentation or DDI if they all understand the concept – and importance of – Representation Information. And subject specialists in vastly different areas (e.g., art, literature, physics, biology, economics) can speak with each other clearly about the needs of their different Designated Communities and describe those needs – however different they may be – using the same terminology for ingest, preservation, and dissemination.

6 Responsibilities (The Functional Model)

In addition to conforming to the OAIS Information Model, an OAIS-compliant archive must fulfill six responsibilities [3.1]. The six responsibilities are:

1. Negotiates For And Accepts Information
2. Obtains Sufficient Control For Preservation
3. Determines Designated Community
4. Ensures Information Is Independently Understandable
5. Follows Established Preservation Policies And Procedures
6. Makes The Information Available

These responsibilities are also essential to the information lifecycle. All of the six responsibilities must be addressed adequately or information will be lost and the lifecycle will be incomplete.

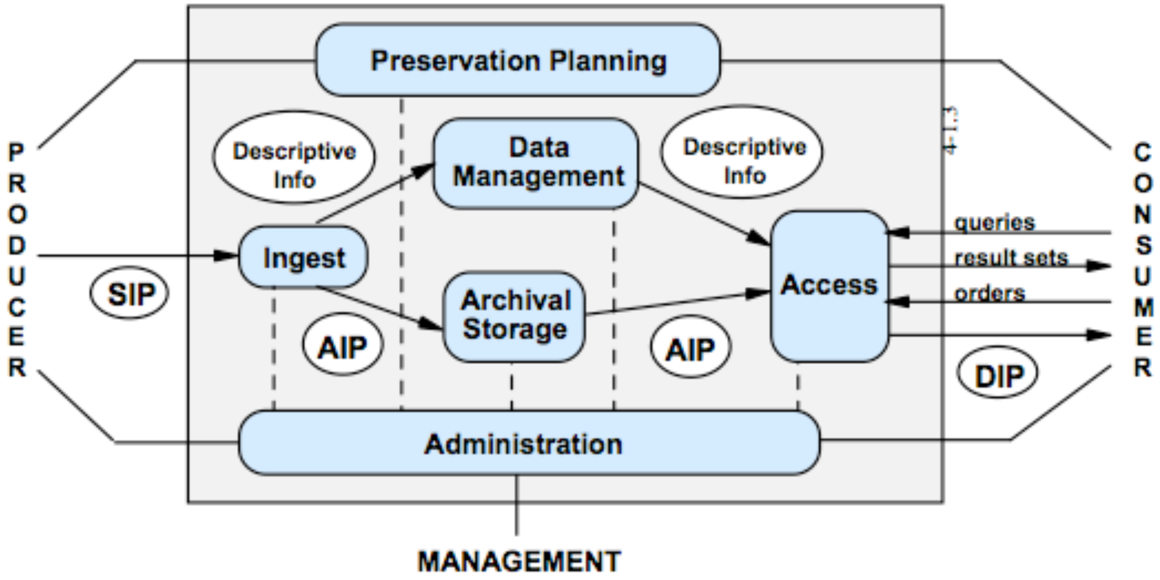
We believe that these mandatory responsibilities parallel the essential, traditional, and continuing roles of a library (i.e., Select, Acquire, Organize, Preserve, and Provide Access To and Service For Information – see Chapter 8). Two of the above requirements do not directly parallel any of the traditional functions of a library: items 3 and 4.

Item 3 (Determines Designated Community) deserves a special mention. In the physical world, it was easy for libraries to think of their designated communities as geographically-based. Physical proximity was a defining characteristic of the "community" in many cases, and in some cases it was the only characteristic. But, in the digital world, the Designated Community need not be physically near the library. This gives libraries a new flexibility in addressing the needs of new communities. Communities can be based on subjects, or disciplines, or type of information, or type of use of information, or almost any other focus.

Item 4, ensuring that the information is "Independently Understandable," is essential in the digital world because "bits" are not understandable alone. We have to preserve the Representation Information that is essential to making the bits understandable.

These two requirements connect us back to the "Long Term" section above in which we noted that information has to be readable, understandable, and usable. These two "new" requirements broaden and focus the traditional functions of a library for the digital age.

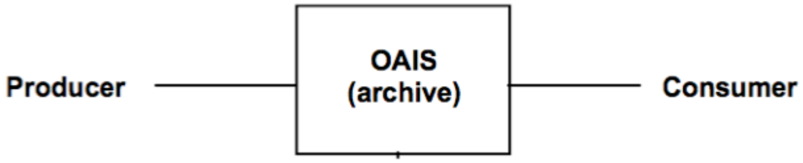
To help archives design systems that adequately meet the requirements, OAIS provides a Functional Model to complement the Information Model.



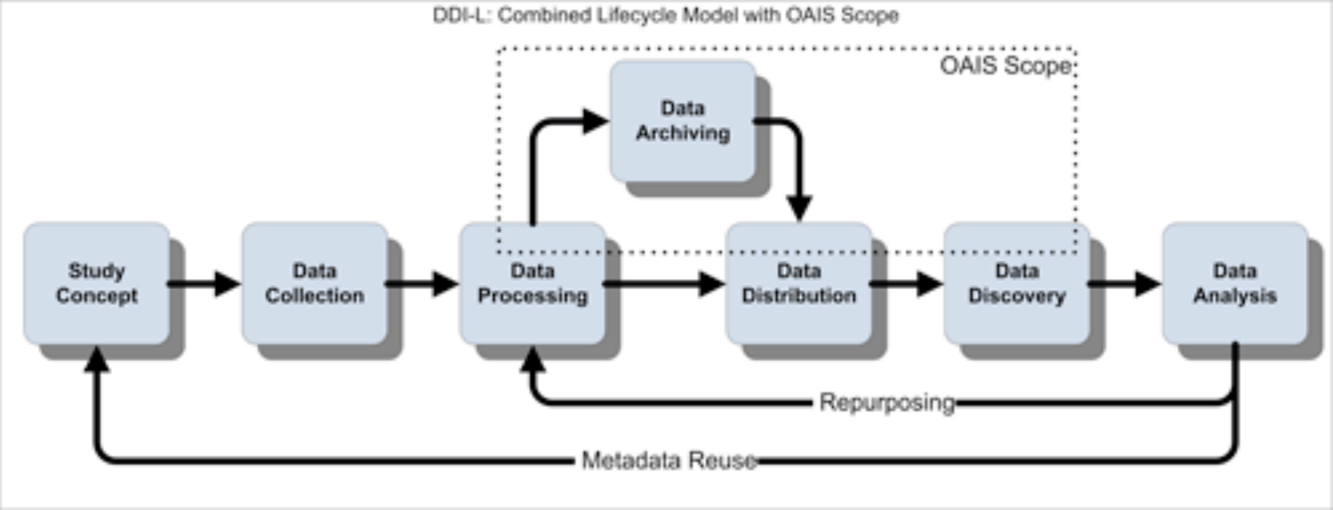
The Functional Model in the illustration above shows the relationships among the “functional entities” of an OAIS and the Information Packages. The Functional Model lists three administrative / management / planning functions in addition to the functions of ingest, storage, and access that we saw above in the Information Model. OAIS describes the Functional Model in some detail, but the key concept here is, again, the simplicity and consistency of OAIS. The information model, the functional model, and the environment are all reflected in these three stages, packages, processes, and functions.

SIP → AIP → DIP

Ingest → Archival Storage → Access



When we design a data service in the context of the lifecycle of research data, we can start by picturing where OAIS overlaps with that lifecycle. The following diagram from the UKDA[5] maps OAIS on top of the lifecycle of research data.



While your design of your data service may not include all of the functions of OAIS, your service design should explicitly recognize these concepts and contexts, define the extent of your role, and acknowledge where these functions will be performed.

Assessing Trustworthiness: Introducing TRAC / TDR

Because the OAIS standard does not tell us *how* to build a preservation archive or prescribe an implementation, we are left with some very practical questions. How do we know if any archive is reliable? How do we know if an archive is OAIS compliant? How can we assure our users that we are preserving their data securely? When we develop a data service that relies on other data libraries, data archives, and data providers, how do we know that the data are being preserved? As you might imagine, these are not new questions. The story begins in 1994.

CPA and RLG set the stage. The Commission on Preservation and Access and the Research Libraries Group (RLG) began exploring these questions in 1994 and issued a report in 1996.[6] That report reached several important conclusions that are still relevant today. These include:

- The need for a distributed system of a sufficient number trusted digital archives.

[5] UK Data Archive. Standards Of Trust / ISO16363. http://www.data-archive.ac.uk/curate/trusted-digital-repositories/standards-of-trust?index=2

[6] Waters, D., & Garrett, J. (1996). *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information commissioned by The Commission on Preservation and Access and The Research Libraries Group* (pub63). Washington, D.C.: Council on Library and Information Resources. Retrieved from http://www.clir.org/pubs/abstract/pub63.html

- The need for a process of certification for digital archives by an independent certifying agency.

The report also recognized that market forces may not adequately respond to a broad "public interest" and may be insufficient to preserve the nation's cultural heritage. Archives, the report said, "have the right and duty to exercise an aggressive rescue function as a fail-safe mechanism to preserve information objects that become endangered because the creator / provider / owner does not accept responsibility for the preservation function."

These were essential revelations: that the digital record of society is going to require a system of trusted repositories that distributes the responsibility of preservation, that trust must be established by a process of independent certification, and that "the market" may not meet the challenge of a broad public interest.

First draft, 2002. RLG and OCLC followed up on the 1996 report by setting up a working group to establish attributes of trusted digital repositories based on OAIS (which was, at that point, still an unpublished, emerging standard). In 2002, the first edition of OAIS was published[7] and RLG and OCLC issued the working group's report.[8] The report enumerated high-level organizational and technical responsibilities and discussed potential models for digital repositories. It paralleled and quoted OAIS and provided a first draft of what certification of digital repositories would become. It said that trusted repositories should comply with OAIS and it went further to specify administrative, organizational, financial, technological, procedural, security accountability. It enumerated "attributes" and "responsibilities" and set out recommendations for further study, but it did not provide specific criteria for assessing and measuring trustworthiness.

TRAC, 2007. A new task force, convened by RLG and the National Archives and Records Administration (NARA), tackled the need for specific criteria and, in 2007, OCLC and The Center for Research Libraries (CRL) released the first audit handbook for digital repositories based on the work of the task force. This document, the *Trustworthy Repositories Audit & Certification: Criteria and Checklist* (TRAC)[9], specifies certification criteria and a process for certification that are

[7] Consultative Committee for Space Data Systems. (2002). *Reference model for an open archival information system (OAIS)*. CCSDS Publications (Blue Book, Issue 1.). Washington D.C.: CCSDS Secretariat, National Aeronautics and Space Administration. Retrieved from http://public.ccsds.org/publications/archive/650x0b1.pdf

[8] Research Libraries Group. (2002). *Trusted Digital Repositories: Attributes and Responsibilities: An RLG-OCLC Report* (p. 70). Mountain View, California: RLG, Inc. Retrieved from http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf

[9] Center for Research Libraries, & OCLC. (2007). *Trustworthy Repositories Audit & Certification (TRAC): Criteria and Checklist Version 1.0* (p. 94). Chicago, IL: Center for Research Libraries. Retrieved from http://www.crl.edu/PDF/trac.pdf

applicable to all kinds of digital repositories including large data archives. (Before TRAC was finalized, it was tested at three repositories, including ICPSR.[10])

TRAC enumerates 84 specific criteria in three broad areas: organizational infrastructure, digital object management, and technology. The criteria are specific without being prescriptive and include examples of the kind of evidence that might be used to demonstrate compliance with the criteria.

Where OAIS provides a framework, TRAC provides an actual checklist of criteria. For example, the OAIS Information Model describes Preservation Description Information (PDI), which includes integrity checks on digital objects in the form of "fixity information" (e.g. checksums). TRAC translates this model into several specific, measurable criteria, mentioning fixity and integrity in five different criteria. For example, it specifies that a trusted repository must have "preservation metadata (i.e., PDI)" to ensure (among other things) that its information "is not corrupted (Fixity)." Another criterion requires the repository to actively monitor the integrity of its content using fixity information such as a checksum and be able to demonstrate that the checksums are stored separately or protected separately from the content they describe.

TDR and ISO 16363. Following the publication of TRAC, the Consultative Committee for Space Data Systems (CCSDS), which had created OAIS, began work on a more formal version of TRAC. This work resulted in the publication of *Audit And Certification Of Trustworthy Digital Repositories* (TDR).[11] The International Organization for Standardization (ISO) has issued TDR as a draft standard.[12,13] A group of experts tested TDR in 2011 by conducting test audits of six archives in Europe and the United States, including the UK Data Archive (UKDA), the Socioeconomic Data and Applications Center (SEDAC) at the Center for Earth Science Information, and the National Space Science Data Center (NSSDC).[14]

Where TRAC has 84 criteria, TDR has 109 criteria that it refers to as "metrics." Metrics were added, removed, and changed from TRAC. The bulk of the changes

[10] Vardigan, M., & Whiteman, C. (2007). ICPSR Meets OAIS: Applying the OAIS Reference Model to the Social Science Archive Context. *Archival Science*, *7*(1), 73–87. doi:http://hdl.handle.net/2027.42/60440

[11] Consultative Committee for Space Data Systems. (2011). *Audit and Certification of Trustworthy Digital Repositories*. CCSDS Publications (Magenta Book). Washington, D.C.: Consultative Committee for Space Data Systems. Retrieved from http://public.ccsds.org/publications/archive/652x0m1.pdf

[12] *Audit and certification of trustworthy digital repositories*. ISO/DIS 16363. http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=56510

[13] The standard is also known as CCSDS 652.0-M-1 and ISO/DIS 16919.

[14] Alliance for Permanent Access to the Records of Science Network. (2012). *Report On Peer Review Of Digital Repositories* (No. APARSEN-REP-D33_1B-01-1_0). Luxembourg. Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_0.pdf

were to the section on digital object management. For example, where TRAC mentions “fixity” in 5 criteria, TDR mentions it in 10 metrics. TDR / ISO-16363 is, however, a refinement of TRAC not a wholesale change.

Self-Assessment and other Assessment Standards

In addition to OAIS, TRAC, and TDR, there are other standards that are relevant to the certification of trusted digital repositories. OAIS specifies a “roadmap” of related standards such as PREMIS[15] for preservation metadata and PAIMAS[16], which specifies an interface between information producers and digital archives. TDR cites standards such as the ISO 9000 family[17] that are complementary or related to TDR.

There are other standards that are very similar in purpose to TDR. The *Data Seal of Approval*,[18] originally created by the Data Archiving and Networked Services (DANS) in the Netherlands, provides 16 guidelines aimed at repositories and at their relationship with producers and consumers. DRAMBORA,[19] developed by The Digital Curation Centre (DCC) and Digital Preservation Europe (DPE), provides tools for an archive to self-assess and manage risks. *Nestor*[20] has 14 broad criteria (and a total of 54 specific criteria) for trusted digital repositories. Germany's DIN 31644[21] defines standardized requirements for the setup and management of digital archives. There are others.

In addition to more formal standards, there are various checklists and tools and lists of questions to help digital archives assess themselves or prepare for a more formal audit. For example, IBM has create the Long-Term Digital Preservation Assessment (LTDPA),[22] which is based on TRAC. The Planning Tool for Trusted Electronic Repositories (PLATTER)[23] is a tool that helps a repository plan and set performance targets. The Minnesota State Archives has written a handbook[24] that

[15] PREMIS Data Dictionary for Preservation Metadata. Version 2.0, PREMIS Editorial Committee, March 2008; http://www.loc.gov/standards/premis/v2/premis-2-0.pdf

[16] *PAIMAS*: [ISO 20652:2006] *Producer-Archive Interface Methodology Abstract Standard*. CCSDS 651.0-M-1. MAGENTA BOOK. May 2004. http://public.ccsds.org/publications/archive/651x0m1.pdf

[17] http://www.iso.org/iso/iso_9000

[18] http://www.datasealofapproval.org

[19] *DRAMBORA*: *The Digital Repository Audit Method Based on Risk Assessment*. Digital Curation Centre and DigitalPreservationEurope (DPE). http://www.repositoryaudit.eu/

[20] NESTOR Working Group on Trusted Repositories Certification. (2006). *Catalogue of Criteria for Trusted Digital Repositories Version 1 (draft for public comment)* (No. 8). nestor-studies (p. 48). Frankfurt. Retrieved from http://www.nbn-resolving.de/?urn:nbn:de:0008-2006060703

[21] Standards Committee on Information and Documentation. *Information und Dokumentation - Kriterien für vertrauenswürdige digitale Langzeitarchive* (2010) http://www.beuth.de/en/standard/din-31644/147058907

[22] https://www.research.ibm.com/haifa/projects/storage/datastores/ltdp.html

[23] DigitalPreservationEurope, (April 2008), “DPE Repository Planning Checklist and Guidance DPE-D3.2" http://www.digitalpreservationeurope.eu/platter.pdf

[24] State Archives Department, Minnesota Historical Society. *Trustworthy Information Systems Handbook* (Version 4, July 2002) http://www.mnhs.org/preserve/records/tis/tableofcontents.html

describes a set of criteria to establish the trustworthiness of government information systems. In 2007, four preservation organizations wrote 10 core criteria for digital preservation repositories.[25] In an excellent overview, ICPSR and Nancy McGovern reprinted a checklist to help institutions assess their readiness to address digital preservation.[26] The UKDA created a useful list of questions for OAIS compliance self-testing.[27]

Europe is working on a Framework For Audit And Certification Of Digital Repositories that would specify three levels of certification. These are: Basic for repositories that obtain DSA certification, Extended for repositories that conduct a successful self-audit based on ISO 16363 or DIN 31644, and Formal for repositories that obtain full external audit and certification based on ISO 16363 or DIN 31644.[28]

[25] http://www.crl.edu/archiving-preservation/digital-archives/metrics-assessing-and-certifying/core-re

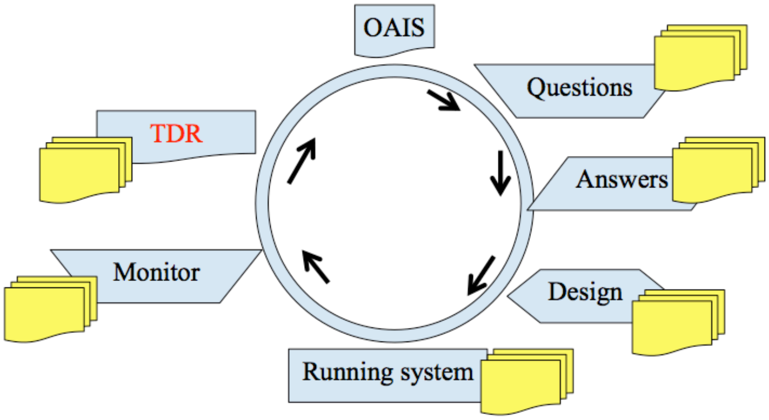
[26] Inter-university Consortium for Political and Social Research (ICPSR), & McGovern, N. Y. (2009). *Principles and Good Practice for Preserving Data* (No. 003). IHSN Working Paper. International Household Survey Network. “Annex E. Survey of Institutional Readiness” Retrieved from http://www.surveynetwork.org/home/download.php?file=IHSN-WP003.pdf

[27] Beedham, H., Missen, J., Palmer, M., & Ruusalepp, R. (2005). *Assessment Of UKDA And TNA Compliance With OAIS And METS Standards* (p. 111). Wivenhoe Park, Colchester, Essex: UK Data Archive. “Appendix 5.” Retrieved from http://www.esds.ac.uk/news/publications/oaismets.pdf

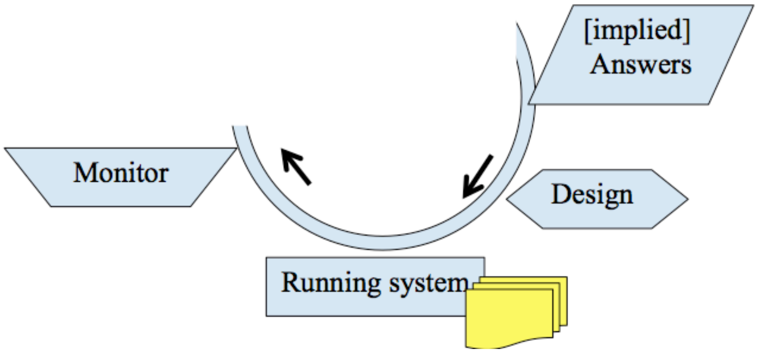
[28] Alliance for Permanent Access to the Records of Science Network. (2012). *Report On Peer Review Of Digital Repositories* (No. APARSEN-REP-D33_1B-01-1_0). Luxembourg. (Annex A) Retrieved from http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2012/04/APARSEN-REP-D33_1B-01-1_0.pdf

Using TRAC for Assessment

The key to using TRAC for assessment is making both OAIS and TRAC an integral part of an archive's planning and development. We can illustrate this (below) as a process that begins with OAIS, proceeds to questions about community and its needs, uses the answers to those questions to design an archive's functions and processes, which result in actual running systems and work-flows that are, in turn, monitored to demonstrate their accuracy and success. The archive documents each stage of this process so that, when it is time for a TRAC audit, everything is ready. The results of the audit provide information useful for continued planning and modification of the archive again relying on OAIS as the keystone for addressing the needs of the community.



This may seem straightforward, but too often archives omit or skip steps and fail to adequately document their decisions, as illustrated below.



TRAC is already being used to assess digital archives. Some organizations have used it to self-assess (e.g., The MetaArchive Cooperative[29]). At the request of its members, the Center for Research Libraries (CRL) has audited digital archives that

[29] http://www.metaarchive.org/sites/default/files/MetaArchive_TRAC_Checklist.pdf

CRL members rely on: Portico,[30] the HathiTrust,[31] and Chronopolis.[32] CRL is committed to using TDR in future audits.

What is it like to be audited? A TRAC audit is not a test in which the auditors have the answer key and the archive hopes to get the “correct” answers. As mentioned above, OAIS essentially defines the questions an archive must ask of itself, but does not prescribe the answers. The auditors expect the archive to have defined its goals, justified those goals and its strategies for meeting the needs of its user community, and designed a system to reach its goals. It then must show evidence that it is meeting its own goals.

The TRAC criteria are flexible enough to apply to a wide variety of types of digital archives. Even for those criteria that are rather specific (e.g., the repository must have “a mission statement that reflects a commitment to the long-term retention of, management of, and access to digital information”), it is left up to the repository how it meets the requirement (e.g., “Mission statement for the repository; mission statement for the organizational context in which the repository sits; legal or legislative mandate; regulatory requirements”).

Many criteria are much less specific and explicitly leave compliance up to the context of the specific repository. For example, a repository must have “the appropriate number of staff to support all functions and services.” It is up to the archive to demonstrate how it knows it has sufficient staff and that staff are successfully supporting the functions and services of the archive.

TRAC does not require a repository to use any specific software, hardware, or file formats. It does not specify how users must be able to search or browse the contents of a repository. It does not specify how a repository will package and deliver information to users. It does, however, require that a repository be able to justify its choices by demonstrating that they adequately meet preservation needs of the information and the information needs of its users. It does require that the repository justify that its discovery mechanisms meet demonstrated needs of the community and that it delivers information in a way that it understandable and usable by the community.

In short, a TRAC audit requires repositories to have put in a lot of work defining, documenting, and justifying its goals and methods. Once that work is done, it is usually a straightforward matter for a successful archive to produce evidence that the it is meeting its own goals.

[30] http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/portico

[31] http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/hathitrust

[32] http://www.crl.edu/archiving-preservation/digital-archives/certification-and-assessment-digital-repositories/chronopolis

The Future of Assessment

Despite the variety of standards and tools and checklists, there are consistent themes for assessing the trustworthiness of a digital archive. The first consistent theme was originally stated back in the 1996 CPA/RLG report: the need for a system of repositories that establish trustworthiness through a standardized, third-party certification process.

The second theme is the recognition that each repository is different and that there is no one-size-fits-all implementation, or set of tools, or method of building a trusted repository.

The third theme is that a long-term repository must be designed with a user community in mind so that it can assure that the information is usable and understandable by that community. Although this allows for all kinds of repositories including temporary ones, it elevates the usability of information to an essential criterion and elevates the user to the ultimate arbiter of usability.

The fourth theme is flexibility. Despite the different approaches and degrees of complexity of the different tools for assessing trustworthiness, they all focus on the essential functionality of the archive as described by OAIS. Some of the tools focus on one or a few aspects of certification and some are intended to help an archive assess itself or plan for an external audit. But all of them aim to help archives assure themselves and their user communities of their trustworthiness in preserving digital information.

What can we expect in the future? There will be more audits using TRAC and TDR and the digital archiving community will learn from these. Those with experience with OAIS and TDR will produce training tools and checklists and procedures for attaining certification. A draft standard[33] for certifying organizations that will provide third-party certification to digital archives is already setting the stage for a formal, international certification process. That process will be finalized and more formal and consistent certifications will become available at predictable costs.

One overall result of all this will be a greater general awareness of OAIS and TDR and a better understanding of the functions and roles of long-term preservation. People and organizations from all parts of the research lifecycle will have a common vocabulary and a common understanding for planning, implementing, and assessing long-term preservation of their data.

[33] Council of the Consultative Committee for Space Data Systems. (2011). *Requirements For Bodies Providing Audit And Certification Of Candidate Trustworthy Digital Repositories*. CCSDS (Magenta Book, Recommended Practice, Issue 1). CCSDS Secretariat. Retrieved from http://public.ccsds.org/publications/archive/652x1m1.pdf

9.18 – Preservation Basics

Data Basics

Developing a Service Plan

A service plan is a statement of who will provide which services. It takes into account the goals of the organization, the needs of its Designated Community, and availability of resources.[1] The development of such a plan is made simpler if it is guided by a service model as described in chapter 7 and includes an understanding of the way collections and services overlap and affect each other in the data library, as described in chapter 8. This chapter will examine the components of a service plan. The following three chapters further explore the making of a service plan through use of the concept of "levels of service" as a means of specifying services, setting limits, and matching resources to user needs.

Benefits of a Service Plan

A service plan provides several benefits.[2] It helps communicate to patrons the types of service that they can expect to receive. It assists in the management of a data service by explicitly charting the landscape of services and providing a rational context for allocating resources. It provides direction for the staff of a data service and helps them maintain service priorities and a commitment to their Designated Community. And it includes a method for systematically assessing services and specifies a scheduled time for reviewing these services.

Setting Limits

As important as it is to specify what services you will provide, it is equally important to set limits and exclude those services that you will not provide. Every data service will find that some services or some communities of users or some types of data will be beyond their technical, financial, or staffing resources. It is best to recognizing this fact formally. An attempt to offer a complete service for all conceivable kinds of users and data would almost certainly over-commit any organization. Providing a formal policy that states which services are offered and which are not is, itself, a service to users. Therefore, an important part of a service plan is deciding which commitments to make and which not to make. This lessens the likelihood of making promises that cannot be kept.

[1] Surprisingly, the literature on service planning in libraries is very sparse. This observation has also been made in Susan Wehmeyer, Dorothy Auchter, and Arnold Hirshon, "Saying what we will do, and doing what we say: implementing a customer service plan", The Journal of Academic Librarianship, vol. 22, May 1996. pp. 173-180. Most of the literature on planning in libraries focuses on strategic planning or the planning of new facilities. One exception is the work Shelia Pantry and Peter Griffiths, *Developing a Successful Service Plan*, London: Library Association Publications, 2000.

[2] A useful presentation about the advantages of a service plan is presented in Wehmeyer, Auchter, and Hirshon op. cit.

Quality and Quantity

When constructing a data service plan, the potential exists to confuse the quality of a service with the extent of a service. The *extent* of service relates to *which* services an organization says it will provide. *Quality* of service has to do with *how well* an organization performs the services that it says it will provide. For example, a data service might provide excellent access to U.S. census data but intentionally not provide consultation services to interpret census results. The presence or absence of a particular service does not in itself establish a service's quality. In other words, the criteria for quality do not depend on the quantity of services offered. Simply put, more levels of data service do not necessarily make a better service. Doing one thing very well would result in a higher quality service than doing a dozen things poorly.

Also, the selection of services that are offered is not necessarily a determinant of the quality of service. Consider an example from the food service industry: an excellent pizza parlor that delivers and an excellent Italian restaurant that does not. Both provide excellent quality but different services. We may think of them as serving different communities as well. Each restaurant provides high-quality food, but one offers an additional service by delivering its products, while the other may offer better dining facilities or a wider menu selection. So it is with different data libraries and the levels of service they offer, as we will see in the following chapters.

When we focus on user communities, it is also important to recognize that any given individual may be part more than one community. A university that provides services to undergraduates and graduate students may define its communities, for example, as users of statistics and users of data. The *type of data-use* is a more accurate indicator for service planning than the academic class of the student.

Components of a Service Plan

A service plan document could take many forms or even have different incarnations or parts. It might, for instance, be useful to have one version or component that addresses users, another for staff, a third for administration and budgeting. This chapter doesn't recommend any particular format or method of presentation of a plan, but instead focuses on the process of making decisions and the kind of decisions that can be documented in a service plan. While the level of detail of a plan could be general or specific or vary between different versions of the document, at minimum a service plan for data services should include:

- Which services will be provided
- The units or departments that are responsible for the services
- The users for whom the service is designed

A service plan could contain much more information for each service. For example:

- Duties, skills, and responsibilities of staff or departments
- Relationship and responsibilities of different service points (e.g., reference desk, computing lab)
- Quantitative limits to services (e.g., storage space, staff time)
- Costs or budgetary information
- Users for whom service is available
- When or if fees are charged
- Methods of assessing quality of services

A complete service plan would, ideally, address services in relation to collections as well, but the specifics of collections (selection, acquisition, organization, preservation) might be better detailed in a separate collection development policy document.

The actual content of any particular plan, or particular document, will depend largely on local requirements for such documents, the audience or audiences for the document, and the level of detail required for your local plan to be effective. For example, a small, college library creating its first data service without any partners, might have a very short, very general document; while a large university creating partnerships between the library, computing center, and a statistical consulting office might require a more formal memorandum of understanding supplemented with a document directed to users to help them navigate service points.

Striking a Balance: Commitment and Flexibility

While the purpose of a service plan is to specify commitments, it is equally important that it avoids being inflexible. A service plan needs to be flexible enough to evolve over time and dynamic enough to deal with conditions that may change rapidly.

Change is inherent in data services and many, if not most, of the changes that affect a service are not within the control of the service. Technological changes (hardware, software, operating systems), administrative changes (computer security enforcement, data license and membership options), academic changes (the introduction of new programs and new focuses of old programs), data changes (file formats, file sizes, documentation formats), and so forth, all create practical issues that the data library must deal with as they occur. A service plan should not be so specific as to be inflexible. A plan that requires a lot of changes to adapt to a changing environment or a plan that cannot be changed or amended without a lengthy administrative process may become unviable quickly.

The method of achieving a balance between specificity and flexibility can be best addressed within the administrative context of the specific organization.

An environmental scan

It is useful to begin the design of a service plan by assessing four key areas that may affect the provision of data services.

- Designated Communities and the Needs of Users
- Organizational structure
- Political issues
- Technical infrastructure

Naturally, this assessment should be done for the local environment such as a university within which the service will operate. The plan should explicitly recognize the larger context of data services and collections available nationally and internationally and how it fits into that context. Knowing the variety of data services available from commercial and non-commercial organizations provides a context for development of local services and provides ideas about the types of services that might be offered locally. An examination of the global context of data services will also provide an opportunity to meet, talk, and discuss service issues with other data professionals.

*Who is the **user community**? What are the **users' needs**?*

A data service should be rooted in users' needs. An assessment of user needs can provide a wish list of services and help set priorities for services. Asking data users what services they want brings users to the foreground of planning and opens a communication channel between them and service providers. A service will also benefit from the goodwill gained by service providers listening to data users and potential users. Getting feedback from users is particularly useful before hiring a new person in data services. Feedback from users can help determine which set of skills best complements existing staff skills and will help position the service to provide the services your users most desire.

Knowing what users want is not enough, though. Users may lack the knowledge of what is technically feasible or be unaware of the possibilities of a service. Users may ask for more of what they have rather than for new services. If users are the only source of service ideas, the resulting service plan could be unnecessarily limited. In surveying users it is important to remember that what is already available may influence what users want next and what is not available yet may limit what users ask for.

It is important for the data library to provide leadership in setting priorities and designing new services. Service providers need to keep informed of developments in the field and be able to imagine what will be coming in the near future. A well-designed assessment instrument can function as an educational tool by asking users and potential users of their interest in specific new services that are technically and financially feasible.

In surveying users, it may be useful to think of users and data librarians as equal partners in the planning process. Each brings different experiences, knowledge, and expertise. A planning process should not be simply reactive -- putting librarians in charge of implementing decisions made by users. Rather, the planning process should be a cooperative venture that brings data librarians and users and their different strengths together to develop a plan that neither might envision on their own.

Rather than assuming that users have all the information needed to plan a successful data service, assume that users have unique knowledge of their subject domain, research techniques, some but not all data sources, and an intimate, but perhaps not explicit, knowledge of their own technical abilities and limitations. Data librarians should bring to the planning process a more extensive knowledge of data sources, metadata capabilities, preservation and sharing opportunities, software and hardware development, and technical innovations that might enhance service delivery.

In contrast to not asking for enough, users may also ask for more than is currently possible. Available technology, staffing, and funding always set limits on what is feasible in any particular environment at any particular point in time. It is important that the assessment process encourages creative ideas from users without raising their expectations unrealistically.

A variety of methods exist to identify the needs of users including formal surveys,[3] informal surveys via e-mail, focus groups, and other consultative approaches. Questions can be very structured, asking about specific data needs or preferred software, for instance, or they can be open-ended, asking about generally desired data services.

[3] See Gaetan Drolet, “Starting a Data Library Service: Where are the Users?", *IASSIST Quarterly*, Vol. 17, Nos. 3&4, Fall/Winter 1993, pp.49-51; and “Conducting a Data Interview” by Michael Witt and Jake R. Carlson (December 2007) http://docs.lib.purdue.edu/lib_research/81/ and “Investigating Data Curation Profiles Across Multiple Research Discipline” datacurationprofiles.org

A key source of information about user-needs is the data library staff that works directly with users. They will have experience with user interests, needs, and abilities. Their experiences of user requests will provide a useful perspective that may be different from that of individual users. They may also help identify potential data users or users who are just starting to use data. For example, in a university library that is just starting a data service, planners may find that reference librarians will have experience with users who ask for or need data. It is also useful to seek the views of administrators who may play a key role in funding or are in a position to facilitate or block collaborative arrangements. These methods work equally well for establishing a new service or for taking stock of an existing service.

It may be useful to conduct a survey of a random sample of your community's population in order to identify potential data users and data users that do not interact with staff.

The users' environment changes continually and this creates a challenge for the data library to stay current with its users. For example, users constantly encounter new computing platforms and different removable media. Upgrades to software may result in new formats for reading and storing data. You may encounter data users from different disciplines with different quantitative data needs in instruction or research. Periodic assessments of users' needs should take into account shifts in technology, user skills, and the instruction and research environment.

Being responsive to changing needs is essential for continued, effective service. In times of fiscal exigency, having well established support from users may mean the difference between keeping doors open or closing a service. In times of fiscal abundance, it may mean receiving funding for an extra computer or an expensive data purchase. The best line of support for the operation of a data service is its community of users.

*What features of the **organizational structure** impact data services?*

A first step in understanding organizational structure is to identify the units that provide data services and the units that might contribute to services. This will be helpful in identifying potential partnerships and in understanding how the organizational structure may facilitate or hinder in the provision of data services.

Each organization will present a unique set of opportunities and potential roadblocks. The key to developing a successful data service plan is to anticipate problem areas and to make use of opportunities. One simple exercise is to answer a few questions about the relationship of units within the overall organization. Do units report to different individuals? For example, what are the ramifications for cooperation if the library reports to a provost and the computing center reports to the vice-president of administration? What happens if a data service is part of a research institute and wishes to collaborate with the traditional library? Within the library, if collections staff in one unit purchases data, are technical services staff in another unit prepared to catalog them? Are there well-established informal working relationships within the organization that mitigate a less than ideal formal structure? What risks are involved in basing services on informal relationships? How can the organizational structure be strengthened to assure stability of service? Is there a way to coordinate activities among organizationally disparate service units?

The long-term development of a data service depends on the stability of its location within the institution and on the permanence of the relationships among data services and partnering units in an organization.

*What are the **political issues** in the environment?*

When planning a new service, knowing about the existence of local political issues is much better than stumbling blindly into a delicate situation. Whether it is some aspect of campus politics, strongly held views of turf by specific service units, outright personality conflicts between key individuals, or a matter of different service cultures, being well informed and anticipating conflict is the best defense against unnecessary obstacles.

Librarians may need to gain the respect of researchers who are used to getting quantitative support elsewhere. Units with a poor service reputation may have difficulty turning around their image. Information Technology staff and staff that provide data reference services may have very different cultures and assumptions about service and these differences may make communicating effectively difficult. Individuals in powerful positions who are uninterested in collaborating with other units may be roadblocks to a service configuration that makes the most sense from an economic and user perspective. Some obstacles may not be immediately changeable and alternative strategies for service may need to be created.

Identifying key players and allies and building coalitions can change strongly held opinions. While seeking the input of users and staff and administrators is essential in developing and assessing a service plan, it is also a prudent and effective strategy to use during a time of major change.

Demonstrating that data services staff are knowledgeable and well informed can increase the respect of those least inclined to give these staff the benefit of the doubt. Particularly in larger organizations that have separate, specialized departments, it is important for the data services specialist to understand the technologies, procedures, and vocabulary of those departments on which data services will depend. Data services staff attempting to establish or expand a service would be well advised to learn as much as possible about the culture and assumptions of those other departments and their own strengths and weaknesses and the frustrations and obstacles that they face. Empathy and understanding goes a long way in negotiations and being knowledgeable and fluent in terminology of another specialized department can facilitate trust and cooperation.

These types of political issues should be factored into assessments, whether they can be easily changed or not. They can serve as real boundaries for the type of service that can be provided. Once identified, strategies can be more easily developed to change or cope with the political roadblocks in an organization.

*What **technical infrastructure** exists in the organization?*

The local technical environment of the organization also needs to be factored into a service plan. This includes hardware, software, operating systems, and networks. It includes the infrastructure that the organization uses such as Content Management Systems, Institutional Repository software, computer security, and digital preservation procedures. Any aspect that is not under the control of data services may present a constraint on service or may have an impact on how the service can be delivered.

Larger organizations tend to rely on multiple units or departments to provide the range of services needed to support access to data. Fiscal realities, historical budgeting and staffing patterns, and existing computer environments will all contribute to the ways in which these units are configured.

One goal of a data service is to ensure that users get the data they need quickly and efficiently. The existing computing and network environment will affect how this can be done and how well it is done. The preferences of users for particular computing platforms and statistical software packages will also affect how easily data can be delivered. The success of a data service will depend in part on the technical environment in which it is located, and in part on the ability to work within these boundaries or to effect change to accommodate new distribution methods of data delivery.

There is no simple rule that defines what constitutes a good computing environment. Any given aspect of a computing environment may present an opportunity for enhancing services or serve as an obstacle. For example, having a site license for a good statistical package at an institution may provide an opportunity for using a well-supported, inexpensive, readily available package as a standard for data exchange. But it could also create a technical constraint if many users choose other packages that are not well supported. It could create a service need (e.g., converting data from outside vendors into the locally preferred software format). It could create a preservation problem (e.g., converting data from the locally preferred software format into a format that better ensures long-term persistence and usability).

Analyzing the technical environment within which a data service must operate is an essential step to identifying what services can be easily provided, what services are needed, what existing resources will support data services, and what additional resources will be required.

Assessment

As noted above, a service plan should address issues of quality separately from the issue of which services will be offered. It should be based on the definition of the service's Designated Community, the needs of those users, and the extent to which those needs have been met. It should, ideally, contain a description of the assessment methods to be employed in measuring an organization's performance of service delivery. The best service plan will include both an inventory of services to be assessed and the method of assessing those services. The plan may include the measurement tools to be used in the evaluation process. The plan should also indicate the time intervals when services will be assessed.

The periods during and immediately following an assessment are also an opportunity to review the mix of services being offered. The results of an evaluation might reveal that the current configuration of services has an impact on the quality of some services. For example, resources may not be sufficient to provide the quality desired for all services. This may result in realigning the service plan to correspond better with resources or reassessing the allocation of resources.

The evaluation should not only consider the performance of individual services, but also changes in overall service offerings. Available resources may have changed between evaluation periods and that may have, in turn, made possible the addition of new services or the loss or consolidation of services. User needs may have changed and created a demand for a new service. For example, the growth of GIS users on a campus might warrant the addition of spatial data to the collection of a data service or indicate an opportunity for a new partnership.

Assessing quality can thus provide a useful measure of how the current service is performing as well as identify changes in the environment that indicate the need for changes to the service plan or opportunities for improving service.

Summary

Developing a service plan for data is potentially more complex that developing other kinds of service plans because of the complex nature of its services and because the necessary resources may be controlled by many different organizational units with different -- and even competing -- missions. Although developing a service plan may seem difficult, time consuming, and frustrating in light of the complexities and potential problems outlined above, these very complexities are what make writing a service plan important. Looking for potential problems and facing them early, enlisting the support and cooperation of different stakeholders, and developing realistic and attainable goals will increase the reliability and sustainability of a service. This can be a very rewarding experience.

A service plan can also help prepare a more realistic budget and help with writing appropriate position descriptions. It can be a foundation for data services promotional literature. It can focus service efforts in an organized and systematic way.

A service plan can be elaborate and detailed, specifying a level of service for many areas, or it can be simple and general, outlining the broad categories of service. Too much detail can make a plan inflexible. It may be easiest to write a first plan broadly and simply and even informally. The plan could be formalized after a trial period and details could be added as necessary and helpful.

Ultimately, a written plan can guide a new service or rejuvenate an existing one. In the case of a plan developed jointly with two or more units, it serves as a contract or set of understandings between the units and specifies commitments of each unit. It can establish boundaries as well as responsibilities. This can minimize confusion and establish expectations for staff as well as patrons. It is an important tool for communicating with users, staff, and management.

There is no template for the ideal service plan. The process of developing the plan and actual document or documents that result from the process will likely differ greatly between institutions. Local conditions, requirements, constraints, and the availability of resources will dictate what a particular service plan needs to address and how best to address it. Nevertheless, every service plan will go through similar stages of development and consider similar issues. This chapter has attempted to give an overview of those issues. The next three chapters expand on this by examining the process of deciding which levels of service are appropriate.

Levels of Collection Services

Collection services encompass the traditional and newer collection building roles of a library. The traditional roles are to select, acquire, organize, and preserve, and to provide access to and service for the contents of the collection. The newer roles include such things as licensing access to information services, and providing access by managing systems that authorize specific computer to view the licensed content. In this chapter we look at how these options interrelate and some specific levels of collection services. In this chapter we begin to apply our understanding of the OAIS Functional Model and the library's role in the lifecycle of research data.

A Matrix of Collection Solutions

In chapter 8, we reviewed the environment for collection building in digital libraries. This includes constraints imposed by information distributors and publishers, as well as service choices of libraries. The key to understanding collection services in data libraries is realizing that every study, every dataset, every selection and acquisition, has the potential to present a unique set of circumstances that requires a unique solution. In practice, there will be solutions that fit groups of studies or categories of distribution. A collection service needs to be flexible enough to deal with different kinds of solutions. The service plan and collection development policy statement need to be written with such flexibility in mind.

The local institutional environment may constrain what the data library can do, but by developing a contextual understanding of the options, the data librarian can make the most effective choices available. The data library should seek to avoid unnecessarily limiting its choices or unnecessarily constraining its services just as it avoids over-commitment of resources or promising services it cannot deliver. Choosing appropriate levels of collection services is a matter of balance and of being creative rather than rigid in deciding which mix of activities best meets broad goals.

The table on the following page presents a matrix of some collection service alternatives. It is illustrative, not all-inclusive. It suggests different ways of dealing with seven different collection situations. Each method is assigned a "level of service" from low to high based on the responsibility taken for each of six possible *activities*.

Data libraries may find it useful before facing collection choices to develop a service plan and collection development policy statement that incorporate these activities and specify the library's preference with regards to each. Then, when the data library faces collection choices, there will be existing guidelines to help choose among alternatives.

Matrix of Some Collection and Collection Service Alternatives

Level of service	Select	Acquire Access	Acquire data	Organize	Preserve	Service
1 (low)						x
2	x			(x)		
3	x	x				
4	x	x		x		
5	x	x		x		x
6	x	x		x	(x)	x
7 (high)	x	x	x	x	x	x

1. Provide service help for users who have found data on the Internet.
2. Select data on the Internet and provide minimum organization of your selections by presenting them on a web page, with minimum cataloging, or by relying on the organizational and finding aids providing by others (e.g., Nesstar tools, Google, etc).
3. Select commercial data services and provide access to your users through subscription, client/server software, IP recognition of their machines by service, etc.
4. Same as 3, but provide additional organization of the services you've selected such as OPAC records, web pages, special finding aids.
5. Same as 4, but add reference service to help users locate needed data, use the commercial service, and download data.
6. Select and preserve data through a membership organization such as ICPSR. Provide access through ICPSR Direct or similar facilities. Organize by adding records for ICPSR studies to your OPAC. Provide service of helping users locate, download, and use data.
7. Build a local collection of data by selecting and acquiring data. Organize by adding records to OPAC and providing specialized search services built on DDI variable level information. Provide online access through web technologies such as Nesstar. Provide reference help in identifying, locating, using data. Provide data discovery and subsetting tools.

Collection Services – Select

Service: Selecting data includes identifying the needs of users, locating data that meet their needs, choosing among different data vendors and distributors, choosing among different delivery methods, and choosing the most appropriate or current or accurate version or edition of the data. Selecting data saves the user time and allows the user to spend more time analyzing data and less time hunting for data.

Skills: Many of the skills required for selection of data are similar or identical to the skills that most subject specialists or bibliographers in a library already have. In addition, familiarity with social science methodology and the collection of data is very important. The activities most likely to require some new training are those involved in choosing formats for data access and delivery. Familiarity with data and statistical software as well as file transfers and storage media will be necessary.

Activities: Selecting data differs from selecting books and journals in a traditional library in several ways. There are no lists of “data in print” or data publishers. Some data files are available from multiple sources. Selection is a combination of identifying data and identifying options for acquiring data. It is of course necessary to provide users a way of knowing what you have selected. This can be a simple as a web page or Content Management System or even creating a paper handout. Other options are described below under "organization." One key component of selecting data in most data libraries is managing a membership or other arrangement with a data archive or other data distributor. In the United States, maintaining a membership in ICPSR is one way for a data library to select a large body of essential social science data for its users. Memberships are described in more detail below under "Acquiring Access."

Collection Services – Acquiring Access

“Acquiring” data can include physically acquiring files as well as acquiring access to remotely stored data, but we deal with these separately because they require very different skills, activities, resources, and commitment.

Service: Acquiring data goes one step beyond selecting data for, while selecting data makes it possible for users to identify and locate data, it does not necessarily make it easier for users to get a copy of the data files and documentation you have selected. While many datasets and data services are freely available on the Internet and therefore do not require any special activity by the data library, many important data sources require membership or licensing or subscriptions for access. The data library can provide an important service for users by negotiating and paying for such access.

Acquiring access to remotely stored data is usually accomplished either through establishing a membership with a data distributor or archive such as ICPSR, or through a licensing agreement with a data vendor. In both cases, the data library takes on a proxy role for the data users by identifying, comparing, and choosing a source of data, negotiating license agreements and contracts, establishing the relationship with the vendor, and paying fees and maintaining the agreements. These are functions that are very similar to those that many libraries provide for subscriptions to electronic journals and indexing and abstracting databases.

Skills: Many of the skills required acquiring access to data will already be available in most libraries. Even some of the more data-specific activities, such as ensuring license agreements are acceptable and enforceable are common in libraries that have subscribed to electronic journals or negotiated access to bibliographic databases. Defining what “acceptable” access is for data does require knowledge, not just of content, but also of how data are used. Understanding the difference between “statistical information” and data suitable for analysis is essential. The complexity of the package of materials that comprise a data order may differ substantially from other types of acquisitions. Understanding the use of statistical software and how data files and digital documentation are managed are, therefore, essential.

Activities: Again, many traditional library activities are appropriate for data acquisitions. In the case of licensed access to remote data collections or services, it will be necessary to monitor the service to ensure it is working and, perhaps, transfer to the service provider tables of computer addresses (IP numbers) that are authorized to use the service or install client software. For PC-based data products, installation and maintenance of software will be necessary.

Collection Services – Building a collection

Services: Acquiring access to data is one increasingly important part of the mix of services that data libraries provide, but it is a complement to, not a substitute for building a local collection of data. Some data libraries may be able to do one or the other; many will need to do both.

There are several reasons a data library may choose to add data to a local collection.

1. *To acquire those data that are needed locally but that are not available remotely*. There are still data that are not accessible online and the only option for local access is local acquisition.

2. *To ensure long-term preservation or access or both*. Often, data are conveniently available on the web, but the organization that provides access to the data does not ensure permanent preservation of or

access to the data. In such a case, adding the data to a local collection relocates control of permanent access to the local data library.

3. *To ensure or enhance short-term access*. This is a service that provides convenience to users where convenience is not ensured by remote access. Some data available remotely may become inaccessible from time to time, or access may be slow or otherwise inconvenient. Adding a copy of the data to a local collection ensures that the data library is in control of the availability.

4. *To provide subsets to meet local needs*. When local users frequently use part of a very large dataset, they may find it more convenient to access a copy of just the subset they need rather than search for, find, and, in some cases, create the subset repeatedly. For example, if the data library finds a most of its users require local census data and not data for all states and localities, it can enhance service for users by providing a local copy of just those files that match users needs and making those files easy to download or otherwise access.

5. *To provide a data product that makes data, which are difficult-to-use, easy-to-use*. Sometimes, just having access to large complex data files is not adequate for local use. For example, the U.S. *Current Population Survey* raw data files are available to ICPSR members and freely available from the Bureau of Labor Statistics, but these files are large, complex and difficult to work with, especially for multiple year analyses. *CPS Utilities,*[1] a commercial product that packages many years of the U.S. CPS into a common environment with special software, makes it much easier to do research across years.

Each of these services is slightly different and brings its own requirements for reference and computing services, which we discuss in other chapters.

Skills: Many of the skills needed for building a data collection are similar to the skills needed for building any library collection. An existing library infrastructure consisting of subject specialists, bibliographers, and an acquisitions department can provide these basic skills.

In addition, an understanding of quantitative research methodologies and data collection, preparation, and distribution practices is necessary in order to understand the needs of users and the options available from data distributors. Data orders often consist of multiple files, documentation (print and electronic), and occasionally special software. The data acquisition librarian needs to be able to understand what these are and, where alternative formats and means of distribution are available, which best fit the collection and service parameters of the library and the needs of users.

[1] http://www.unicon.com/, Unicon Research Corporation, Santa Monica, California.

Data that are bundled with software will require staff skilled at installing and configuring the software and resolving software conflicts on public access machines. Depending on your local environment, this job may be a responsibility of an I.T. department. Management of public service computing facilities may be subject to existing service policies that may not be friendly to such software.

For long-term, permanent collections of data, the data library will require skills in metadata creation and maintenance, and computer file preservation, migration, and conversion.

Activities: There should be procedures in placed to help users identify which data meet their needs. Users may request data by subject rather than title, or a requested study may be available from different distributors with different utility. It may be necessary to identify alternate sources for the data in order to select the best and most appropriate source. It is necessary to compare the formats, costs, services, and restrictions on use of each different source. Licensing restrictions must be examined to see if the data library is equipped to meet the conditions of use imposed on the data. Trained staff can ensure that data arrive in a format that is usable by the local community of data users.

Some data come with licensing restrictions that are so specific that a library may want or need to require each individual user to sign a licensing agreement or otherwise manage access and permission to use the data. As always, it is essential to understand the user community and its needs. For example, is the data needed and used by a few individuals, or is it needed for a university class?

Collection Services – Organize

Service: Organizing the data collection makes it easy for the data user to quickly locate the data you have selected. The more data you have selected, the more important it is to provide useful organization and finding aids for users. A wide range of access strategies are possible and the choice of a strategy for organization may depend on the how the item is selected and if it is “acquired.”

All authorized data users should be able to learn easily what data files are locally available and how to get a copy of or access to using these files. Available data files must be cataloged, described and listed. Access to a study requires access to documentation and often to multiple files and sometimes to software for using the data.

Note that a library can take an active role in providing access to data by adding catalogue records to its OPAC[2] for data that are available to its local community of users. Keeping in mind the access issues discussed in chapter 17, catalogue records for locally held data can provide valuable information to potential users of data. This is true even if the data are not in the library, but are in, for instance, a computing center, or a survey research center, or an academic department. Providing catalogue records for data that are accessible on the Internet adds selection and order to those data. In our experiences, the mere fact of adding records about data to a library OPAC increases the use of these data.

Skills: Most of the skills required for these activities are already present in today's library. Consortia that share cataloging already have many data titles in their databases of catalogue records. Cataloguing rules[3] exist for data and their documentation. The increasing experience and familiarity that catalogers have working with records of other types of computer files should ease the inclusion of data files in the OPAC.

Building specialized indexes and databases requires technical skills of creating and managing such services, an understanding of data and data documentation, and enough knowledge of the subject content and how data are used to provide a useful utility.

Activities: Information in a format suitable for loading into library OPACs is available from cataloging consortia such as OCLC, from data archives such as ICPSR, and from commercial vendors such as MARCIVE. In addition, ICPSR makes available to members its Metadata Records[4] that describe its holdings. Some libraries incorporate these records into their OPACs.

Adding catalogue records for data to the OPAC can also be the first step in providing data services within a library. Even before a library provides other services, this simple step can begin the process of integrating access to data with access to other information. Reference librarians, subject bibliographers, and catalogers can all contribute with little additional training or expense.

Here is a scenario that we have seen several times with slight variations. A collection of data files exists in a computer center, data archive, social science research center, or even an academic department. However, these data are not accessible to all potential users because there is no central listing that enables the discovery of these data. The library may offer or be asked to

[2] Libraries are rethinking the traditional library Online Public Access Catalog (OPAC). We use the term here in a generic sense to refer to whatever method the library uses to manage and make available to the public a record of its holdings and resources.

[3] Anglo-American Cataloguing Rules, Second Edition, 1998 Revision with Amendments 1999 and 2001, Published jointly by Canadian Library Association, Library Association Publishing, and American Library Association.

[4] http://www.icpsr.umich.edu/icpsrweb/ICPSR/or/metadata/

catalogue the data files or the accompanying documentation or both and to include these records in its OPAC. If there is no easy access to the accompanying documentation, the library might also offer to house and maintain these materials, or the library might choose to buy copies for its collection thereby adding another point of access.

Other methods of organization include building specialized search facilities or databases for data, providing a variable-level search, and providing web pages that list popular data sets with information about how to access and use them. Chapter 17 examines these options in more detail.

Collection Services – Preserve

Service: It is useful to differentiate between two kinds of preservation services, which, though they have much in common in practice, have different purposes. One purpose is to preserve data for as long as the data are needed and used locally. That could be as short as a few months or as long as decades. A separate purpose is to preserve data as the copy of last resort, the copy that will be preserved in perpetuity even if there is no local use of the data. It might be useful to think of the first as a *library* activity and the second as an *archive* activity.

In a traditional archive, there is literally one copy of an original document, but in a data archive environment, many essentially identical copies may exist in other data libraries. A data archive, therefore, must take the additional responsibility of maintaining the one, guaranteed-authentic copy of the data and ensuring its survival, usability, and integrity over time. As OAIS tells us, the scope of an archive's responsibility is determined by the needs of its Designated Community. (See Chapter 9 on Preservation.)

In traditional archives – and even in data archives in the twentieth century – it was typical to think of the "life span" of materials. This encompassed the time from the "birth" of the item through its useful life to its retirement when it was no longer useful or usable or needed. In the twenty-first century, archivists must think of the *lifecycle* of information. In this process data are not just analyzed, but are re-analyzed, reused, repurposed, and recompiled in teaching, learning, and research environments. Data files are not just stored in a repository in a linear birth-to-death model, but data and metadata are managed as part of a cyclical process of data collection, discovery, sharing, and repurposing.[5]

[5] Green, Ann G., and Myron P. Gutmann. "Building partnerships among social science researchers, institution*." OCLC Systems & Services* 23, no. 1 (2007): 35 - 53. http://www.emeraldinsight.com/10.1108/10650750710720757. http://hdl.handle.net/2027.42/41214.

The choice between the role of library and the role of archive is a significant one and has strong implications for the types of materials collected, the services provided, and the resources required. A data library collection might include only materials that are available from stable distributors, while a data archive might seek out and accept unique data. An archive accepts an obligation to ensure the integrity of its collection over time with careful preservation techniques, while a library might not take the extra steps necessary to “archive” the data since the assumption is that it could always get another copy if needed. A library might keep data in the most conveniently usable format even if that format was not easily preservable.

Every item in a collection does not have to be treated identically. A policy could, of course, define an entire collection as either an archive or a library and define identical procedures for all items in the collection. Alternatively, however, a policy might be flexible and allow for differential treatment of materials. A particular data library might choose to use a "library" collection policy in general, but occasionally accept the donation of locally produced data that are appropriately documented and apply archival procedures to these data. Similarly, a data archive might, as a matter of convenience to its users, include in its collection some data that are readily available and archived by another institution. Such policies and exceptions to policies should be as specific and explicit as possible.

Skills: Data archiving requires data management skills as well as traditional archiving skills. Traditional skills include record keeping, documenting provenance, creating administrative records about the data and data handling, legal issues, disaster planning, and so forth. Implementing these will require data management skills including risk assessment and risk reduction, examining data files, understanding file formats, ensuring that the needed Representation Information is available (see Chapter 9 on Preservation), and working with a variety of operating systems, software, and various media. These latter skills of working with files across operating systems and media and of transferring files without loss of integrity are essential to migrating data across evolutionary changes in computing. A data archivist needs special skills to ensure that data will be readable and usable in the future and that the integrity of the content remains consistent irrespective of changes in technology.

Being able to create and manage metadata is essential to preserving data. Some of the cataloguing skills mentioned above apply here. However, other archival record description standards need to be considered in the production of the metadata. These include documenting the relationships between files and the chain of custody of files as they are migrated, refreshed, and reformatted over time. There are many taxonomies of types of metadata.

Anne J. Gilliland-Sweatland lists five types of metadata that need to be managed for long-term preservation.[6]

1. Administrative
2. Descriptive
3. Preservation
4. Technical
5. Use

OAIS describes metadata as types of information necessary for preservation and use of content information. (See Chapter 9 on Preservation for more about OAIS metadata.)

The DDI version 2 divided the description of data into 4 components:

1. Document Description
2. Study Description
3. Data Files Description
4. Variable Description

DDI version 3 uses a conceptual model that is both more complex and more flexible because it is modular. The conceptual module provides a way of documenting data across the life cycle of its conceptualization, creation, use, and preservation.

DDI is not the only metadata standard for documenting statistical data. Other standards include SDMX (Statistical Data and Metadata Exchange)[7], and XBRL (eXtensible Business Reporting Language)[8]. Increasingly, data libraries will use XML as the format of choice for storing and maintaining metadata. An understanding of DDI requires an understanding of how to work with data files and use documentation. Working with XML will increasingly require the use of XML software.

Activities: A data archive preserves the content of the data in a way that permits easily moving data across operating systems and storage media ensuring that the data are usable. The metadata must record any changes that are made to accomplish this and to document the relationships among the various files and documentation.

To preserve data and then not provide the necessary support to access it would remove the value of having saved the data. Therefore, a data archive

[6] Gilliland-Sweatland, Anne, "Different Types of Metadata and Their Functions." In, Introduction to Metadata: Pathways to Digital Information, ed. By Murtha Baca. Getty Information Institute, 1998.

[7] http://www.sdmx.org/

[8] http://www.xbrl.org/

should also have or enable the functionality of a data library. In this way, the levels of service that define the reference and computing functionality of a data library apply to the data archive, too.

A traditional archive will operate with a clear statement about what it will and will not accept. Data archives need to follow similar practices. For example, no data archive can accept data files without comprehensive, accurate, and usable documentation. No OAIS conforming archive can accept a SIP without adequate descriptive information.

Collection Services – Providing Service

Service: Providing users with help in locating and using data is a topic in itself. Briefly, services can range from helping users search for and find data, help with understanding documentation, help with downloading or otherwise getting data ready for analysis. In OAIS terms, service delivery incudes discovery and delivery (DIP). Services can also include software consulting, statistical consulting, and even statistical analysis. We discuss these services in much more detail in the next chapter.

Skills: Obviously, the skills needed depend on which services are provided. Examples of skills used for service include: basic reference , searching for data as described in chapter 4, understanding and describing the use of data documentation, and using statistical software.

Activities: Examples of service activities include: helping a user find a known study, helping uses with specialized catalogs indexes and Internet tools, and helping users with specialized software used by licensed data products.

Summary

By retaining the flexibility to provide different levels of collection service for different datasets and different user groups, the data library can provide optimum effective service without the constraints of a rigid, one-size-fits-all policy. Providing services in the context of the data library’s mission and goals keeps the data library focused on users, collections, and services and helps ensure that long-term goals will be met. Developing policies and strategies in the context of collections and services provided by other data libraries, data archives, and even commercial vendors, adds to the options a data library has at its disposal.

11.12 Levels of Collection Service

Data Basics

Levels of Reference Service

Reference services for data can be provided whether or not your library has a local collection of data files. In fact, as more data become available on the Internet, it is becoming increasingly necessary for libraries to provide services locally for data that are stored remotely. Libraries that have locally stored collections have additional service options and responsibilities. For example, having a local copy of a dataset may enable a library to provide its users with more reliable access or an interface that better meets local needs. Doing so may involve other costs, but may also increase efficiency of the service.

To provide reference service for data requires some special skills. As noted repeatedly below, every level of reference service requires a commitment on the part of the organization and the staff to continuing, regular staff training, professional development, and the allocation of time to work with data and data users.

When planning for and designing reference services for data, it is just as important to identify local users' needs as it is when building local collections of data. Because any given community of data users may have a wide variety of needs and expectations, the development of a service plan that clearly articulates what expectations can and cannot be met is critical to maintaining a practical and sustainable service.

This chapter presents some examples of different levels of reference service. These services are not necessarily hierarchical, but each "higher level" service does require more skills or resources or both. We emphasize that these examples are meant to guide the planning process for a service and are not presented as a prescription. They may also provide ideas for staged growth of a service over time. In practice, a "service matrix" (similar to the "collection matrix" in Chapter 11) might be used to identify different services available for different sets of data or different user communities.

Level One Reference Service: Data Identification

Services: Perhaps the most basic kind of question that any data reference service should be able to answer is a known-item request. Data users frequently ask for data from a particular study. They may have used the data at another institution, had a particular study recommended to them by a colleague or professor, or found a citation to the data in a publication.

A thorough answer to a known-item request should include determining the availability of the data, clarifying any conditions on the use of the data, and identifying the procedures to access the data. Such activities can quickly escalate to higher levels of service, though, and can require computing resources and more advanced skills, so care should be taken in describing this level of service.

In some instances, data users may not require or want an entire data file or may need additional help in identifying which part of the data they want. For example, some studies consist of multiple files or parts -- often organized around social units, time, or space (see Chapter 5). Longitudinal studies consist of separate "waves" of interviews. Economic data often consist of time-series for thousands or even tens of thousands of different variables and users rarely want more than a few at a time. Assisting the data user to identify just the data that she needs is a helpful part of a known-item data identification service.

Data files are often available in more than one format and ensuring that the user gets the data in a format with which she can work is essential. Different versions or editions of data may be available and there may be more than one source or distributor for the data. Government data may be available, for example, directly from the government and from private sector vendors who repackage the data. Being able to identify the alternative formats, versions, and sources of data, describe them to users, and help the user identify the one that will match her needs is an important aspect of this basic level of service. (See level-four, below, for additional service opportunities.)

Skills: Three kinds of skill sets aid and inform this level of service. First, general information searching skills (e.g., ability to search library catalogues and databases, familiarity with online guides and even older, published special catalogues and guides) are fundamental. Corresponding skills with searching the Internet, particularly knowledge of and experience using specialized web search engines, are essential. Additionally, the special searching strategies relevant specifically to data, which are described in Chapter 4, are also needed.

Second, reference interview skills apply as well. Known-item requests for published materials (books, laws, journals or journal articles) are common at library reference desks and many of these same skills apply to the data reference interview. (Note that a known-item request in a data context is often complex. See Chapter 16, "Reference Strategies for Data Services" for more about the data reference interview.)

Third, there is no substitute for a general familiarity with popular studies and datasets. Staff who provide data services regularly will pick up much of this familiarity on the job, but an active, regular perusal of the substantive literature in which data findings are published is a very helpful way of learning more. Data users often refer to data by common names (e.g., the twins file, the death data, "kids"), acronyms (e.g., PSID, SLID, GSS)[1], the name of a principal investigator (e.g., the Parnes data, the Espy file, the Terman study), or other short-hand conventions (e.g., "the census", the Eurobarameters). While such shorthand references alone may not provide enough information to find the data, reference providers should be able to recognize such requests and have strategies to find

[1] See Appendix B for a list of some common names and acronyms.

answers easily. Often, this is a matter of experience. The more one works with data and data users, the more familiarity one will have with common and well-known data sources. An organization wishing to provide even this most basic level of reference service should commit adequate staff time to it just as it would to any other specialized collection or archive.

Activities: A basic and yet excellent place to start with this level of reference is adding records for social science data to whatever catalogs of resources your library maintains. Most libraries belong to one of several consortia that share catalogue records including records for many data products. Adding catalogue records, which is a routine activity of a library, can be done for data whether or not the library is the location of the data, though care should be taken to adhere to any policies for what is appropriate for inclusion in the local catalogue.[2]

Providing a “ready reference" collection for data that includes quick links to key web sites, published guides that list or describe data, guides to products of data producers and vendors, and even copies of data documentation will give users and service providers resources for identifying studies. While printed guides to data and data archives and libraries are much less commonly produced today than they were a few years ago, it is still helpful to have on hand some of the older guides to data collections because they will help identify older studies. Published guides to statistical resources will help track down sources of data. Published collections of statistics are often a useful starting point for locating data sources behind popular statistics.

Using whatever local tools are available (e.g., specialized databases, simple web pages of bookmarks, social bookmarking sites such as delicious.com, etc.) for providing quick pointers to the major web-based catalogues of data, data libraries, data archives, and data vendors will also help users and service providers alike. Creating special guides can be a simple, effective service. Such guides can list local holdings, popular studies, and studies known to be used locally, and can even include instructions about how to access data files.

Level Two Reference Service: Finding Data by Subject

Services: At least as common as known-item requests are requests for data on a particular subject. In some environments, subject requests are more common than known item requests. New users of data, in particular, are more likely to ask for data by subject. A data reference service that efficiently responds to such requests provides an additional level of reference service. Again, there are parallels between this kind of data reference question and similar inquiries for research output resources (i.e., books and journal articles).

[2] As noted in Chapter 11, we use the term OPAC (Online Public Access Catalog) in a generic sense to refer to whatever method the library uses to manage and make available to the public a record of its holdings and resources.

Example: A patron asks "Do you have some statistics on income?" A basic level-two reference interview should certainly be able to identify if the user needs data or statistics, as explained in Chapter 1. If the user needs data, there are (as is explained in Chapter 4) specialized reference tools and techniques for searching for data. One immediate way of answering such a question would be to search a catalogue of locally held datasets by subject. Such a search might locate a variety of popular surveys (e.g., in the United States, *National Longitudinal Survey*, the *Current Population Survey*, the *Census of Population and Housing*, the *Survey of Income and Program Participation*, or in Canada, the *Survey of Labour and Income Dynamics*, the *Survey of Consumer Finances*, the *Survey of Family Expenditures*), all of which include variables that are useful for analyzing income. Where available, a search of survey questions and data-set variables may be worthwhile.

A level-two service that would be a bit more comprehensive would provide some guidance on the differences among those studies that include variables about income. To do this, service staff should know the difference between aggregate data and microdata, should understand the difference between cross-sectional and longitudinal studies, and be comfortable talking with users about sample sizes, choice of sample, unit of observation (e.g., household or individual), geographic coverage, and so forth. (See Chapter 5, “Data-speak: A Search Vocabulary for Data.”)

Skills: To provide this kind of service, reference staff will need a working understanding of the methods and practices of quantitative social science. This knowledge enables a data librarian to communicate with users in an intelligent and informed way in order to help them locate and use the data that they need.

Examples of specific skills in this category include the ability to:

- Understand the process of collecting social science data and constructing data files;

- Ability to use social science research terminology (e.g., dependent and independent variables, panel study, time series);

- Interpret data documentation.

Activities: All of the activities in level-one reference service will provide a foundation for level-two reference. The addition of consultation services to help users locate studies by subject expands service into level two. While level-one reference service can be integrated into a traditional library reference service, level-two service must allow for longer consultation times. Although catalogues of data and subject entries in library catalogs can facilitate finding data by subject, many data questions will require the assistance of a service provider who is familiar with a broad array of data sources and finding aids that provide more detail than the typical OPAC.

Identifying a study that may match the researcher's needs is only a first step. After identifying one or more studies, the researcher will need to work with the documentation to see if the data really meets all of his criteria. The data service plan might give guidelines for the extent of involvement of the data librarian in examining data documentation at this level of service.

Because there is no "books in print" for data, because data are often not cited adequately, and because the distribution of data does not always depend on listings in catalogues or databases, one often needs to rely on more informal sources to find data. Experienced data librarians rely on mailing lists that reach communities of subject specialists, personal contacts, and other data librarians to track down data not easily locatable with traditional tools. An essential activity for this level of service is, therefore, ongoing professional development. Attending meetings of professional data-relevant organizations, maintaining memberships on mailing lists, and cultivating personal contacts within the data community build this type of social network.

Locating difficult to find data may require a willingness to make phone calls to likely producers, distributors, or creators of data. Government researchers, for instance, often are willing to share their extensive knowledge of a subject area and sources of data. Having quick access to online directories and a ready reference collection of print directories will facilitate this.

At this level of service, the contemporary data librarian should be creative and use "Web 2.0" technologies to record and share information. Since data-reference can be time consuming, one should look for ways to record answers found so that they can be shared with colleagues and end-users – and so no one will have to re-discover information that was found last month! Some reference departments are developing web sites and blogs to record answers to general reference questions and James R. Jacobs at Stanford University is managing a Google Custom Search Engine[3] to index several of these.[4]

Level Three Reference Service: Data Content Recommendation

Services: Providing service for large datasets is somewhat like providing service for a collection of manuscripts. Typically, a data file will record information on dozens or even hundreds of topics and, all too often, few of these topics are indexed. Just as an archivist may remember a letter from an ex-slave buried in the private papers of an Ohio school teacher, a data archivist may remember that

[3] http://www.google.com/cse/

[4] "Library Questions and Answers" http://tinyurl.com/6ylz4o. For an example statistics-related reference question see "Historical Foreign Direct Investment (FDI) statistics" (http://tinyurl.com/5j6g2q). See also "Help build a library question and answer custom search" by James R. Jacobs, 2008-06-16 http://freegovinfo.info/node/1888

a particular poll asked a question about day care availability for single parents. As data libraries are increasingly able to provide searching for individual variables (see Chapter 4) the data librarian will increasingly be expected to provide help with such searches and to provide help when full-text searches are not adequate. As a data service provider becomes more familiar with the contents of studies, she will be better able to help the data user locate a study relevant to the user's research interest.

Skills: The ability to provide this kind of service comes from a greater familiarity with the content of studies. Experience in quantitative research as either an academic or data professional is one way to achieve such knowledge. The best preparation for this service level, however, is working each day with data users. These experiences in searching for studies, in reading and re-reading data documentation with particular research questions in mind, and in working with data dictionaries or codebooks and with data files all provide a greater understanding about the data of a particular research community.

It is not necessary to wait for a question from a user to become familiar with the data from a particular study. Data librarians should anticipate the kinds of questions that a study might address by reading data documentation, by examining bibliographies of research based on particular studies[5], and by reading some of the research emanating from particular studies. When a study results in published statistical reports, these can be a useful overview of the content of the study.[6] Some, large studies have their own meta-literature – books about using the studies.[7] Some of the major data projects, such as a national census, are guaranteed to attract the interest of data users. Since national statistical agencies usually require one or more years to release the statistics and data from a census, the data service provider has plenty of time to investigate basic characteristics of these products in advance of their release.

[5] Some large, well-known studies have their own bibliographies and newsletters. (e.g., *Annotated bibliography of papers using the general social surveys* by by Tom W. Smith and Bradley J. Arnold [Ann Arbor, Mich.] : Published and distributed by the Inter- university Consortium for Political and Social Research, [1990], 8th ed., and *The NLS Annotated Bibliography*, NLS User Services at the Center for Human Resource Research, Ohio State University http://www.nlsbibliography.org/. Some studies can be found in the *Social Science Citation Index* (Web of Science).

[6] For example, the U.S. Census Bureau issues a series of printed reports based on its *Current Population Surveys* data. The series, "Current Population Reports" (also known as the "P series") includes P20, *Population Characteristics*, P23, *Special Studies,* and P60, *Consumer Income and Poverty*; see, http://www.census.gov/main/www/cprs.html.

[7] The U.S. Census is a notable example. See, for instance, *The concept of veteran status in the U.S. decennial censuses: 1960-90* by Diane C. Cowper, Lynne R. Heltman, and Stephen J. Dienstfrey [Princeton, N.J.] : Association of Public Data Users, 1994; and *Analysis with local census data : portraits of change* by Dowell Myers, Boston : Academic Press, c1992. Another notable example of a user guide to data is *The NORC general social survey : a user's guide* by James A. Davis, Tom W. Smith, Newbury Park, Calif. : Sage Publications, c1992. Also see the NORC website (http://www3.norc.org/GSS+Website/Publications/GSS+Reports/) for more reports on GSS use.

An essential skill at this level is the ability to read and understand data documentation. Often, only by reading the documentation carefully will one determine the universe, time period, geographic coverage, and unit of observation of a study. Frequently, only by reading the text of a survey's questionnaire will one know if the data can be used in a particular analysis.

Activities: Providing a consultation service for users is the primary activity of this level. As suggested above, the more one works with users, the more familiar one becomes with data and the better equipped one is to provide this kind of service.

Level Four Reference Service: Data File Advisory Services

Services: While level-three service deals primarily with the content of studies, level-four service deals more with the structure of the data as distributed in computer files. Data files may have complexities that make them difficult to understand and process. For example, a survey about travelers may include separate files containing variables about the travelers, the trips they took, and the methods of transportation used on these trips. To assist data users with the complexities of the structure of data files, a service must go beyond simply recommending data based on the documented contents of a study. This fourth level of service helps users to resolve problems about complexities in data format, structure, and processing, and to understand the analytical opportunities that exist within data since complex structures often create opportunities for complex analyses.

While technological advances have made it increasingly easy for users to get data files that are ready to use, this has not obviated the need for data services at this level.

The technological advances are impressive and very useful. Where once it was common for users to have to jump through many technical hoops just to open a data file, today three technological developments have reduced barriers to use drastically. First, data files are often downloadable in a statistical software format rather than as raw ASCII data files with separate documentation. For the user with the appropriate statistical software this means that, once the file is downloaded, the user can begin working with the data at once. (In the not too distant past, users would have to create SPSS syntax files or their equivalent for other statistical software before the data file could be opened.) Secondly, software such as Stat/Transfer[8] makes it very simple to convert a data file in the format of one statistical software into the format for another. Thus, a user who does not have SPSS, but does have STATA, can download an SPSS file and convert it to STATA quickly and easily. Third, popular datasets are now available on the Web embedded in web-applications that allow users to do quick analysis

[8] http://www.stattransfer.com/

without any statistical software on their own machines. For example, ICPSR provides many datasets for online analysis using SDA and NESSTAR software.[9]

While these developments allow users to jump over many hurdles quickly and efficiently and make it possible for many users to do everything they wish to do easily, they do not solve all problems for all users for all datasets. There are at least three categories of service issues that may indicate the need for level four data services.

1. First, and most obviously, there are those datasets that no data library has yet configured to be ready for analysis. There are still lots of "raw" datasets – particularly older datasets – that will require writing statistical software commands in order to get the data ready for analysis. Ironically, as more users are attracted by ready-to-use data, there will be more users who will have experience analyzing data, but little or no experience getting data ready for analysis. Data services can help users by providing training and by providing software consultation and programming.

2. Second, users may need training and assistance in understanding online statistical services. As these services get more sophisticated, users new to data analysis will need help understanding concepts of recoding, subsetting, variable and case selection, and so forth.

3. Third, some users will need to restructure datasets and combine datasets from more than one source and this will require data management skills that many users do not have. A common example is linking social data to geo-referenced data. A user may have experience with one or the other type of data, but not both and may be new to the concepts of geo-codes and how different software handles those codes. Data services can provide instructions, training, and programming help in such situations. A data librarian might also provide advice on the use of weight variables to adjust for the sampling methodology.[10] Although weight variables are usually documented, new data users may not know to look for them. Moreover, some studies have complex weighting schemes with multiple weight variables that are essential to generalize results to specific subpopulations. (See also, Appendix A. "Understanding Weight Variables.")

As noted above in level-two service, when helping a user choose a data file that will answer a specific question, a data services specialist can help the user choose files that meet their technical knowledge and needs as well as their content needs. For example, a particular study may require specific software because of the unusual way the data are stored.

[9] http://www.icpsr.umich.edu/icpsrweb/ICPSR/access/sda.jsp

[10] Many surveys contain special variables for adjusting results and these "weight variables" must be used to obtain generalizable results. See Appendix B for more about weight variables.

In addition, data may have limitations that are not obvious from their documentation. For example, U.S. foreign trade data contain limitations because of the way data are reported (data released for one month actually included trade from earlier months), because of changes in the system of classifying industries, because exports are not counted as carefully as imports, and because aggregate figures are revised, but industry level figures are not.

And, of course, as users increasingly rely on ready-to-use datasets they may neglect to read the documentation adequately. A dataset that is preloaded in statistical software and ready to analyze may not provide any indication within the statistical software of the universe of a particular variable or the distinction between two different measures of income. The data librarian who is knowledgeable about data documentation can help the user track down why their analysis is yielding unexpected results by helping the user understand and make use of the documentation.

As a data librarian works with a variety of researchers, data users may discuss with the data librarian the various problems they encounter with the data. This shared knowledge and experience can result in the data librarian knowing more about the difficulties in working with a particular data file than any individual researcher. By sharing this knowledge the data service provider can save data users valuable time in conducting their analyses.

Skills: Many of the skills of this fourth level are largely gained from the experience of working with data and from exchanging information with data users and data providers. Special computing skills are necessary for those activities that require working directly with the data or providing computer-based solutions to working with complex data files. Knowledge of documentation and social science terminology and methodology can solve many problems at level-four.

Activities: As with level-three service, the provision of consultation services is the primary activity at level-four. Offering this level of service, however, usually takes more time per consultation than the previous two levels of reference service. Both sharing known information and working with users to gain new information takes time. Resources must also be available to the data service provider to permit her attendance at conferences or meetings where other data service professionals convene to share this type of information.

In some cases, the most efficient way to help users work with complex data files may be to provide computer-based solutions. For example, if many users need to use a complexly structured data file with a particular piece of statistical software, pre-loading the data into the preferred software could provide a service that is both better (saves users time, allows users to begin analysis without having to learn unnecessary data-processing steps) and more efficient (saves repeating the same consultation multiple times). A data service that provides its own online analytical service has the advantage of choosing which datasets to make available in ready-to-use format.[11] Computing and software resources and software or programming skills are necessary for such activities and may require a larger funding commitment.

Level Five Reference Service: Data Extraction Services

Services: A reference service for data extraction will provide users with the information they need to create a subset of a data file rather than working with the entire data file.[12] As noted above under level-one service, many users will neither need nor want a complete data file. This is particularly true of large, complex studies and of users who wish to use spreadsheet software. For example, a student may require only a few variables from a survey that asked respondents hundreds of questions. Another user might only want a subset of cases; for instance: only bus riders from a survey of commuters, or only residents of selected zip codes from a census.

Support for data extraction may involve written guides, consultations, or both. A handout or web page might be as specific as a list of statistical package commands that when processed create a subset for a particular study, or as general as a tutorial about reading documentation to select variables and cases. Consultations may involve discussions about how to conduct the extraction or even entail helping the user do the extraction. Being able to help users understand whether a particular subset can be created easily or will require complex programming will save everyone time in the long run.

Skills: A basic understanding about data file structure and how subsets can be selected and created is needed to provide this level of service. The ability to read and understand data documentation is essential, especially in determining if it is even possible to extract the data the user desires. (See Chapter 6, "Reading Data Documentation.")

Some datasets are still distributed as "data products" on portable media such as CDs and DVDs and are often intended for use by an individual on a personal computer. (Examples include *CPS Utilities*[13] and Geolytics U.S. Census data[14].) Others are web-based with unique user-interfaces. Examples include the Census Bureau's *American Factfinder*[15]) Such data products often bundle the data in a proprietary format with special, often proprietary, software or otherwise hide the raw data behind the interface. Such products make it easy for an individual who uses that data product frequently to use the data. But, it often

[11] There are several affordable online analytical tools available for hosting locally or through shared resources. Examples include Nesstar (http://www.nesstar.com/), SDA (http://sda.berkeley.edu/), and Dataverse (http://dvn.iq.harvard.edu/dvn/).
[12] See also Chapter 13 for more about providing subsets.
[13] http://www.unicon.com/
[14] E.g., http://www.geolytics.com/USCensus,Census-2000-Products,Categories.asp
[15] http://factfinder2.census.gov/

increases the complexity of providing service for a data library where the dataset is one of many. Providing data service can also be made more complex when essentially the same data are available from more than one service since different services usually have different interfaces and even different functionality. For example, the U.S. *General Social Survey* is available for online analysis from ICPSR,[16] from University of California Berkeley,[17] and from the National Opinion Research Center,[18] but the interfaces are different and what you can do differs. When such data products are part of the mix of data for which a data library wishes to provide service, it will be necessary for staff to have an understanding of that special software or interface to provide this level of service. As noted above, a conceptual understanding of how subsets of data are created in general is useful; in this environment, such an understanding helps in the use of such products.

Being able to use one of the major statistical software systems[19] or one of the specialized data extracting and reformatting software packages[20] to subset cases, or variables, or both is a mandatory skill if performing data extractions is one of your service offerings. Simple analyses using frequencies or cross-tabulations may be necessary when determining the size and definition of a desired subset. Even if computing or programming assistance is not provided as part of a reference service, understanding conceptually how statistical software can be used to create subsets is necessary. This knowledge can help assess if the data that are wanted will be easy to extract.

Activities: Helping users identify and create data extractions is the primary activity of this level. This may entail writing guides or tutorials about the process of creating extracts or teaching workshops for patrons that describe the basic steps in subsetting data. It may involve teaching users to use specialized data products on stand-alone machines or on the Internet. Consultations remain an important activity and draw upon skills in interpreting data documentation and in working with data files. It is essential to provide staff time for experimenting with software tools to perform actual extractions.

Another activity at this level of service is to provide a public service computer for data acquisition. Such a computer could be equipped with statistical software, statistical translation software (e.g., Stat/Transfer), the capability of downloading data, and capability of transferring work done on the computer using portable media, FTP, email, or cloud-based services. Users could use this machine to get the data they need, create subsets, and reformat the data for use with particular software. Such a service might be particularly indicated at a college where some

[16] http://dx.doi.org/10.3886/ICPSR31521.v1
[17] http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss10
[18] http://www3.norc.org/GSS+Website/Data+Analysis/
[19] e.g.,, SAS, SPSS, Stata, S, R.
[20] e.g., Stat/Transfer

students need computing resources and the college does not have adequate student facilities in support of statistical acquisition or analysis.

Level Six Reference Services: Data Analysis Advisory Services

Services: This level moves data services beyond information management to information processing and analysis. The kinds of support offered may include advising users on appropriate statistical techniques, helping users choose statistical software, writing code in the language of a statistical program, debugging statistical programs, and interpreting statistical results.

Skills: Staff providing this service require training in quantitative methods, including both descriptive and inferential statistics and modeling techniques. In addition, knowledge of statistical software and methods of transforming data are required to conduct actual analyses.

Activities: To provide this advanced consultation service, specialized skills in statistical computing are required. Combining data selection and retrieval services with statistical consulting provides users with a comprehensive, even ideal, service. Note that this level of reference service requires a large allocation of staff time and resources.

Statistical consulting is one activity that can be rather easily separated from the previous five levels of reference service and delegated to statistical computing specialists. A data service that removes barriers to users for data selection and acquisition provides an excellent complement to a separate data service of statistical consulting offered by a separate unit in your organization. At smaller colleges, instructional faculty may provide statistical consulting along with their classes. With the data delivered conveniently by a data service, a statistical consulting service can concentrate on all aspects of supporting data analysis. There may need to be only a small amount of contact between the departments to ensure a coordinated service.

Level Seven Reference Services: Comprehensive Data Analysis Services

Service: This level of service would do everything for the user. Staff would analyze data as requested and deliver finished output in the form of charts, graphs, tables, maps, and other analytic products. While most data services do not provide this level, a few do. As is true of all these levels of service, some users will expect this service. It is prudent to anticipate this kind of request for service and have a clear policy delineating the services that are provided and those that are not. For those that contemplate providing such services, they might consider defining precisely which services will be offered, to whom those services are available (e.g., faculty? Graduate students? Undergraduates? non-affiliates?), and if there are fees charged.

Because these services are comprehensive and complex and expensive, non-commercial organizations have rarely chosen to provide this level of service.

One way of providing some of the features of level-seven service without actually providing analysis is to host web-based analytical tools that make it easier for users to do their own analysis. It has become relatively easy to do this with off-the-shelf software.[21] These tools can provide users who have little or no computing or statistical software experience with a way of creating charts, maps, tables, and graphs with little effort. Some schools are beginning to find it easier to rely on online analytical tools than to maintain computer labs for statistical computing. Instructors are beginning to discover that when they use such services they can spend more time teaching statistics and discipline-specific concepts and less on “how to use SPSS in the computer lab.”

Such changes in administrative policy and teaching methodology may affect the level of data service provided by the data library. It can change the model from doing the work for users to providing tools that enable the user to do their own work. Inevitably, such a service must strike a balance between flexibility and ease-of-use; the easier it is to use the service, the less flexible it tends to be and vice versa. For data libraries that have a very focused collection, or whose users fit a rather specific profile, or for institutions that support distance education and off-site users, providing such services may become increasingly attractive. This is a trend to watch.

Skills: This level requires the skills of all the previous six levels. In addition, expertise in organizing and displaying statistical results in the form of reports and graphs becomes an important skill. Management of the service and billing and accounting may also be necessary.

If web-based services are offered for users to do more advanced analytical procedures, special software skills, web-hosting skills, and systems administration skills will be needed.

Activities: This level of service introduces an aspect of contractual liability that is not necessarily part of previously discussed service levels. Therefore, having a carefully detailed policy that stipulates what will and will not be done, for whom, at what costs, and under what deadlines is important. There may also be restrictions imposed by data providers on the ability of the data library to charge for the use of licensed data.

[21] As noted earlier SDA, Nesstar, Dataverse.

12.14 Levels of Reference Service

Data Basics

Levels of Computing Services

Data require computers for storage, delivery, and use. Data users must have access to computing resources in order to use data. A data library must address both who will provide computing services and which services will be provided. It may seem obvious that a data library needs to have computing resources, but it is important to realize that each data library has significant options among different levels of computing service it might provide. These choices are more complex and varied today than they were even a few years ago. Organizations starting a new data service need to pay particular attention to computing services and carefully consider options in order to avoid either over committing or providing insufficient services. This chapter addresses some of the service options.

In order to make appropriate choices, the data library should understand what computing resources are available to its users. Data users today may have access to their own personal computing resources, or shared computing resources, or both. These provide the context within which a data library decides which computing services its user community requires.

Data libraries interested in equal access to information should not overlook those people who do not have access to their own computing resources. A data service's commitment to its community should account for all users, not just those who can afford to own their own computer and software. A service plan should address how all its users will access data. While libraries that serve the general public most often face these issues, it is, we believe, important for all libraries to examine their user base carefully and not assume that, for example, "all freshmen have laptops these days." Examining the computing needs of users also provides an opportunity for libraries to examine potential users of data as well as existing users. For example, those who use spreadsheets for simple tasks but who do not have expensive statistical software might require a different level of computing support than those who do have their own computing resources. Finally, for many situations a high speed network connection is essential for downloading large files and not everyone has high speed access yet.

As computing hardware, software, and networks change and evolve, so the computing services that data libraries offer must evolve and change. When centralized, mainframe computing was the de facto standard, data libraries could be (and often were) part of a centralized computing environment or provide a service of delivering data to such an environment. With the advent of powerful desktop computers and software, there was less need for data libraries to support large computer servers and many assumed the role of delivering data to individual users, often on portable media. With the growth of the World Wide Web, the increasing availability of broadband, high speed network connections, and the evolution of web-server based data-delivery software, many data libraries have moved to sophisticated, web-based services that again require powerful server computers and related data management and programming resources. As data vendors and distributors increasingly target their products to individuals and create data products that are bundled with proprietary software and designed for use by an individual on a stand-alone personal computer, data libraries have to face a plethora of interfaces and difficult compatibility problems. As government agencies, data vendors, and data distributors move to providing web-based data services (as opposed to simple delivery of data files), data libraries must carefully examine their roles in the delivery of data and in the long-term preservation of data.

While the data library does not control these kinds of changes, it can design computing services that are flexible enough to accommodate change. Here are some guidelines for making choices.

1. When possible, avoid computing services that require a particular technology. Operating systems change; statistical software that is popular this year may go unused next year. Proprietary data formats may be (almost certainly will be) unreadable in the future. Portable media are notoriously short-lived -- not necessarily in terms of the physical readability of the media, though that is often the case -- but in terms of the availability of hardware and the software "drivers" that enable the hardware.

   This guideline applies to the user environment as well. It would be unwise, for example, to design a delivery system that requires users to have a particular computing environment, or that assumes every user will have the same computing platform in 3 years that they have today.

2. Where particular technologies are useful or unavoidable, assume that they are temporary. For example, do not assume that today's cutting edge storage media is a long-term storage solution. Between the late 1980s and 2010, we have seen a variety of storage media become unusable: half-inch tape, tape-cartridges, five and one-quarter inch and three and one-half inch floppy disks, "zip" disks, and flash drives. While we can still read most CD-ROMs and DVDs, we are already seeing the early stages of obsolescence of these media. "Standards" for data documentation have evolved from printed codebooks, to OSIRIS data dictionaries, to ASCII codebooks, to Adobe Acrobat Portable Document Format files, to DDI 2, to DDI 3.

3. For data preservation, rely only on published, well-documented, preferably open, formats for both data and metadata. Avoid instantiating metadata in software unless the metadata are separately documented as well. For example, a dataset created with statistical software in 1977 may not be readable with today's versions of the same software and thus variable names, locations, labels, and data types would be lost without external documentation. Formulas used in some spreadsheets are a form of metadata and data encapsulated in formulas embedded only in a spreadsheet file will be lost when used under a different OS or when the spreadsheet software file cannot be read at all.

4. Make use of a specialized, unique, or proprietary data delivery solution only as a temporary service, not a long-term solution. As noted above, there are more and more data products that provide useful features and make data easy to use. There is nothing inherently wrong with such products as long as the data library does not assume they will be usable in the future. Acquiring such products and making them available can be a useful short-term service, but is almost certainly not a long-term solution. This guideline applies to web-based services as well as PC based products.

5. For long-term solutions, focus on preservable data and metadata first and design a modular service delivery system that relies on replaceable software for delivery solutions. There is rarely a reason to spend lots of resources creating a service that will be obsolete or unusable in a few years (or less). Instead, think of the data and metadata as permanent and the delivery software as temporary.

Computing service issues influence and are influenced by preservation strategies, reference services policies, and collection service issues. One approach to this situation is to review computing resources and skills first, and then ask, "What can the data library do with these?" A better approach, we believe, is for the data library to focus on the services it wishes to provide first, then examine the computing resources it has at its disposal, and finally determine a way to best meet its service goals with those resources.

In general, there are two key computing service challenges that face data libraries today.

The first challenge is how to deliver data to users in formats that are compatible with users' machines and software through networks or on digital media that allow users to load the data physically onto their machines. While remote computing, web-based statistical analysis, and “cloud” and “grid” computing are all significant emerging technologies, most users will still want and need at some point in the research cycle to get a usable copy of the data they use.

The second challenge is how to provide services that allow users to locate, explore, and acquire data or statistics easily. The levels of computing service described below provide examples of ways to address these challenges. As in previous chapters, a "matrix" approach, which specifies different levels of service for different data or groups of users, gives the data library more flexibility than a "one level fits all" model.

Level 1 Computing Services: Pass-Through Data Delivery

Services: Provide mechanisms so that users can get data directly from remote data providers when the library provides access to those providers through memberships, licenses, and purchases. An example of such a service is ensuring that students and faculty at an ICPSR member university can get data from *ICPSR Direct*. This level of service does not require local storage of data in a data library, but facilitates data passing through directly from a vendor to users. Such pass-through services are an increasingly important part of providing data services.

Skills: Communicating authorized computer addresses (IP numbers) to vendors is currently a standard way of authenticating access to remote data. Another way involves distributing and helping users install client software that connects to the vendor's server software. Managing passwords is sometimes a part of such services, but may or may not involve computing skills. Maintaining proxy servers or Virtual Private Networks (VPNs) may also be part of the mix of authenticating users as being part of a subscribing institution. In the near future, other mechanisms of managing authentication (e.g., Shibboleth[1]) for access control and resource sharing will become more prominent. For the data library that is part of a larger library or network-based organization, such skills will probably already be part of the existing infrastructure for managing subscriptions to electronic journals, indexing and abstracting databases, and other leased digital content.

Activities: Much of the activity to provide this service is not computing intensive at all. It includes compiling and maintaining lists of computer addresses, emailing or otherwise transferring the lists; obtaining software from vendors; communicating with vendors; and so forth.

Level 2 Computing Services: Computer Consulting

Services: Providing data consulting at a staff computer is one way to avoid the costs of maintaining public service computing facilities. A staff computer, perhaps in an office or other secure area, could have data, statistical software, format conversion software, and other tools useful for finding, locating, downloading, and using data. Designated staff with proper training could provide instruction, demonstrations, and hands on assistance at this computer. Instruction for small groups and classes can be provided if a computer that is portable or otherwise available for classroom use is available.

Skills: The computer used for this level of service will probably be configured differently from other machines in the organization and will therefore probably require special attention. Skills in hardware and software installation and maintenance will be needed. Service providers will need skills in using the selected statistical and other supported software. Familiarity with data and ability to work with documentation and raw data files as well as with specialized software are essential.

[1] http://shibboleth.net/

Activities: Although a wide variety of specific activities can be supported at this level, they all fall into one of two broad categories: instruction, and mediated service. Instruction can include demonstrations, classroom presentations, and consultations for individuals.

At a simple file-handling level, instruction in how to download data, how to transfer files, how to use local proxy servers, and so forth may be quite helpful for novice users. More data-oriented instruction might include demonstrations of how to open a raw data file and configure statistical software with information from a codebook, how to create a subset with a particular software package, how to convert data from SPSS to Excel, and so forth.

Mediated service includes working directly with users helping them get the data they need and make it ready for use. This takes instruction one step further by actually working with the files the user wishes to use. This is more time consuming and more personalized and as such is a fairly high level reference and consultation service, but the amount and complexity of computing service is rather low. The particular mix of activities supported can be narrow or broad, though, to match the needs of users and the resources of the organization. For example, consultation service might not be made available for all datasets from all sources, but might be available for one database or data-source or provider that is used by many.

The administrative structure of organizations sometimes erects service barriers between resources and users. For example, it may require special agreements to make business data acquired by the business school available to social scientists who are not part of the business school. An important responsibility of data libraries is to remove or minimize barriers. This may translate into working through the existing bureaucracy to forge administrative agreements, partnerships, licensing agreements, and so forth.

In some organizations, the data library may be the first to deal with service issues that will later confront other areas of the institution. At many libraries, it has been data services that have pioneered dealing with providing access to large files, plotters, specialized software, and so forth.

Increasingly data libraries will need to deal with different kinds of data, not just social science data; e.g., GIS, E-Science data, and linguistic corpora.

Level 3 Computing Services: Providing access to data files in a local collection

Services: Once the data library establishes a local collection of data, the most basic and essential service it has to offer is a way for users to get access to those files. Simply providing mechanisms for users to use files mounted on local or networked computers, or to download or otherwise move files to their own storage space is sufficient for this level of service.

Skills: At the low end, no computer skills are needed to simply hand over CDs and DVDs to the user. A moderate expansion of skills can make data through “cloud” services, FTP, and similar networked-based solutions. With personal-computer skills, providing stand-alone computing facilities is possible and with basic Internet service skills, data files can be made easily available by ftp or http.

Activities: This level of service can be provided in a number of ways with different levels of computing support and skills. Examples of ways of providing access to data files include:

- Delivering files on portable media;
- Providing copies of data files on stand-alone publicly-accessible personal computers with files ready for access or transfer;
- Providing simple network file access via FTP or HTTP or cloud-services;
- Providing access to files through an institutional repository or digital assets management system.

Many data files are distributed with contractual conditions restricting who may use the data and specifying the purposes under which they might be used, such as for teaching or scholarly research. In many instances, for example, data will be leased to a university with the restriction that only members of the university may use the data. A data service providing access to these data over a network will need to ensure that proper authorization of users is enforced and that users are informed about the contractual conditions in which these data may be used.

For provision of public service computers, networks, and internet-accessible computers, the skills needed include those of maintaining the computer hardware, its security, its operating system, and software. Even for simple, stand-alone personal computers, these are specialized tasks that require experience, ongoing training, and regular attention. A public service PC will have all of the support issues that the staff PC in level-two service has plus more complex issues associated with publicly available computers. Public service computers for data may require different software and hardware configurations than other public service computers in a library. As mentioned earlier, data-products designed for single-person use may present a technical; see Level 4 Computing Services.

Many libraries are establishing Institutional Repositories (IR)[2]. While these are often set up to enable researchers to deposit copies of articles published in journals and other research outcomes, the basic functionality of file deposit, file management, and file retrieval can be used for larger data files and accompanying metadata files. The data librarian will need to work with the

[2] For an overview of IRs and pointers to IR software see Institutional Repositories, Tout de Suite by Charles W. Bailey, Jr. (2008) http://www.digital-scholarship.org/ts/irtoutsuite.pdf

managers of the IR to establish IR file format standards, file naming, and so forth that are appropriate for data files and metadata files.

Similarly, libraries that have other kinds of digital assets management systems may be able to use these for storage, management, and user-retrieval of data files and their metadata.

Although names like "institutional repository" and "digital asset management" imply some sort of long-term preservation, no system becomes OAIS compliant without explicit planning and design. If long-term preservation of data is part of your service, you should explore the OAIS compliance of any existing infrastructure you wish to use.

Level 4 Computing Services: Providing computing for data with bundled software

Services: For data products that come bundled with special software that make the data accessible on a personal computer, the data library can provide one or more personal computers with the software and data installed and ready to use. Providing reference service to help users understand and use the software and the data is desirable and providing computing services for transferring their work to portable media or other machines is essential. Although less common than a few years ago, such products do still exist.

Skills: Ability to install the special software and data and maintain the computers is essential. It is desirable to have staff with enough familiarity with the software and the data to provide instruction and help. As with level-three computing service, this level requires the skills to maintain public service computers. It can be a complex matter to manage public machines with many users with software that is designed to work on a machine with a single user with administrative privileges.

Activities: Installing software and data; maintaining public service computers; providing support for users.

Level 5 Computing Services: Data conversion and subsetting

Services: Most social science data files do not come bundled with special software; they require statistical software for use. The data library that is not prepared to provide statistical software computing support may find general-purpose conversion software to be a reasonable alternative that provides users with many useful features. Such software[3] can read many data file formats, including plain ASCII, and convert these to other formats and create subsets. Making such general-purpose data management tools available on a public

[3] for example, Stat/Transfer http://www.stattransfer.com/.

service computer can provide an easy way for users to get the data they need in a format they can use with little difficulty.

Skills: Similar skills to level-four service with the addition of familiarity with the particular software supported and a general understanding of data structures and terminology and ability to use data documentation.

Activities: Similar to level-four with the addition of providing support for the supported conversion software, and helping users understanding data formats and use data documentation.

Level 6 Computing Services: Providing statistical computing facilities

Services: The data library that wishes to go a step beyond level-five service can provide computers with popular statistical software so that users can load data, do basic data manipulation such as merging and splitting data sets, and even do statistical analysis. This could be done on a small scale with a single public service computer -- essentially extending level-two computer consulting to providing a place for users to do their work.

Providing statistical software and computing resources can be a significant way for a data library to remove barriers between users and data. This can be done without providing statistical analysis services. See, in Chapter 12, the sections on Level Five Reference Service: Data Extraction Services and Level Seven Reference Services: Comprehensive Data Analysis Services.

Another alternative is to provide a web-based service using software such as Nesstar, SDA, or Dataverse.[4] These provide a range of data collection services covering levels 3 and 5 as well as analytical services for level 6. This kind of service differs from providing *general* statistical computing services, which provide an environment for users to work with *any* data file because these services provide specific services that are supported by the software for those files that the data library pre-loads. There is a higher up-front labor cost to providing access with this model and less flexibility for the user than in a general statistical computing service model. Web-based services can provide excellent, extensive, standards-based level 6 service to users. This is likely to become an attractive alternative for many data libraries.

Skills: Staff will need all the skills of level-five service, plus an ability to use the supported software and help users with their data management questions. The ability to provide statistical consulting is optional. If many users are likely to use such a service, the data library might consider managing a statistical computing lab. This would require skills in managing a computing environment for multiple

[4] Nesstar (http://www.nesstar.com/), SDA (http://sda.berkeley.edu/), Dataverse (http://dvn.iq.harvard.edu/dvn/)

users including the administration of user accounts and work space, maintaining hardware, performing software installation, managing product licenses, and more. This is not a trivial undertaking.

Activities: Similar to level-five service with the addition of support for statistical software and data management.

Providing a computing environment for statistical computing requires system administration and is a substantial overhead to a data services. These services may be outsourced to another department that provides contract system support. Regardless of the administration, the service plan needs to determine the right scale of computing resources to match the size and complexity of the data to be analyzed, the skills and needs of data users, and the size of the data user community. Models of activity could range from providing a publicly available, fairly powerful PC with one or two statistical packages and a few data sets, to a small computing center or lab with shared disk and user accounts. At least one method of delivering the results of analysis to users is essential. Possibilities include high-end color printers, portable media, FTP, and cloud-based services.

In planning the appropriate activities it is important to anticipate how the service will scale if demand increases. A staged approach to growth can provide valuable flexibility. When setting up a new data service, it may be appropriate to "start small" (e.g., providing a single public service computer) to see what kind of use it gets. In our experiences, providing such services encourages greater demands. Be prepared.

Level 7 Computing Services: Statistical analysis

Services: The computing support for statistical analysis is much the same as for level-six service: providing hardware, and statistical software. The real cost of this level of service is not the computing support, but the staff support (see Chapter 12 "Level Seven Reference Services: Comprehensive Data Analysis Services.") Such services might be divided across different units within the library, or even in other departments. For instance, quick fact retrieval might be done on a PC in a library, while a data extraction service might be performed on a Unix workstation over the local network and large analyses might be done on a central compute server or in a statistical computing lab managed by a social science department or academic computing unit.

Skills: Providing the computing service to support data analysis activities requires all the skills of level-six service, plus advanced analytical and statistical skills. This is beyond the scope of many data libraries.

Activities: As noted in Chapter 12, this is an advanced service that most data libraries will not offer. In considering whether or not to do so, the data library should consider the following kinds of activities likely to be required: large scale data file management, data analysis, statistical consulting and choice of statistical methodologies, delivery of reports, graphs, statistical tables, and maps.

Collection Strategies for Data Services

Collection Development

The terms "collection development" and “collection management” are routinely used in library environments, but has applicability in any setting where data files are stored for use by others. In this chapter, we examine issues specific to collecting, managing, and providing access to data files.

These "strategies" are more specific than the general discussions in Chapters 8 and 10. Where those chapters gave the overall goals and ways to define objectives for collections, this chapter revisits some of the same issues and some new issues with an emphasis on strategies for reaching objectives.

The following is a list of issues that should be addressed by those responsible for a collection of data files. The one assumption implicit in this list is that, in a university environment, a campus should treat its data files as a collection accessible to all authorized users.

Collection Development Policy Statement

In OAIS, setting policies is part of the management function of a long-term preservation archive and in TDR there are several metrics that require explicit statements of missions and policies and preservation plans. TDR metric 3.1.3 specifically requires an explicit collection policy. (OAIS and TDR are described in Chapter 9, Preservation.) To the extent that a data service take responsibility for long-term preservation of any data, it should, therefore, have a written collection polity statement.

Many of the issues identified below could be included in a policy statement for the development of a collection of data. A collection development policy may also define or help define service policies. There are several advantages of writing and distributing such a policy statement:

- It helps the service providers address issues and make decisions before they become problems.
- It is useful as a means of conveying to users the extent (and limitations) of services and collections.
- When the structure of the organization includes collaboration among different units in an organization, it provides a place to specify those agreements.

A policy statement may be as important for specifying the *limitations* on services and collections as for specifying which services and collections are provided.

Similarly, a policy may specify obligations of cooperating departments and limits to those obligations.

The collection policy and the actual practice implementing the policy should be reviewed periodically to insure that the collection still reflects the needs of its users, unwanted files are not being acquired, and needed files are not being missed.

A sample list of items to include in a collection development policy, "Elements of a Collection Development Policy" by Dan Tsang is provided at the end of the chapter.

Archive or Library?

As discussed in Chapter 11, it is important to recognize the different roles of a data library and a data archive. An *archive* is a collection that includes items that may not be available in other collections and that are intended to be preserved indefinitely. Typically, a *library* differs from an archive in that none of the materials in the library collection are considered to be the "copy of last resort" (the last available copy). A library copy of data might be kept only as long as it is required by local users and might be discarded if no longer needed.

As noted in the Preservation Chapter, in the digital age, access and preservation are inevitably connected. Even traditional “libraries” that have not previously accepted long-term preservation responsibilities should examine their place in the research data life cycle and determine appropriate tasks that meet the needs of their designated communities for the long term.

In practice, any particular data service may find it necessary to have a policy that allows it to treat its collections with flexibility – applying archival procedures to some materials and library procedures to others. A data library might, for example, choose to use a "library" collection policy for most of its collection, but occasionally accept the donation of appropriately documented, locally-produced data. Similarly, a data archive might, as a convenience to its users, include in its local collection data files that are readily available and archived by another institution. A data library might choose to apply archival techniques to data in its collection because it has no guarantee that the producer or distributor of the data will provide long-term access to the data. Such policies and exceptions to policies should be as specific and as explicit as possible. Flexibility does not obviate consistency.

One way to approach this question day-to-day is to ask yourself several questions when you obtain or are offered data. If the answer to any of these questions is "no" you may want to archive the data yourself.

- Will I be able to get another copy of this data later from the same or another source?

- Will the distributor continue to distribute these data indefinitely?
- Is the distributor an archive? Is it taking archival responsibility for this data?

What formats will be used to acquire, store and deliver data?

There are many media available for transfer and storage of computer-readable information and these evolve and change quickly. The most common removable media today are Compact Disks, DVDs, and flash drives. For extremely large data collections transferring data on removable hard drives is not uncommon. New technologies emerge rapidly. Data are often transferred using HTTP (and less often today, FTP), cloud-based file-sharing services, and even electronic mail. Data archives sometimes make files available over small local area networks, with NFS-mounting and AFS-mounting of remote disks, and other shared-file system solutions. There are different standards for writing onto and transferring with these media: densities, block sizes, character encoding schemes, character and binary FTP, and so forth.

The question of media can be divided into how data will be:

- Acquired
- Preserved
- Delivered

It is not necessary that the same format or medium be used for all of the above activities. The OAIS standard (see Chapter 9) explicitly separates these into three “packages”: a Submission Information Package (SIP) for acquisition, an Archival Information Package (AIP) for preservation, and a Dissemination Information Package (DIP) for delivery. For instance, you might acquire data from a vendor on DVD, but store data locally on a file server. In a more complex situation, you might acquire data from a vendor by downloading files (delivered over the Internet), store files on a public service personal computer for immediate use by users, use separate networked machines and disks for preservation (backup copies), and allow users to take data to their own machines by writing data to DVDs or memory sticks.

The particular mix of media that one chooses to use, will, of course, depend on local hardware, software, computing environment, and needs and computing environment of local users. It pays to be inventive in this area and not assume that if you acquire data on one medium, you must store it on the same medium or that users will have to use that medium. While local conditions may limit the range of options open to any particular data library, it is best to keep as many options open as are practical and useful rather than be limited by a particular technology.

An interesting aspect of this question arises when one looks at data acquired on DVD. Once acquired in this format, is this the best format to distribute the data? What options do you have locally for disseminating the data? You might find that the medium for delivering the data effects the services you are able to provide or the services that users expect you to provide.

Preservation should be addressed also. How will the collection be stored for day-to-day use and will provisions be made for backup or safety copies? Is it necessary to provide long-term preservation of data in your environment? What preservation metadata will you have to create and store securely?

A data-library may choose to use several media and formats, but it is still important to define:

- Which formats are used and which are not;
- For what purposes each format is chosen;
- What are the service implications of each format choice.

In what formats will documentation be collected and maintained?

Aside from the question of *data* storage and dissemination addressed above is the question of the storage and dissemination of *documentation* and related files. While you may come across older studies for which all or part of the documentation is available only in paper format, increasingly you can expect documentation to be available as computer files of some sort.

ICPSR distributes codebooks in PDF and DDI formats. Other common kinds of machine-readable documentation files include data dictionaries, frequency files, "data map" files, SAS and SPSS control cards, and even SPSS export files. Data that are distributed on CDs and DVDs and other media intended for use on a microcomputer, may have documentation that is available in machine-specific formats or software-specific formats. Examples include:

- Documentation that is viewable only through use of a particular piece of proprietary software that runs on only a Microsoft Windows machine.
- Documentation that is part of the data file itself (e.g., dBase files, Excel files).
- Printed codebooks converted to computer-readable scanned image formats that can be viewed on screen with appropriate software but that cannot be edited or easily searched.

It is also important to be sure that you have complete documentation. Sometimes you need more than one format to obtain complete documentation. Issues to address in deciding on formats for collecting documentation include:

- If codebooks are available in more than one format, will you prefer one over the other, or collect more than one format?

- If computer-readable dictionaries and statistical software "control cards" are collected, which statistical packages will the data library support?

- Is the format chosen appropriate for the intended use of the codebooks and do the primary users of the codebooks have adequate access to them?

- Traditional “codebooks” were intended to be used directly by humans, but more contemporary data documentation is intended for machine processing before being used by humans or statistical software.

In considering codebooks, it is important to analyze the services you wish to provide and how codebooks fit into those services. To use just one example, codebooks can be useful tools beyond their utility as documentation for a particular study. For instance, codebooks can:

- Serve as models of how to conduct survey research;

- Be used as a reference tool when the codebook includes frequency tables.

Will the collection be reactive or proactive?

A reactive collection policy would authorize acquiring a data file only when a user requires it -- the collection builds in reaction to the users' specific requests. A proactive collection policy would authorize acquiring data in anticipation of future needs of users, much the way a traditional library collects books as they are published in anticipation of demand for those books. These options are sometimes expressed as acquiring information "just in time" (reactive) or "just in case" (proactive).

This distinction may not always be clear, however. For instance, an otherwise "reactive" collection might regularly acquire the *American National Election Studies* as they are released without waiting for users to request the newest release of data. An otherwise reactive collection might acquire all available parts of a study when only a single part is actually requested (e.g., acquire the "work-history" file of *Panel Study of Income Dynamics* when a user requests only the "family" files).

Shared Acquisitions and Remote Access

As discussed earlier, remote access to data is becoming increasingly common and an important part of the mix of services offered by data libraries. In addition to relying on existing remote access solutions, it is also possible and sometimes quite practical to acquire data at one location and have it available at several locations. Such a situation could be as simple as several graduate schools at a university cooperating in their acquisition of data and sharing data access. A more complex arrangement might involve a Federated Membership in ICPSR that allows for sharing among institutions.

Who will select data to add to the collection?

This question may arise particularly in an academic library setting where there are several selectors or bibliographers who have traditionally purchased materials for particular disciplines. Should the responsibility for selection and ordering be in the hands of several selectors or one data-selector? To help decide the answer one might want to look at the training required for accurate and efficient selection and the volume of selection and ordering likely at the institution.

Funding

There are, it seems as many models for funding data purchases as there are data libraries. For example, funding for data purchases might come from a library collections budget, but funding for ICPSR membership might come from an academic dean. Within a library, there could be a data-budget or data could be purchased out of subject or discipline funds. Expensive purchases and licensing agreements may be subject to additional review and approval processes.

Fees for Service

An issue related to funding is whether you will charge for services. You need to address questions such as:

- Will fees be charged at all?
- How will the amount of fees be determined?
- If fees are charged, will all users be charged?
- Will different users be charged different fees?
- For which services will you charge?
- How will the income from fees be used?

Addressing the possibility of assessing fees may open possibilities for providing services for which there would be no funding, or might help clarify priorities. Knowing the costs of providing services may also help open discussions for funding partnerships within the university or other larger institutional environment.

Elements of a Collection Development Policy Statement for Data Files

The elements listed below are adapted from the article "Academic Libraries and Collection Development of Nonbibliographic Data Files" by Daniel C. Tsang, *IASSIST Quarterly* 12 (Fall 1988): 52. Along with the examples of collection development policy statements and criteria for accepting data for deposit in *Data Basics: A Reference Manual,* these elements can serve as a guide for building your own data collections.

1. **Subject Scope:**
   Relevance to research and instruction at the university.
2. **Temporal Domain:**
   Is the time period covered of relevance to research and instruction at the university?
3. **Spatial Domain:**
   Is the region or location covered of relevance to research and instruction at the university?
4. **User needs:**
   Does the user need to manipulate data or just use manipulated data?
5. **Uniqueness of data:**
   Are the data available in print format? Is it necessary to get the data in computer-readable format? Are they only available in computer-readable format?
6. **Currency of data:**
   Are the data from an ongoing study that will be quickly superseded by more recent revisions? Is it important to acquire quarterly updates or just annual cumulations?
7. **Confidentiality of data:**
   Is there a need to restrict personal or proprietary information in the data set? Will acquisition violate privacy?
8. **Physical format:**
   Is the medium compatible with available hardware?
9. **Software compatibility:**
   Are the data accessible by software currently available, or are they software dependent?
10. **Documentation:**
    Are the data supported by adequate documentation?
11. **Data quality:**

Are the data sufficiently "cleaned" so that the data set can be added to the collection without further processing?

12. **Access:**

    Is the data set accessible to all users? Are there any restrictions? Is it accessible online?

13. **Producer reliability:**

    Is the distributor/producer of the data reliable? Are its products well regarded?

14. **Historical importance:**

    Is the data set worth preserving even if use is limited in the foreseeable future?

Acquisition Strategies for Data Services

Chapter 11 discussed collection issues; this chapter addresses issues of acquiring data once the collection decision has been made. Acquiring data entails a number of tasks, such as deciding what data to get, determining from which vendor to order the data, picking a format in which to receive the data, and choosing how to have the data transmitted. For each of these tasks and others we discuss below, a number of choices may exist. This chapter reviews the categories of choices that data librarians confront when acquiring data. The number of choices will likely vary from data order to data order, even when dealing with the same data distributor. Ideally, the data that are acquired will match the research needs of your patrons and the technical requirements and preferences of your institution. In some instances, no options will be available. Both OAIS and TDR say that the archive and the information producer should negotiate an appropriate SIP (Submission Information Package). Sometimes, however, no negotiation will be possible or successful and the the set of choices will not include any that you prefer. Nevertheless, if the data are still required, you may have no choice but to accept less-than-ideal format or delivery method. In these instances, the data librarian will still have to deal with the data when it arrives. Ways of coping with this situation will also be discussed in the chapter.

Data Acquisition, the Collection, and the Reference Interview

Although this chapter focuses on the acquisition of data, acquisition strategies draw on and complement reference strategies and collection strategies. Ideally, a data service integrates these related functions to ensure consistency, accuracy, efficiency, and reliability.

Many of the choices discussed below will most likely surface during the data reference interview or consultation. Fuller coverage of the reference interview is provided in Chapter 16. The material below focuses on the content needs of the researcher arising out of the reference interview.

Instrumental to a thriving collection is a collection development policy. This statement should outline the criteria used to assess items for inclusion in an established data collection. Individual acquisitions should be guided by such policy. Other criteria, however, must also be considered when evaluating data for acquisition. Chapter 14 considered broad policies that deal with the overall collection whereas this chapter deals with decisions that arise when considering acquiring a particular data set based on specific research needs.[1]

[1] This distinction may be more obvious by comparing the collection development elements listed in chapter 14, in particular, the reprint of the article "Academic Libraries and Collection Development of Nonbibliographic Data Files" by Daniel C. Tsang, and the discussion presented in this chapter.

Options and Choices

The data librarian is often faced with an almost bewildering number of choices when deciding to acquire data. Below, we address several common options and choices for acquiring data.

Which Source?

One common challenge is deciding among more than one distributor or vendor of the same data.

Unlike books and periodicals, most data are not “published,” listed in catalogs of publishers, or even sold. Data collected by the U.S. government, for instance, is often available from multiple vendors since, for the most part, vendors can redistribute such data without having to pay the government royalties. This creates a situation where different vendors acquire the same data from the government, repackage the data differently, and then offer their product for sale or lease.

Often vendors will "add value" to the data. For instance, *CPS Utilities* from Unicon Research Corporation provides multiple years of the U.S. *Current Population Survey* in a convenient package that makes it easier to work with variables across years. Several commercial products incorporate public domain, basic economic data collected by the government with data from the private sector and add software that simplifies the selection and analysis of the data.

Data from more than one source may appear to meet your patrons’ needs. In such cases, the reliability of the vendor and the authenticity of the data may be worth investigating. Knowing the original source of the data that each vendor uses is also important. For instance, are the data new or old? Economic data are often revised and corrected; does the repackaging include these changes and are they documented? Does the vendor specify the source at all? Or does the vendor use vague terms (e.g., “BLS” or “Census Bureau” rather than a specific survey or report) to describe its provenance?

In the case of international data, source information is particularly important. Are the data from an official statistical agency or are the data estimates from another country or perhaps imputed from other data? In the case of international trade data, are imports and exports all based on a single country's data collection or are they based on both trading partners?

For instance, the IPUMS projects at the University of Minnesota do an excellent job of making public-use census microdata data easy to use across years, but in doing so they had to make decisions about how to combine datasets and they had to re-code variables to make them match across years. This puts an additional burden on those who use these data to ensure their results are compatible with the original data and their analyses are accurate and well documented.

As another example, if you try to identify the source and authenticity of the facts in the *CIA World Factbook*, or compare variables from the International Monetary Fund with the same variables from the OECD, you may find inexplicable variations and undocumented assumptions.

It is also important to evaluate how clearly the data product communicates to the data user the sources, changes, enhancements, comparability, and reliability of the data.

Acquire a copy or get access?

When looking for a source of data, you may encounter a service bureau that provides access through a subscription fee or a fee-for-data basis. These distributors charge a fee for access rather than offering the option of acquiring and keeping the data. Some economic and financial data in particular has been increasingly offered in this way in recent years. *Datastream*, a large international financial and economic service bureau, operates this way. Statistics Canada's CANSIM service, a large time series database, is available openly on the Internet but imposes a fee for the retrieval of each series.

There are advantages and disadvantages to each option.

Cost is a likely to be an important influence on choice. A pay-per-use or pay-per-data point service may be unacceptably expensive if many users will be using lots of data, but may be less expensive if only a few users will be using small amounts of data. If your service model incorporates or enables passing costs along to users, using a service bureau may be attractive when you face unknown or unpredictable demand for data.

It may be easier to use a subscription service for highly volatile or frequently updated data. Ensuring long-term access to data is rarely possible with a subscription service. Reliance on a service bureau’s interface to the data is also a service issue that needs to be considered carefully. Is it easy to use, accurate, and so forth?

Get all data files or just selected ones?

Many studies consist of multiple data files. For instance, data from the U.S. Census is organized in individual files for each state (organized by spatial units); the Canadian *General Social Survey on Time Use* has one file for diary-level data and another file for questionnaire data (organized by different units of analysis), and the U.S. Department of Justice's Immigration and Naturalization Service releases separate files each year for all immigrants admitted to the U.S. (organized by time). An option may exist to get the entire collection of files belonging to a study or only a part of the study's file collection.

Files may also be organized by content variation. For example, the U.S. *Census of Population and Housing* has several different "Summary Files" each of which contains different levels of subject and geographic detail. If income data are needed, some specific files will need to be acquired; if income data are not required, some files may be omitted. The level of geographic detail provides another option as different files cover zip codes, blocks, and states.
Overall, such choices should be straightforward based on the content needs of your patrons, but attention should be paid to ensure that the data ordered do indeed contain the content required.

The choices will also depend on your collection and service policies. For instance, if a user wants county level census data, but you anticipate other users will want block level data, you might acquire both at the same time.

Get complete data file or a subset?

In some cases, decisions that have already been made about your services, data delivery, and collection will guide the acquisition choices you make.

Perhaps a subset of a very large file can be extracted over the Internet where only the variables and cases required by the patron are acquired. For instance, a patron may require census microdata, but only for one state and only for those individuals with high incomes, and from these cases the user may require only a handful of variables. A file much smaller than full microdata file with all variables for all states would meet this particular patron's request. Note that, in this example, there is an intersection of collection policy and service policy. Perhaps, for instance, your collection policy precludes adding small extracts of data to your collection, but your service policy encourages helping patrons get subsets of data from the Internet. Or, perhaps your service policy would encourage getting the subset for the user at the time of request, and your collection policy would indicate that the entire data file should be added to your collection so that it will be available locally for future requests.

Which physical format?

Data may be available in different physical formats such as "card image" or "logical record length"; hierarchical or "rectangular." Most statistical software can handle different physical formats, so this choice is often based on the preferences of your patrons or on the availability of statistical software on your campus. Increasingly we see data available in ready-to-run form, formatted for specific statistical packages. Although any of several formats may be technically acceptable, your users may have a preference for a format they have used before. Again, the task here is to identify what format options are available and to make choices compatible with your service plan and your patrons' desires.

Which Data Documentation?

The quality, kinds, and format of data documentation differ greatly among data distributors. Occasionally a data distributor will offer data with more than one kind of documentation. Choices of documentation are often complex and correct choices are always crucial to the accurate use of the data. It is particularly important to ensure that you get complete documentation.

Different editions of the data may exist and each edition will likely have its own documentation. Naturally, a major concern is to ensure that the documentation matches the data that have been acquired!

Furthermore, the same edition of a study may be available in more than one format. For instance, a data file might be available in raw data file format and in SPSS system file format.

It is, of course, essential that you get the documentation that match the data you get, but you may have other considerations as well. In recent years, many data distributors have begun making documentation available as digital documents (e.g., Adobe Acrobat PDF files). With any digital documentation, one must consider long-term access to the documentation. A lovely codebook written with ancient word processing software that is no longer available or no longer runs on contemporary operating systems or hardware is as useless as no codebook at all. A “set up” file for and old or unsupported version of statistical software is not very helpful.

It is important when acquiring documentation to determine if it is complete. This is something to investigate particularly when a survey instrument was involved in collecting the data. Sometimes, documentation of the data file will seem complete, but no copy of the survey instrument is included or a separate document is required to obtain the instrument. One may have enough documentation to read and process the data and later discover that a researcher needs some additional documentation in order to interpret the data. (See Chapter 6, “Reading Data Documentation,” for more details about what to look for in documentation.)

Sometimes the option exists of getting data as a SAS dataset or an SPSS system file, or in other software-specific formats. Researchers often value these formats because they can save much time in getting the data ready to analyze. Specifically, the patron does not have to go through the tedious job of defining the data for the software. This is very handy if the kind of analysis to be conducted is readily available in the software for which the file-ready format exists and if your patron can use this software easily. While much of the information normally included in data documentation is often incorporated in such files, additional documentation is almost always necessary. One rather common situation is to find a data file with a codebook and an SPSS system file. Although the SPSS file is technically ready to be used, the codebook is still essential because it contains information that the SPSS file does not; e.g., the complete text of survey questions, survey question skip patterns, survey methodology, the universe and sampling methodology, explanation of weight variables, and so forth.

Knowledge about how statistical-software-specific files are created can also be important. In some instances, these files may have incorporated assumptions about the data, or re-coded values, or sub-groupings of the data, or other changes or additions to the original raw data and these changes may not be documented in the original codebook.

Note also that there are many files called “data” on the Internet in Excel format. When finding such files, it is important to consider if these files are really “data” or if they are simply “statistics” as discussed in Chapter 1. We often come across government agencies, for example, that use Excel to format statistics for visual presentation. Such files are often difficult to handle as “data files” even if the aggregate statistics in them is suitable for analysis.

Archiving Data

In 2010, the U.S. National Science Foundation announced that it would begin requiring a “data management plan” as part of proposals for funding.[2] Other agencies (e.g., NIH, NEH, IMLS) and other countries have begun to follow this trend. The result is that grant-seekers are highly motivated to plan for the entire lifecycle of the data they gather. At many academic libraries, moves are underway to position the library as a participant in data management planning and, in some cases, in data archiving.

In addition, some researchers may seek to deposit data in a university’s institutional repository, or include data with a donation of their “papers” to the university archive. Some libraries may already have facilities for preserving data and may make this available as a service locally.

For these and other reasons, original research data from your home institution or from other researchers may be available for acquisition. It is essential when acquiring such data to ensure that the data are adequately documented, that the privacy of respondents has been protected, and that access restrictions, if any, are clear and consistent with your service and collection policies.[3] It is important that your collection policy spell out these requirements. A researcher who

[2] National Science Foundation. (2010, May 10). Scientists Seeking NSF Funding Will Soon Be Required to Submit Data Management Plans - US National Science Foundation (NSF). press release. Retrieved December 10, 2011, from https://www.nsf.gov/news/news_summ.jsp?cntn_id=116928&org=NSF&from=news

[3] For more information on data preparation and what to expect of documentation, see: *Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle, 5th Edition*. Inter-university Consortium for Political and Social Research, (2012) http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/access/deposit/guide/

wishes to donate her data to the data library deserves to have a clear, written statement of requirements for depositing data.

It is important to know if the deposited copy is the only copy of the data and if it is considered the copy of last resort

A service that a data library can offer is helping a researcher prepare data for deposit in a data archive. This could be a minimal service, providing instructional materials, or it could be part of a campus data management policy, or it could be a service of translating documentation into DDI. (See Chapter 17 for more about DDI.)

Cost

Costs may be associated with any or all of the above considerations. Some will be obvious. For example, some government data will be available directly as part of the government's depository program or by paying a small fee to a government agency. On the other hand, the same data may be provided from a commercial vendor with additional features but at a much higher fee.

Some costs are hidden or less obvious. For example, will the "free" data from the government result in an investment of time and other resources to convert the data or get the data ready for analysis? Will the easy-to-use commercial version be preservable or even usable in a few years?

Naturally, these can be complex considerations. Perhaps time is more readily available than dollars and spending time preparing free data is more of an option than spending dollars to purchase packaged data. The various costs will have to be weighed along with all of the other considerations discussed above.

A checklist of considerations

For any particular data acquisition, one or more options may exist, or there may be none. When options exist, the choices will often interact with each other. The best data may be the most expensive, for instance; or the most reliable data may be the hardest to use or acquire. In practice, each data acquisition seems unique, although there are patterns that make choices easier. Experience is an excellent guide. Here are some things to consider while building experience.

- Make sure that the data acquired meets your patrons' needs.
- Make sure the data acquired fits with your service plan.
- Make sure the data acquired fits your collection policy.
- Make sure the data acquired is technically usable in your computing environment.
- Your service plan and collection policy should include guidelines that balance access and privacy and these considerations should be important in your decision-making.

- Weigh the tradeoffs you have to make among different choices.

What to do when no data match user needs?

Sometimes (all too often, in fact!) the research request will be for data that no single, existing study will be able to address. This situation does not mean that no choices exist, but it does mean that the available choices are very different from those discussed above. Data service providers must recognize the options that are available and provide researchers with choices when it appears that no data are available.

Often, a search for data involves looking for a study or collection that matches multiple criteria. The reference interview can help identify the different elements of the data request, which elements are essential, and which elements are preferred or expendable. For instance, if a patron asks for monthly data for trade between two specific countries over a twenty-five year period, the reference interview might determine which of these four criteria (frequency, subject, geography, time coverage) could be dropped or modified if no data match all four requirements.

For example, one way to structure the reference interview is to categorize the variations in data that might be acceptable to the researcher:

- **Acquire a different *Unit of analysis*.** If monthly time series are not available, perhaps quarterly or annual series will suffice. If a researcher is looking for data on individuals and none can be found, maybe data exist for households.
- **Acquire a different *universe*.** The universe, which is a sampling concept, defines the inclusive membership of objects in the study about which data have been collected. If a researcher wishes to analyze all labor unions in the U.S. and such data are not available, will data on Southern labor unions or data on factory worker labor unions be an acceptable substitute for the universe? If a twenty-five year time period is requested, will a ten-year period be acceptable? If no data exists for people of all ages, will data for adults suffice?
- **Acquire different *variables*.** If an ideal data file for a study about economic inflation would contain a variable for "seasonally adjusted consumer price index for all urban consumers," would non-adjusted data or data for "urban wage earners and clerical workers" satisfy the researcher's needs instead?

It should go without saying, but we will say it anyway: do not overlook the opportunity to construct variables out of existing variables. For example, if a researcher wishes to examine “teenagers,” there need not be a variable that defines respondents as “teenagers” as long as there is a suitable variable (e.g., age, date of birth) that can be used to construct a new variable or select cases based on the “teenage-ness” of respondents. By the same token, if a researcher needs to analyze respondents based on their age, do not assume because there is a variable called “age” that this variable meets the user’s criteria. It could record single year ages or could it group people into age categories (e.g., 0-10, 11-20, 21-35, 36-50, 51-65, etc.). Always read the documentation.

Sometimes the variables needed for a research project exist in more than one study but not all together in a single study. One solution to this situation is to combine the data from different studies if the data are compatible. A typical example is to combine census data with other aggregate data at the same geographical level. Including data from the census file might allow determining rates or percentages based on the total population. For instance, one study may have data on all votes cast in an election by congressional district, but not have the total population. Combining census data with voting data would permit a calculation of the voter participation rate. Another example is to create a file for cohort analysis from survey data in which the same questions were asked but at different times and of different samples. This possibility exists with some of the *Canadian General Social Surveys* in which questions are repeated in a five-year cycle with different samples.

If no raw, digital data can be located for a particular research need, another approach to pursue is statistical summaries: monographs, journal articles, reference volumes and web sites, and other sources that may have statistical tables containing information relevant to the researcher's needs. Such statistical tables may also reveal the data sources from which the statistics were derived, thus providing another lead to a source of raw data suitable for analysis.

When no option exists that will satisfy the researcher's original needs, another tactic is to see if some existing, known data source contains data of interest to the researcher. In our experiences, we have found this a particularly useful approach to take with undergraduates. Often undergraduate researchers have fascinating ideas and research interests, but there are no data available that address those ideas directly or in precisely the way the student first poses the question. The enthusiastic student is often surprised to learn of other sources that are rich in data that invite equally interesting new analyses.

Data collections and data services

Throughout the late twentieth century, “acquiring data” was almost always synonymous with adding data to a data library collection. As noted in other chapters, in the twenty-first century, this is no longer a given. Today, users have direct access to a wealth of data. Government agencies, private organizations, commercial vendors, and others make data easily available on the Internet. Changes in computing environments have made high-density portable media common in most desktop machines so that increasing amounts of data are easily available directly to end users. Some of this wealth of data is free or inexpensive; some is very, very expensive. Some come with fancy software that makes use of the data easy (or at least easier than writing programs to access raw data files). Some come in rather raw formats that require fairly sophisticated users and sophisticated software.

All of these changes have one thing in common. They make it easier for individuals to acquire data directly from a vendor or distributor. That is a Good Thing and we welcome it.

It also means that when users come to the data library, one option that the data librarian has is to help the user acquire data without adding the data to the data library collection. This creates new service issues with old problems. Chances are that many of the issues discussed above about acquiring data for the data library are also issues for the individual. In fact, a data library may have more flexibility in acquiring data than an individual does.

It is not uncommon today, therefore, for a data library to provide a service of acquiring data for individuals. Such a service may help a user identify, select, subset, reformat, and physically acquire a data file. This “pass-through” acquisition is of growing importance to data service providers and the issues around it should be watched closely as technological issues and data marketplace issues evolve.

Reference Strategies for Data Services

“Reference service” has over a century of tradition in the library world and consequently, carries specific connotations. It is the business of connecting patrons with their informational needs, organizing information in a way that makes this easier, teaching patrons strategies to find what they need themselves, and pursuing this in the spirit of respect for each person and her or his question, regardless of what it may be.

Computer center “help desks” or “consulting services”, which have developed over the last twenty-five years, provide another service model, but specifically for technical information. This type of service also involves an interview process to determine the technical information needs of a patron and methods of finding answers to technical questions.

We find reference services in the library the preferable model of these two approaches. Reference services in the library has a stronger tradition and, we believe, better addresses the complexities and referrals necessary to provide data services for social science research.

In some sense, all of the direct patron services we talk about in this course are some variety of our expanded notion of “reference service.” Some of them are traditionally practiced in libraries, some are not, but quantitative social researchers may reasonably expect all.

Evolution of reference service in the library

At one time in the not too distant past, the day-to-day functions of a library reference service could be categorized (overly simplistically) as consisting of:

1. Helping users locate known items.
2. Helping users find facts.
3. Helping users find materials on a subject.
4. Teaching users to use bibliographic tools.
5. Referring users to other libraries, services, or information sources.

This greatly oversimplifies reference service by omitting the means used to accomplish these ends. A wide variety of methods are used in many libraries. These include:

1. Individual consultations.
2. Individual instruction.
3. Classroom instruction.
4. Creating handouts, guides, signs, bibliographies, "pathfinders," and other finding aids and instructional materials.

5. Building and maintaining a collection of reference materials containing facts, bibliographies, indexes, and other materials.

In addition, many library reference services provided at least two levels of service, which we may broadly categorize as "ready reference", and "consultation services." These are broad categories and may be described formally by library policy or may be implemented more informally based on the needs of library users and the resources (particularly time) of the library and the reference librarian. We can make some broad generalizations about these two categories, however.

"Ready Reference" service includes on-demand and point-of-use service. Most typically, ready reference is associated with the "reference desk" where library users go to ask questions and with "telephone reference" that provides short answers to questions asked over the telephone. Often such services have an explicit time limit imposed by policy or an implicit time limit imposed by the quantity of questions asked in a short time frame. Ready reference typically provides short answers to specific questions (e.g., a fact, a recommendation of which reference source to use for a question, a quick instruction in the use of a particular reference source).

"Consultation services" are characterized by their length, depth, and breadth. Often, consultation is available only by appointment and may consist of more than one meeting with a librarian or with one or more subject specialists. Consultations may consist of in-depth one-on-one instruction and guidance, or a more active participation by the librarian in the research project to the point of collaboration between the researcher and the librarian.

In addition to the above broad categories, many reference librarians also have explicit responsibility for formal instruction and the building and maintenance of the library's collections.

The technological changes in libraries have affected reference service in several ways. Most libraries have Online Public Access Catalogs (OPACs) and most have some sort of web presence. Libraries now have many more digital reference sources such as bibliographic indexes that are available on the Web or, in some cases, on CDs or LANs. Some libraries have multiple bibliographic indexes integrated (in a variety of ways) with their OPACs. Most libraries provide access to the web, often on the same machines that provide access to their OPAC or other databases. Many libraries now have access to non-bibliographic electronic resources on CD-ROM and DVD, including full-text publications, "statistical" publications (e.g., tables of numbers), and even numeric data.

Because of the addition of digital materials to libraries' collections, many libraries offer various levels of assistance with the use of computers and software. This is still a relatively recent development for libraries and many libraries are still struggling with the challenges created by these changes. Until the advent of digital collections, most library services were able to assume that their users understood the "user-interface" of the library collections -- stairs, shelves, card-drawers, books with tables of contents and indexes, magazines, and so forth. Now, libraries often offer assistance and even instruction in using a computer, using specific bibliographic software, using web resources, searching the web, printing and saving files to disk, and similar digital "user-interface" issues. In some libraries it is the reference services staff that provide this instruction and assistance; in other libraries, services are split between staff with different technical, subject, computing, and disciplinary skills. Most libraries rely on licensed access to online services of various kinds – the most common is services that index and abstract magazine, newspaper, and journal articles. For each different vendor or product, there is usually a different interface that may require instruction and assistance.

Many of the computing-assistance services libraries offer are similar to the kinds of service offered by computing center "help desks." These services tend to be technical, procedural, and often not related to content. Some libraries resist offering some of these more technical, computing-related services.

The trend toward providing assistance in the use of software and computers has made it easier for some reference services to incorporate more advanced kinds of computer assistance and advice into their services. Some libraries offer instruction and advice in using bibliographic management software. Many libraries are now offering or considering offering assistance with the non-bibliographic materials and software in their reference collections. These include statistical and numeric data, GIS files and software, data-extraction software, spreadsheets, and even statistical software. Many users require more than simple "user-interface" assistance with these types of materials and software. Some reference services are considering providing technical assistance that goes beyond the technical and procedural.

In this environment, *content* is as important as it is in more traditional reference service. What we are seeing, then, is a hybrid of sorts -- a new kind of service that combines the traditional values and skills with technical expertise and experience. Defining this hybrid service and its extent and limits in the social science data context requires considering more than technical and bibliographic skills, however. It also requires considering subject expertise and knowledge of social science methodology.

In Chapter 12, we outlined seven levels of reference service and we have mentioned various kinds of "reference" service in the other chapters in order to give a broad overview of kinds of data services. The intent of Chapter 12 was to provide a context for making broad service policy decisions. The rest of this chapter examines some specific reference services.

Data Reference Service

Preliminaries. Careful planning and preparation can make it easier to launch a data reference service.

A simple first step is to develop a small collection of data-reference tools. Since many such tools are now online, this “ready reference” collection of data-reference tools may start as a simple web page with pointers to good starting places. There are still printed guides and tools that are useful, however, and they should not be overlooked. These include older guides to data archives and collections and studies (e.g., *Inventory of Longitudinal Studies in the Social Sciences*. By Copeland H. Young, Newbury Park, Calif: Sage Publications, 1991), directories of information providers and distributors (e.g., *Information Industry Directory*. Detroit, MI: Gale Research), indexes to data collections and series (e.g., *Index to International Public Opinion* [1979-1999]). Westport, Conn: Greenwood Press), directories of statistical sources (e.g., *Statistics Sources*. Detroit, Mich: Gale Research Co.; *Statistical Sources for Social Research on Western Europe 1945-1995: A Guide to Social Statistics*. By Franz Rothenbacher, Opladen: Leske + Budrich, 1998), and useful compilations of statistics (e.g., “Statistical Abstracts” from different countries and areas[1]).

Additional printed tools to add to the collection include social science dictionaries and glossaries, guidebooks and sourcebooks to statistically-rich literature (e.g., business, finance, economics, demographics), bibliographies or inventories of locally produced data, and directories of local experts and statistical consultants and data producers.

Online tools are myriad and there are many libraries have good web pages of starting points for data reference. For example, see ICPSR’s own “Other Criminal Justice Sites” page[2] and its Substance Abuse & Mental Health Data Archive “Other Data & Sites” page.[3]

Data service providers can use the basic collection of reference materials and begin to provide a number of basic data reference services almost immediately. These include:

- Providing assistance in locating data to meet user’s needs.
- Identifying data not available locally.
- Providing help with understanding data and social science terminology.
- Promoting data as an important information resource.

[1] “National Government Statistical Web Sites” Indiana University. http://www.archive-it.org/public/collection.html?id=317

[2] http://www.icpsr.umich.edu/icpsrweb/content/NACJD/links.html

[3] http://www.icpsr.umich.edu/icpsrweb/content/SAMHDA/resources/links.html

Collecting information about local data use is another good way to prepare for providing reference service. Even a few informal chats with key researchers can lead to a better understanding of who is using data and who would like to use data. In colleges and universities, determining if there are class assignments that use or would like to use data can be a very useful starting point in planning for service. There may already be services or collections that overlap or complement the services you are planning. Many data librarians have found hidden caches and informal collections of data in their institution even long after the initiation of a formal data collection. And many organizations may already have a statistical consulting service or a survey research center that may provide data service or house a collection of data.

Creating a service plan is a very important preliminary activity. If the data service is designed to include a collection of data, a collection policy should also be created. Service and collection plans are useful for communicating to colleagues as well as data users and potential users what services and collections to expect. Writing the service plan provides an opportunity to formalize organizational decisions and support.

See Chapter 10, "Developing a service plan," for more detail on collecting information and writing a plan. See Chapter 14, "Collection strategies for data services," for additional information about writing a collection development policy.

The Data Reference Interview

Data reference begins with the reference interview. During this conversation between the data user and the service provider it is the responsibility of the service provider to determine the data needs of the data user and to offer assistance and choices to the data user.

There are as many different kinds of data reference interview as there are data users. We can, however, list several broad categories that recur with variations quite often.

1. The data user wishes to use a specific study or data set. Because of the nature of social science research, there are many studies that are well known to researchers. These studies may be analyzed and re-analyzed frequently. So, a simple and rather straightforward data-request is for a particular study. The service provider’s key tasks in such a situation are to recognize the study and determine its availability. Recognition of the study is often a matter of experience as many studies are referred to by common names and short hand and abbreviations. Appendix B provides a list of many commonly used names and acronyms that will help with this process. Accurate and complete citations to data make the process of identifying studies much easier.

2. The data user requests a particular study not because of an interest in the study but because of a subject interest. It is not uncommon for a researcher to assume that the best source of raw data to answer a particular research question is a known study. Sometimes, however, a different study may be as good as or better than the familiar source. It is difficult for the service provider to distinguish between question type 1, above, and question type 2 without additional information about the research question. One approach that may help distinguish between these two types of questions is to answer the request as a "type 1" but offer support for any other related data needs.

3. The user requests data on a particular *unit of analysis* or subject or data that contains a particular *variable* or examines a particular *universe*. This is, perhaps, the most common kind of request to reach the data service provider. The "decision chart" on the next page outlines a method of approaching this complex subject.

Checklist of Data Reference Services

Here is a checklist of what we broadly characterize as "reference services" for data.

- Provide instruction and promotion for services and holdings.
- Identify local data holdings.
- Make appropriate referrals to other service providers or data sources.

- Provide assistance obtaining data.
- Provide assistance interpreting codebooks.
- Furnish proper citations for data.
- Facilitate understanding of the structure of data files.
- Facilitate understanding of the contents of data files.

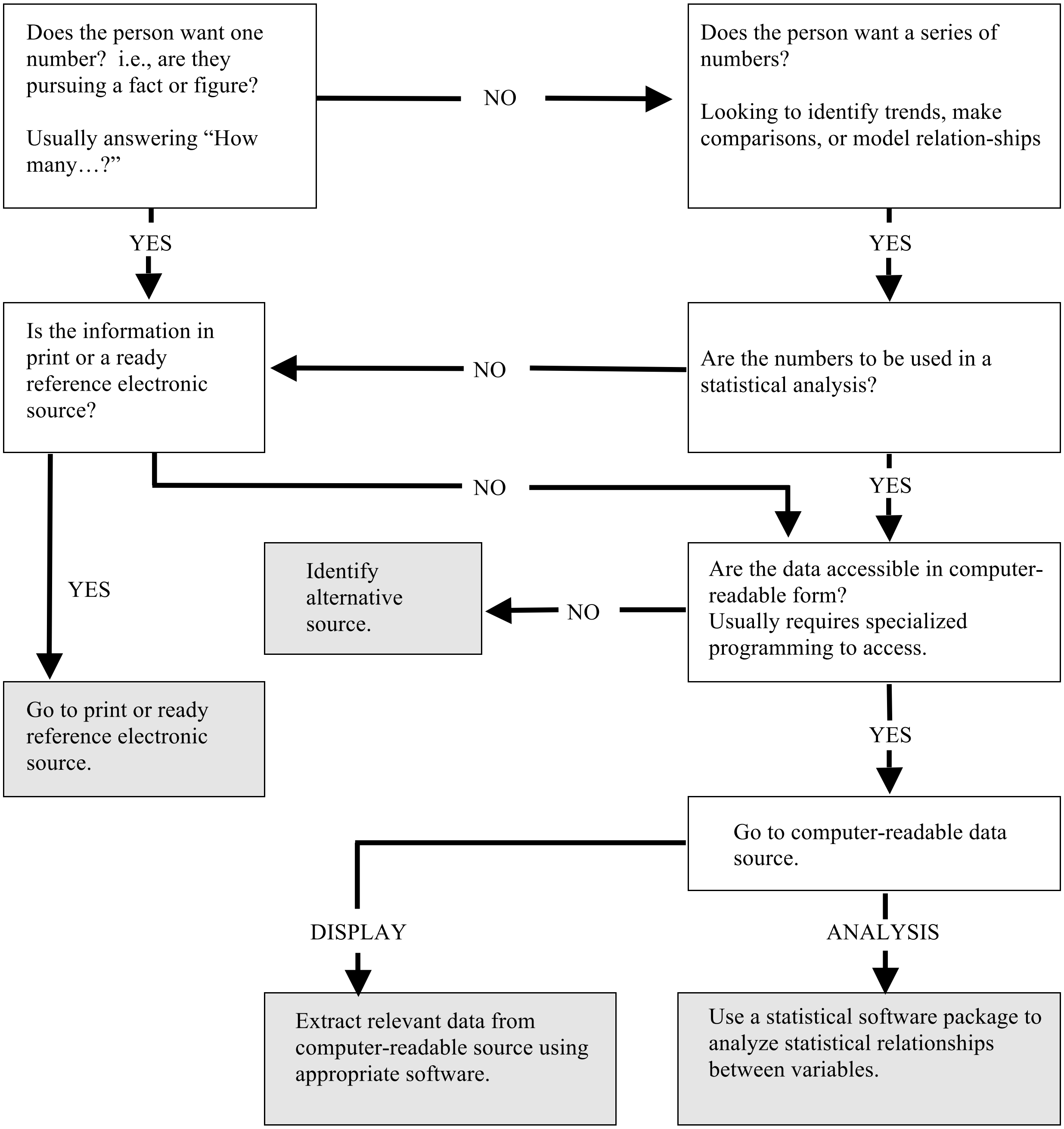
- Provide assistance with "self-service" computer programs, e.g., U.S. Census Extract; local or remote web extractors, etc.
- Subset data.
- Copy data or subset from one medium or machine to another.
- Move data from one software format to another.
- Provide computer-consulting services.

- Provide assistance with preparing data and documentation for deposit.

- Provide assistance choosing appropriate software for analysis.
- Provide assistance using statistical software.
- Provide assistance writing statistical programs.
- Provide statistical consulting services.

Choosing a Data Source

This decision chart can be used to assist reference desk staff in making informed referrals to data services and in becoming confident in working with data service providers.


Does the person want one number? i.e., are they pursuing a fact or figure?
Usually answering "How many…?"
NO
Does the person want a series of numbers?
Looking to identify trends, make comparisons, or model relation-ships
YES
YES
Is the information in print or a ready reference electronic source?
NO
Are the numbers to be used in a statistical analysis?
NO
YES
YES
Identify alternative source.
NO
Are the data accessible in computer-readable form?
Usually requires specialized programming to access.
Go to print or ready reference electronic source.
YES
Go to computer-readable data source.
DISPLAY
ANALYSIS
Extract relevant data from computer-readable source using appropriate software.
Use a statistical software package to analyze statistical relationships between variables.

16.8 – Reference Strategies

Data Basics

Access Strategies for Data Services

Providing access to data locally requires a number of organizational and technical decisions. Access strategies encompass a combination of organization (how you organize, list, and index your collection) and how users will actually obtain copies of or access to data files and documentation. Access strategies overlap with computing services (see Chapter 13) and delivery mechanisms. Together, these comprise information retrieval systems. In this chapter, we focus on supporting local data collections with a metadata-centered approach that frees the data library from the constraints of particular information retrieval systems and provides excellent long-term preservation of metadata.

Access Points

In designing an access strategy, it is necessary to consider the different access points available for social science data.

We can conceptualize the structure of social science data as a hierarchy, consisting of four levels of detail. Each of these provides a potential way of organizing a collection, a way of listing studies and their components, and a way of users looking for and retrieving data. They comprise a hierarchy of detail, with each level providing more detailed information about the data. These can be thought of as the possible levels of access, both in terms of how users find data and how they retrieve data.

1. *Study-level access*. A "study" is similar to a "book" in that it is a whole work with an author (or principal investigator). Typically, the data in a study have been collected with a single methodology and focus on a defined unit of observation and defined social units (time, space). A study may be described in a "bibliographic" citation with an author, title, broad subjects, "publisher," "distributor," and so forth.

2. *Dataset-level access*. Some studies consist of many data sets; in some cases providing access to the broad contents of each data set is one step more detailed than access to the study as a whole. For example: at the study-level, geographic access to the U.S. *Census of Population and Housing* might list only "United States" while data set-level access might list each state.

3. *File-level access*. Data sets are often composed of more than one file including one or more data files, codebook files, files of statistical software commands, and so forth. The relationship between files is very important because a user must know of, and be able to use, all the documentation associated with a data file and must not get the wrong documentation.

4. *Variable-level access*. In social science research, each item of data (e.g., age of person, income of family, consumer price index) is called a

"variable." Each variable in a study is a potential access point. The detailed content of a study is defined by its variables. Variable-level access to data is the holy grail of discovery in social science data collections. While it is nice to know that the U.S. *General Social Survey* is a public opinion survey of the United States (study level access), what researchers really need to know is that it has eight questions on gun control and twenty-nine on abortion (variable level access). The best access to variables is access to the full text of questions and answers in a survey or the descriptions of every variable in an administrative dataset.

Each study is different

While we can use the four levels listed above to easily conceptualize the different possible kinds of access to social science data, actually implementing a system to provide access is not as straightforward as the model might suggest. One reason for this is that studies are not all alike. The table on the next page outlines some attributes of very different studies. Creating a single information retrieval system that would optimize retrieval of information from all these studies would be difficult and possibly not even a wise thing to attempt.

Access Issues: Different studies have different requirements (table)

	Study	Data Sets	Files	Variables	Special Issues	Other
General Social Survey	annual, cumulative; special studies	1	2	2000+	sample size & representation of universe	bibliography of studies; book of tables
1990 U.S. Census	various STFs; PUMS, etc	50+ per STF; 2 or 3 PUMS	250 per STF	60 per STF	geographic coverage	paper copy census
Congressional Roll Calls	1, but 2 per Congress	202	400+	thousands		on-going series
California Polls	6 per year	200+	350+	thousands		on-going series
EuroBarometers	periodic, each with a "theme"	32+	175+	1000+		special topic to each, plus repeating questions
Current Population Survey	monthly with recurring topics	March is about 30+	March is about 50	March 600/year	exists over time but not time series	published report series
International Financial Statistics	economic time series	1	1	25,500	annual, quarterly, monthly	updated monthly, countries of the world
Survey of Income and Program Participation	multiple panels, multiple waves	9 per panel	9 per panel	4000+ per wave	8 units of analysis	Topical Modules done periodically
National Longitudinal Survey	panel study	6	50+		5 cohort groups	
Panel Study of Income Dynamics	panel study	9	40+	700+ for individuals	families & individuals	special samples (e.g., Latino, work history)

Limitations of Information Retrieval Systems

Today's data libraries use software to provide access to data. Indexes, catalogs, databases, and other kinds of systems can be built on one or more of each of the four levels of detail listed above. As with any system, as detail and quantity of information increases, so does the potential complexity of the system. The best information retrieval systems are designed to match the complexity and nature of the information content they are meant to manage. Mismatches between system design and information content can result in systems that are difficult to use and that deliver misleading or even false results. Most general-purpose systems have built in constraints (e.g., maximum length of title, limits on number of authors and subjects, etc.).

Many data libraries will have a choice among existing systems already in use to manage information retrieval. These include OPACs or their equivalents, Content Management Systems, Institutional Repository software, or off-the-shelf data-specific data management systems. Cost will often drive this decision, but, even given cost constraints, it is best to choose the best access service that fits available resources rather than to try to use an available system for services it does not support.

For example, a library OPAC built around MARC records is not an appropriate choice for variable level access, but works perfectly well for study level access. Conversely, using Nesstar to provide only study level access is overcomplicating an otherwise fairly simple problem.

It may seem that choosing an information retrieval system will inherently lead to compromises, but it does not mean that we can't have our cake and eat it too. The question is, what is the cake?

Metadata-Centered Model of Access

The “cake” in this analogy is Metadata. Perhaps it would be more accurate, however, to compare metadata to bread -- the staff of life of a data collection. When a data library has the best metadata, it can use and re-use the metadata with different systems, for different purposes, at different times. This means a data library that concentrates on creating and acquiring rich metadata does not have to be limited to a single information retrieval system. It can choose one for study level access and another for variable level access, if it wishes. It can choose one today based on what software is available or affordable and another tomorrow when new software becomes available or affordable. This fits in well, too, with the frameworks and phases of innovation discussed in Chapter 7, because, with a metadata-centered information retrieval strategy, the data library will be able to do today what it can while being prepared to do tomorrow what it cannot yet imagine. The data library will provide optimum service today and be prepared to meet new user needs tomorrow.

This metadata-centered model is ideal for any digital library project that needs to evolve over time to meet changing needs of users and changing contexts imposed by data vendors. It allows for changes in software, hardware, and operating systems over time. It makes preservation easier by freeing content from the constraints imposed by specific software or hardware. Because it is flexible, it allows the library to provide those services it can afford to provide and evolve services as new needs develop and new software becomes available.

The metadata-centered model can be easily understood in contrast to a software-centered model. In a software-centered model, software is acquired first and digital objects and metadata about those objects are created or otherwise obtained to conform to requirements or limitations of the software. In a metadata-centered model, the collection begins with digital objects stored in their most preservable format with metadata designed to accurately and completely describe those objects for long-term accessibility and use. The appropriate metadata are then selected, transformed as necessary, and loaded into software selected for a particular application.

By designing the system around the metadata instead of designing the metadata around the system, the data library can ensure that the metadata are complete and accurately describe the data. The metadata will describe a data file generally (bibliographic-type, study-level information), and technically (dataset-level, file level, relationships among files), and at the variable level (location, encodings, names of variables, full text of questions, etc.). From such a metadata file, the data library can extract (“cherry pick”) just the information needed for particular applications. The same set of metadata can be used in different ways – not just for information retrieval systems, but for data explorations, statistical applications, data visualization, subsetting, data delivery, and so forth.

For example with complete metadata available it is possible to extract from that metadata:

- bibliographic information to build MARC records for an OPAC.
- the text of survey questions to build full text indexes at the variable-level.
- information on variables to create a SAS job for extracting and reformatting a subset of data.

This model has three advantages.

1. No limits. It does not limit the data library to any particular software. It can use and reuse the same metadata for different applications, both now and in the future.

2. Built for the future. It allows the data library to provide new services that we cannot even imagine today.

3. Ideal for Preservation. The metadata-centered model creates metadata that are comprehensive, complete, uncompromised, unambiguous, usable and reusable, sharable, and non-proprietary. In short, this model ensures the long-term persistence and usability of the data collection.

The principles of metadata in a metadata-centered model

The word "metadata" has become a buzzword and a vague term that means different things to different people. While its most generic meaning of "data about data" is accurate, it is neither informative nor precise. When the word can refer to everything from a minimal-level MARC record to a DDI 3 XML file describing a complex dataset, its imprecision makes it lose any useful meaning.

One way we can facilitate our understanding of various kinds of metadata is to develop some principles of metadata and ask, for any given application of the term, to what extent the metadata in question lives up to these principles.

Ideally, then, in order to have metadata that support a metadata-centered model of information management, discovery, retrieval, use, and preservation, metadata should be:

- Preservable across time and space, across changes in technology.
- Comprehensive (can describe every important aspect of a digital object)
- Complete
- Uncompromising (No limitations on which fields, or how many, or how long. Metadata must not be trimmed, truncated, munged, abbreviated, aggregated, or otherwise compromised in order to fit into a particular software application, transport format, service-delivery mechanism, etc.)
- Consistent (Two similar digital objects are described in the same way)
- Flexible (Allows for expansion; preserving backward compatibility)
- Unambiguous
- Sharable (give and take)
- Usable and Re-usable
- Documented (well and completely documented for people and for machines)
- Non-proprietary
- Easily parseable (and, preferably, understandable) by machine, now and in the future.

In addition, a metadata-centered model should recognize and support some related principles:

- Different kinds of objects will require different kinds of metadata.
- A library of metadata should be designed for storage and exchange of metadata content; it should facilitate the implementation of different applications rather than be designed to implement any one application.

- Conversely, no application or service or software package should dictate the content of the library of metadata.
- Metadata need not be stored in the most compact way. The criteria for metadata should be that they are complete, parseable, and well documented, not compact.
- Metadata should meet international standards, where available
- A library of metadata should allow "minimum records" but should accommodate "full records." The ability to create or acquire complete, full metadata may be limited by financial or other practical constraints, but a library of metadata must ensure the ability to store everything even if everything is not stored. ("All of the metadata some of the time, some of the metadata all of the time.")

DDI

The social science data community has a standard that meets these criteria. It is the Data Documentation Initiative (DDI) standard. DDI is an international effort to establish a standard for technical documentation describing social science data.[1] A membership-based Alliance is developing and maintaining the DDI specification, which is an XML Schema. It is becoming a de-facto international standard. SPSS and SAS are working on tools to directly import data documented by DDI. Web-based data-service projects are making use of DDI. These include Dataverse, the Census Bureau's DataFerrett, the ICPSR catalog of data, the ICPSR database of variables, the Social Science Statistical Data Finder at Yale, the SDA software, and Nesstar.

The DDI 3 XML Schema is modular and is capable of describing many types of data. The DDI is not just replacing the old "codebook"; it is providing a way to build new functionality into the use and exchange of data.

The "Bow Tie" Model

The metadata-centered approach works particularly well for a number of reasons specific to the nature of social science data.

As noted in Chapter 1, data libraries, data archives, and researchers have been sharing, exchanging, preserving, and using social science data for more than three decades. Data archivists have had to migrate data over multiple generations of hardware, media, and other changes in technology to ensure the usability of the data. The two keys to doing this have been to keep data in a 'neutral' format (e.g., plain ASCII text files instead of software-specific formats) and to keep the metadata that describe the data files separately. Early social science "metadata" was simply printed in books called "codebooks" (so called

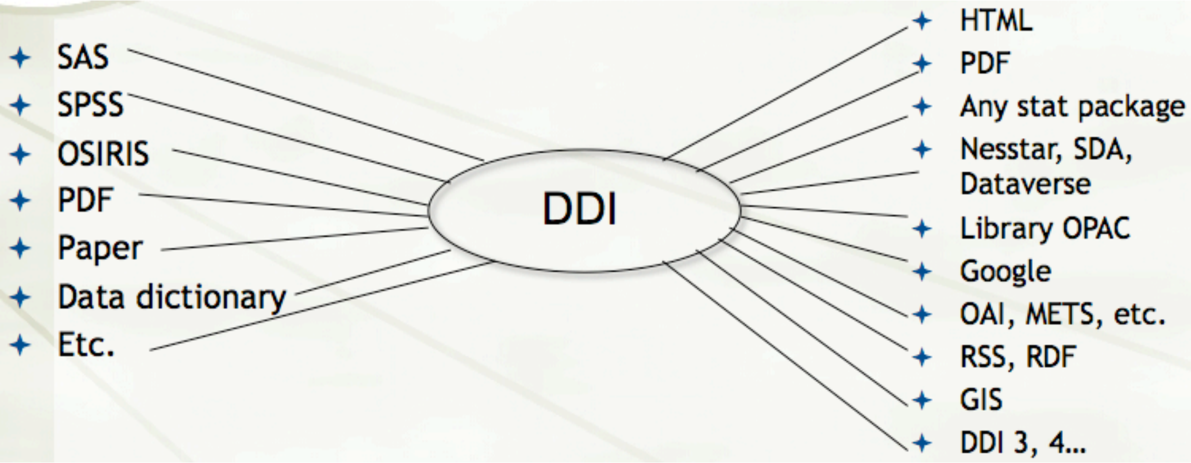
[1] http://www.ddialliance.org/. As noted in Chapter 11, there are related metadata standards: the Statistical Data and Metadata Exchange (SDMX) and the eXtensible Business Reporting Language (XBRL).

because they listed the “codes” for each variable in a data file, e.g. 1=male, 2=female). Other machine-readable “standards” have been used, but none was more than a widely used convention and none was truly ‘neutral’ and preservable. Examples of such formats include the files that statistical software use to read data (SPSS “syntax” files and SAS job files and OSIRIS “dictionary” files). In 1995, work began on developing a truly neutral, cross-platform, complete metadata format in XML and this work, funded by NFS, resulted in the DDI format.

The existence of the DDI format allows data archivists to avoid having to choose proprietary software and metadata formats and makes it possible to choose a format that is not software-specific and is, most importantly, easily parsed and transformed by machine.

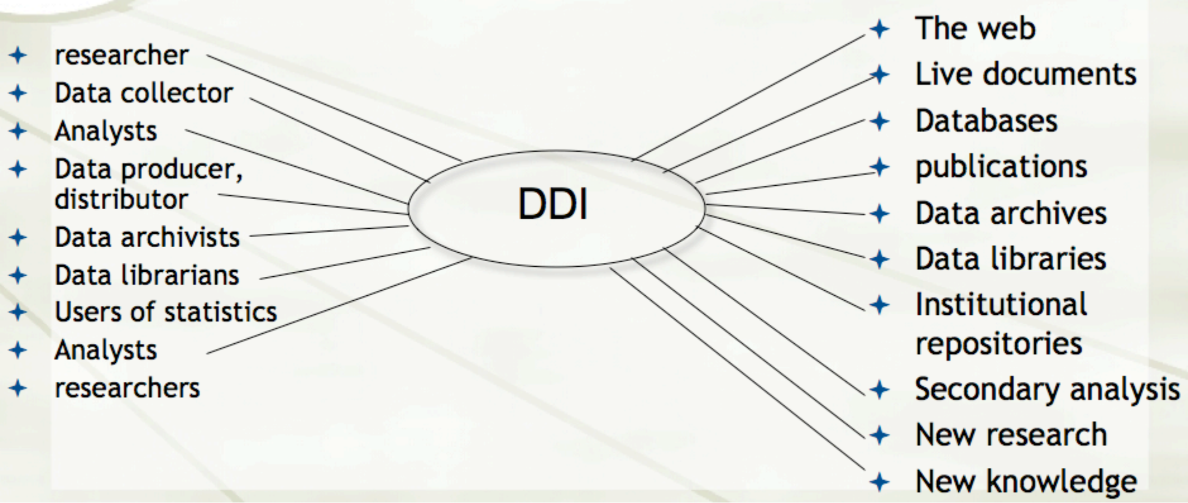
Data libraries face two related metadata problems. First, we have various sources of information that document social science data files and those sources are in a wide variety of formats and media. Second, we want to be able to use and reuse that information in a number of very different ways that rely on different software and computing environments.

The DDI metadata-centered solutions to these problems suggests a picture of a bow tie. With this model, we get information (on the left side of the bow tie) from numerous sources in numerous formats, convert it to DDI (the knot of the bow tie – a library of XML files), and then use and reuse the DDI metadata in a wide variety of ways (on the right side of the bow tie).



This illustrates how, once metadata are converted from legacy formats to DDI/XML, they can be used and reused by different applications for different purposes including applications that are unknown at the time of the creation of the metadata. Once a tool is available to convert DDI to PDF, for example, PDF codebooks can be created on the fly from the metadata for any study. DDI can be used to load the OPAC and to create Google sitemaps and OAI-PMH files for indexing by robots. And so forth.

The bow-tie also mirrors the SIP-AIP-DIP Information Model of OAIS. With it we can envision (see below) new, complementary roles for people who are involved in the lifecycle of research data and new and different products as part of the lifecycle of information.



As an example of the utility of DDI, see the study: Martinez, Luis. “The Data Documentation Initiative (DDI) and Institutional Repositories.” DISC-UK DataShare, February 2008. http://www.disc-uk.org/docs/DDI_and_IRs.pdf.

17.10 – Access Strategies

Data Basics

Computing Strategies for Data Services

"What kind of computing power do we need?"

Those beginning a data service frequently ask about the type of computing equipment needed to support data in a library. This question is asked with the usual expectation of being told to buy a specific brand of computer. However, the best advice begins by redirecting the inquirer's focus back to the level of data service envisioned. Once the service model is clearly understood, a computing strategy can be devised to complement this plan.

For example, if the data service plan consists of only ordering and passing data along to clients, the computing requirements will be less than if the plan includes extracting and subsetting data for clients. If preserving research data for the campus is part of the service plan, the computing requirements will be different yet again. In the end, a computing strategy should identify -- within the context of your university's computing environment -- the hardware, software, network connectivity, and network services needed to support a specific level of data service.

In all likelihood, the computing strategy of the data service will need to be built using the computing resources available through your organization. For example, a data service may have to rely on the IT unit in the university library, while this IT unit may in turn depend on central computing services offered on the campus. The larger your organizational environment, the more complex your computing environment will likely be. Each of us must live with the computing environment dealt to us. Very few of us have enough influence to steer the direction of central computing at our institution, although we may have a little more success in our immediate workplace. Thus, we all try to match our computing needs with the resources provided locally.

The Elements of a Data Services Computing Strategy

The elements of a typical data services computing strategy include processing power, storage space, statistical software, utility software, network connectivity and Internet services.

Desktop computing power to support a variety of applications. It will be useful to have a variety of application software to provide services for data. A typical office suite with a word processor, spreadsheet, presentation and database applications should be part of this mix. Common desktop computer office suites include Microsoft Office, Apple’s iWork, and OpenOffice. The trend toward Internet-based office applications is providing an increasingly viable alternative to installing software on a local workstation. Examples include suites by Google[1] and Zoho.[2] Each office suite has its own strengths and weaknesses. The best choice is the one that fits your service strategy. Functionality, ease of use and management, affordability, and compatibility with your organization and your users are all factors to consider.

Desktop versions of one or more major statistical packages (e.g., SAS, SPSS, STATA, R, S-plus) should also be part of the software configuration. Depending on the level of data service being offered, statistical software will provide different functionalities. They may be used to create internal system formats for one the major packages, such as an SPSS sav file or a SAS sd7 dataset. They can also be useful to combine datasets, subset cases, and manipulate variables. Some data services may even perform basic statistical analysis. Like office suite software, statistical systems have strengths and weaknesses. *The software you choose should meet your service needs and be compatible with and complementary to the statistical computing needs of your users*.

In addition, a collection of general desktop and network utilities will be needed. Internet tools, such as a Web browser and secure-shell utilities for telnet and ftp, are essential for online data discovery and access. File compression software, such as Zip, is needed to mange file “archiving” formats[3] that contain bundled and compressed files. Software that can read PDF files (such as Adobe Reader) is essential because of the volume of documentation now available in PDF format. If spatial data are part of the data service, GIS software, such as ESRI’s ArcGIS, should be included on the list of software. If you rely on XML formats (including DDI) you may need XML editing or programming software.

The above software applications should be considered when selecting a desktop computer for data services and the choice of machine should have the memory and disk space to store and run these programs. This workstation, however, may not be the only computer for processing data but rather may be a service-point from which data are disseminated after having been prepared on another processor, such as a central Unix system. In such a situation, statistical software would need to be installed on the central system. This example demonstrates the need for a computing strategy to take into consideration software required on local desktop machines as well as the software available on other systems used by your data service. A comprehensive software list should be identified that includes both local and remote processors.

Internet services and tools to support data services. The Internet is an integral component of the computing strategy of a local data service. For some

[1] Google applications are online at http://documents.google.com.

[2] Zoho applications are online at http://www.zoho.com.

[3] System managers typically refer to "packaged files" as "archived files". Thus, Zip and the Unix tar utility are viewed as archiving tools. For those working in data services, archiving entails far more responsibilities than storing a bunch of files in one file and then compressing its. Consequently, we prefer to call the use of these utilities as "packaging" rather than "archiving" tools.

time, the Internet has been used as a way to facilitate the discovery of local data holdings. Using lists that could be browsed or searched on local Web pages or including records for data titles in the library's online public access catalogue, clients were provided with a resource to find items in a local data collection over the Internet. As data providers develop online access services to their data collections, the Internet has become a means of delivering data to clients. ICPSR Direct and MyData are examples of a major social science data provider offering Internet access to its holdings and to an online analysis system. The licensing of access to data through such services is typically managed the same way that libraries manage access to the ever-expanding collection of bibliographic databases, electronic journals, and other digital content. The local data service typically arranges the addition of the link from a data provider to their institution's list of online database subscriptions. Through proxy servers and VPNs, clients can authenticate themselves and access these data resources over the Internet. The computing strategy should take into account the tools to be used for discovery purposes as well as a list of online data vendors that provide access to data resources.

The computing strategy should also identify the Internet services and tools to be provided by the local data service. Offering such services locally may require access to a Web server, a content management system, blog software, wiki software and a database server. Increasingly, these services are provided on a shared system running a virtual server. The determination of who will be responsible for running the server will be part of the overall computing strategy. For example, a campus-wide computing center may operate the server while the local data service is responsible for maintaining the content on the Web site. The Library may run its own Web server and provide the data service with content management tools for space on its Web site. Ideally, the data service will not have to provide system support for a Web server but will manage its content on the site. Some coordination will be required between the Web service and the content provider. For example, the choice of a content management system could be dependent on the Web site provider's system.

Common Web services supporting library OPACs, bibliographic databases, and the storage and retrieval of small files may not be adequate for data services. Staff in the data service will need to work closely with those who provide common library and campus Web services. In many cases, these services may need to be modified to accommodate data services or, alternatively, different Web services will be needed to meet the requirements of data services. Some data services may face a situation in which either desired Web 2.0 tools are not provided locally or the content management system is so closely controlled by others that the data service cannot provide dynamic content. In such situations, some data librarians have found that publicly available services are reliable enough and enable them to add flexibility and functionality to an otherwise static Web site.

Software and web-based tools that enable a data service to provide sophisticated online services are readily available and should be considered as part of an overall computing strategy. Three of these are the *Nesstar server*, which continues to be developed through a partnership between the UK Data Archive and the Norwegian Social Science Data Service, the *Survey Documentation and Analysis* (SDA) system, which is maintained by the Computer-Assisted Survey Methods Program at the University of California, Berkeley, and the *Dataverse Network Project*, which is housed at The Institute for Quantitative Social Science (IQSS) at Harvard University.

Nesstar and SDA systems run off locally installed HTTP servers. Data are managed in the Nesstar system through the Nesstar Publisher, which is an editor that produces DDI metadata and write files to a Nesstar server. The Nesstar Publisher is an MS Windows application, while the Nesstar server operates under either the Microsoft or Sun operating systems. SDA operates from an ASCII data file and a Data Description Language metadata file that are transformed into an SDA dataset. A separate HTML codebook must also be generated and together these products are added to the SDA web space. The utilities supporting the preparation of SDA file formats operate under MS Windows, while the server can run off Microsoft, Linux or Sun operating systems. These system dependencies will also contribute to the overall computing strategy that is developed. Dataverse can be installed on your own machine running Linux, but you can also use IQSS to host your data on their installation of Dataverse while creating and managing the “branding” of your collection.

If a data service has responsibilities for preserving data produced by researchers on its campus, the computing strategy should consider the use of local institutional repository services. Preferably, the data service will not be responsible for the operation of the institutional repository. However, the functionality of the institutional repository should be assessed before making a commitment to use it. For example, if the support for research data in the institutional repository is deemed inadequate or insufficient, a data service may choose to run a Nesstar, SDA or Dataverse service either to complement or bypass the use of an institutional repository.

As part of a computing strategy, the data service should also take into account public, collaborative Web services for data sharing and data visualization. These services may become another viable method of providing access to data as they grow in popularity with users and as they add and improve functionality. Services that show promise include:[4]

- Many Eyes http://www-958.ibm.com/software/data/cognos/manyeyes/

[4] Two useful reports on these kinds of services are: MacDonald, Stuart, and Luis Martinez Uribe. “Libraries in the Converging Worlds of Open Data, E-Research, and Web 2.0.” *Online* 32, no. 2 (March 2008): 36-40; and MacDonald, Stuart. “Web 2.0 Data Visualisation Tools: Part 1 - Numeric Data.” DISC-UK DataShare, January 2008. http://www.disc-uk.org/publications.html.

- Data 360 http://www.data360.org/
- Google Finance http://www.google.com/finance

Large, readily available quantities of fast disk space. The desktop computers in a data service should have large hard drives with fast internal data transfer rates or burst rates.[5] Fast hard drives are desirable because of the type of data management tasks typically performed by a data service. This type of processing involves a lot of input-output activity. For example, processing files in a compressed format, such as Zip or tar, generates a lot of disk activity as records are read, uncompressed and written to a drive. Saving the internal file format for a statistical package will also produces a lot of disk activity as the file is written. A trade off tends to exist between the largest hard drives on the market and internal data transfer rates. Usually, the largest drives do not have the fastest rates. Given the relative size of today's hard drives, a faster internal data transfer rate is more desirable than the largest amount of disk space. Consumer-grade computers and data centers are starting to make use of fast-access Solid State Drives as a complement to traditional, mechanical, spinning hard drives. The architecture of the hardware and OS are usually designed to make use of SSD for caching recently used, often-requested information and to speed up active applications

Network files systems are valuable resources for storing data files.[6] The data transfer rates over network file systems can be problematic, however, for a data service. Data input-output will now depend on both the rate of the network disks and the speed of the network. Therefore, depending on these two factors, one may decide to store files on the network system but process them on the hard drive of the data service's desktop computer. One strategy would be to localize all processing of files on the local desktop computer. This might entail copying files off a network drive to the desktop's hard drive, processing the data and after the data have been written to the hard drive, copying the output back to the file server.

Network connectivity that permits high-speed file transfers. As just mentioned, the speed of the connection to a network file system is a factor in the overall data service computing strategy. This connectivity should allow fast file transfer on the campus network as well as the wider Internet. Steer clear of a connection to the network where bottlenecks typically occur, such as large local area networks sharing a common router. Another situation to avoid is sharing a connection with a large cluster of OPAC stations. Once again, the default

[5] The internal data transfer rate is the speed that a hard disk can read data from the surface of the drive's platter and transfer it to an internal buffer or cache. See http://www.pcguide.com/ref/hdd/perf/intRate-c.html for a discussion about internal data transfer rates in PCs.

[6] Storage Area Networks typically allow remote devices, used as disk arrays, to be attached to servers in ways that the operating system makes the storage device appear as thought it is part of the local computer.

configuration by your organization may be inadequate for data services. For example, a library might provide network connections for a dozen PCs on a single switch, which may work well for the applications being run by colleagues sharing this connection but be inadequate for the type of network requirements for a data service PC. The staff in the data service may need to work closely with the network administrators to ensure adequate bandwidth.

Securing a stable network connection can be a challenge and might require using services at another location of the campus network with better connectivity. For example, a central Unix service may be better situated on the network. You may then decide to use your local workstation as a terminal connected to the Unix service for managing large file transfers. In such a situation, an X Window application can provide a graphical user interface to remote Unix systems and would be a recommended addition to the list of software for the local workstation in data services.

Access to high-performance computing running statistical applications. There are times when access to SPSS or SAS on a high-performance computing configuration can be very useful. This type of computing resource typically consists of computer clusters that are centrally administered on a campus. Access to a cluster machine usually requires an account on a high-performance system running Unix. A version of SAS or SPSS for Unix will need to be installed by the cluster provider. Clusters are often connected to institutional file servers, which allows their applications to read and write files accessed elsewhere on campus. This machine may be used to subset data from very large files, such as the CRSP daily stock exchange file or the public use microdata from one of the Canadian or U.S. censuses. These systems are also helpful in performing file verification tasks on very large files to ensure that a file has the proper number of records and that the record length matches the data documentation. The overall computing strategy should taken into account the availability of high-performance computing and how it could support the desired level of data service.

Implementing a Computing Strategy

One of the challenges of implementing a computing strategy arises from the variety of products and services in each of the areas discussed above. Here are some practical guidelines to follow when acquiring new equipment or software for a data service.

Always investigate the computing support at your institution. Take advantage of the site licenses or educational discounts your campus has for software and hardware. There may be even further advantages to site license software than just saving money. Central support is often provided with this software, including help with installing and debugging programs. There still may be instances, however, where you discover that the options provided by your campus' site licenses are too narrow and that you need to purchase other software or hardware. For example, your institution may only have a site license for STATA when you feel you need either SPSS or SAS.

When buying hardware consider purchasing equipment that is compatible with your university's computing environment. This will often allow you to share peripheral devices or services without a great deal of hassle. Furthermore, you are more likely to find people with experience who can help with installing systems and diagnosing hardware problems.

A corollary to this guideline is to **acquire equipment that is compatible with the majority of your clients.** If the largest percentage of your clients use PC's, purchasing a PC would be wiser than a Macintosh even though you may prefer the Macintosh operating system. By working with equipment similar to your clients, you will be more able to address their inquires about using data on their personal workstations. The selection of removable media devices should also match those most commonly used by your clients. Over recent years, such preferred media have migrated from diskettes to CDs to USB memory sticks and DVDs. Increasingly, clients working with spatial data files will carry an external hard drive filled with digital boundary or cartographic files. These typically require fast USB or Firewire ports on the data service’s workstation.

A guideline to follow when purchasing hardware is to **buy as much memory and fast disk storage as you can afford.** Both of these are good investments.

Never underestimate the amount of time and skill it takes to administer a computing system for data services. If you do not have the staff nor skills to provide a great deal of system administration, you will want to keep the computing environment in your data service as simple as possible. Currently, maintaining a PC and Macintosh system is easier than supporting a Unix machine, which requires a great deal more systems administration. Unless you are willing to invest in the skills to support a Unix system, you are better off with a PC or a Macintosh.

Ask your data service colleagues at other institutions about their experiences. While the variety of solutions may seem overwhelming, many of the experiences are common. Just knowing how others have confronted their computing problems can often help.

18.8 - Computing Strategies

Data Basics

Promotion Strategies for Data Services

Marketing your Service

There will always be new members of your institution who are unfamiliar with your services, and long-time members of your institution who never before needed data until now. Marketing can sometimes feel like you are repeatedly informing the same people over and over. But until that moment of need, it is as if you have never said a word.

Ultimately, your best advertising may be the result of providing quality services to your users. Increasing your visibility and the number of users who rely on your service may mean more support in the long run in terms of budget and staff lines. Keeping an eye out for new academic degree programs or research centers that are developing and would benefit from including data services support as part of their proposal is a good way to get involved with new user groups from the ground up.

The overall goal is to make your services visible and data easily obtainable. Here is a list of promotional activities to consider for your environment

- Make services and/or data and documentation easily available via your institution's *web* pages.
- "*Mainstream*" access to and visibility of your data holdings and resources by incorporating your data holdings (at the study level) into appropriate databases, lists of licensed information, OPAC, etc. (Be sure to investigate adding ICPSR records to your catalog.[1])
- Incorporate data resources into appropriate traditional *library instructional sessions*.
- Use existing digital communications to reach your user communities. These may include existing listservs for departments or disciplines or other groupings.
- Don't overlook print. A nice *brochure* or a paper and ink *newsletter* are useful ways of reaching new users at local events and at point-of-use locations. Make a brochure available to new or prospective faculty, researchers, and graduate students in relevant departments. Include sample newsletters in welcome packets distributed to new community members.

[1] http://www.icpsr.umich.edu/icpsrweb/ICPSR/or/metadata/. "ICPSR shares its metadata records with the membership to promote wider awareness and use of ICPSR's social science data resources. In particular, we encourage members to integrate the records into local Online Public Access Catalogs (OPACs) intended primarily for the use of faculty, staff, and students at their institutions."

- Use social media to give your service visibility and to keep in touch with your users.
- Keep users up to date digitally. For existing users, use social media, a data-blog with an RSS feed, a data-users email list, and similar low-cost methods for announcing new developments, new data acquisitions, training sessions, ICPSR Summer Program, and so forth.
- *Survey* users to identify individuals with immediate or potential needs.
- Establish an *advisory group* of researchers with representation from a broad spectrum of departments.
- Send regular news submissions to other *campus publications* (newsletters, e-mail listservs, blogs) such as faculty/staff publications, student newspapers, computer center or library newsletters, departmental or campus-wide listservs.
- Keep *relevant campus offices* apprised of your services, e.g., sponsored research, graduate affairs, and institutional research.
- Hold general *workshops* about your service for various groups at your institution to ensure proper referrals, e.g., computer center, library, departmental graduate organizations, other user groups.
- Hold workshops on *particular data sets*, e.g., census, CPS, General Social Survey.
- Provide training and information sessions for *library staff* to make them aware of your services, collections, and users.
- Organize a *speaker series* of researchers to share results of research.
- Circulate appropriate print and electronic newsletters, e.g., *ICPSR Bulletin*[2] and other ICPSR promotional materials[3]*, DLI Directions*,[4] to colleagues in related units such as Government Documents, Collection Development, etc.
- *Promote data-deposit* with ICPSR or your own data archive and provide information to your researchers about the data life-cycle, DDI, and preservation issues.

Integrally associated with promoting data services is how you justify your ICPSR membership. With the fluctuations in educational support that routinely occur, you can expect at some point, if not annually, to be asked for budgetary justification. Having the foresight to monitor and track certain measures makes this task an easier one. Ask ICPSR and other data distributors and services about reports of usage that they can generate for you.

[2] http://www.icpsr.umich.edu/icpsrweb/ICPSR/org/publications/bulletin/

[3] http://www.icpsr.umich.edu/icpsrweb/content/ICPSR/org/publications/index.html

[4] http://www.ddialliance.org/resources/publications

Justifying Your ICPSR Membership

For many of us the ICPSR membership is only one of the avenues through which we acquire data, but it also may be one of the most costly. In the ***Survey of ICPSR Official Representatives, 1997*** conducted by the ICPSR staff, the annual membership fees of the 236 respondents were paid in by a variety of units:

Value Label	**Per Cent**
A single academic department	25%
By the college/university library	25%
By the office of the Dean or Vice Pres.	19%
By more than one academic dept.	10%
Other (inc. Federated memberships, Multiple campuses, State System)	10%
By the Computer Center	6%
By the Graduate School	2%
Grant funds	1%
Missing	2%

How the budget process works in your institution and who pays the membership will differ, but it is common for some justification to be necessary. One of the first questions administrators often ask is, “How many people use ICPSR data?” but this is just one measure of worth. Here are several factors to keep in mind as you think about ways that make sense at your institution to justify your membership.

- **ICPSR Direct Reports**
  ICPSR provides reports of use of *ICPSR Direct*, the facility that allows members to download data and documentation directly. These reports can be vital in quantifying the amount of use of ICPSR membership by department, study, etc.

- **Number of Uses of ICPSR Files Available Locally.** Where data files are downloaded and stored locally for local use, it is often possible to obtain automatically the number of times a file or study has been accessed. If you provide local extraction services for your clientele it is probable that these uses can be automatically recorded. While a crude measure since you may not know what the usage represents, it does provide another count of activity.

- **Recruitment Activities.**

Attracting top candidates for faculty and research positions in the social sciences will often depend upon the research services available at an institution. The ICPSR membership is a key asset in this area and can be used successfully in the recruitment process.

- **Cost of Membership Services "On the Street"**
  ICPSR staff can provide an Official Representative with a *Utilization Report* for your institution that includes the number of data sets ordered, number of summer program attendees, and bulk mailings to the OR. It includes a calculation of the non-member charges that you would have incurred for these services if you were not a member. As an example, one average year, Binghamton University sent 2 people to the Summer Program and ordered 370 data sets for a grand total of $102,636. This total does not reflect a number of calls to User Support or e-mails to staff who virtually provided assistance with data, but it certainly made the point.

- **Uses of Various ICPSR Web Services**
  There are other uses of ICPSR data that do not directly require membership, such as the *General Social Survey* and *American National Election Study* accessible from the ICPSR web site that provide extraction capabilities for anyone. Also, you may be asked to retrieve a study and documentation from the Publications-Related Archive, a free, but important service for those required to deposit their data as a condition of publication so that another scholar may reproduce their research results. You can also assist your researchers in depositing data with the Publications-Related Archive as well. These are important services for which the ICPSR may get external funding that directly benefits the membership. These types of uses are also worth recording when you assist users with them.

- **Local Technical Support.**
  If you staff a data services point or an e-mail data service, you can track the types of queries you handle from a simple measure of how many questions were asked or the number of people assisted to a more detailed level that corresponds to your own data-service levels-of-service policy statement. For example:

  - ✓ *Ready Reference* (e.g., locating a known data set or codebook)
  - ✓ *Extended Reference* (e.g., identifying appropriate data to address a research question, or in-depth questions about a specific study)
  - ✓ *Technical Assistance* (e.g., subsetting data, or moving data between platforms or formats)
  - ✓ *Research Assistance* (e.g., statistical programming, or research methodology)

- **Local Instructional Support.**
  Data Services staff will be frequently called upon to provide general overviews or more detailed instruction in the use of ICPSR resources. Doing a classroom session for a graduate statistics or methods course is a good strategy for gaining support as well as a line of defense against twenty students individually coming in for basic assistance. Also, working with the instructor of an undergraduate class to devise a workable quantitative assignment for students new to data analysis and survey research is another way the resources from the membership can be used and promoted.

- **ICPSR User Support and Archival Support Services**
  ICPSR staff are very knowledgeable about many of the data sets in their collections. They can answer questions about the technical requirements and difficulties associated with manipulating certain data sets as well as provide expert advice about the content of various studies. They can also provide referrals to other ORs who have had similar problems. Another invaluable area where they can provide assistance is in data preparation and archiving for those wishing to deposit data with ICPSR or another archive. The "ICPSR staff as resource" for the social science community should be included in the support the membership receives.

- **ICPSR Summer Program.**
  The ICPSR Summer Training program, founded in 1963, is an internationally and nationally recognized program of instruction for basic and advanced training in the methodologies and technologies of instruction and research in the social sciences. This unique resource for the social science community, supports training and retooling of graduate students and researchers around the world. Membership fees contribute to this program and over 200 colleges and universities send individuals from their campuses to attend each year.

- **Grant Applications.**
  The success of a grant application often depends on the institutional support available. ICPSR's extensive collections of low cost access to data for secondary research analysis can provide the raw data needed for a grant application.

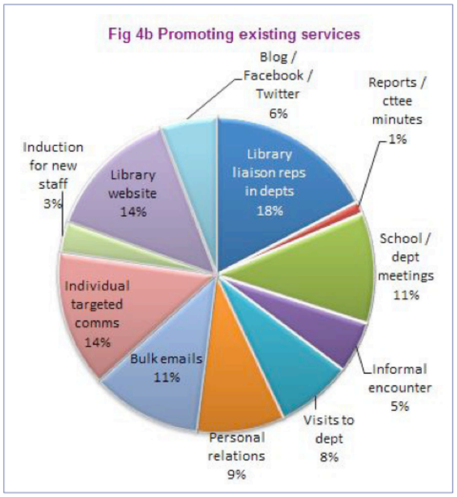
- **Investment in the Future of Quantitative Social Science Research.**
  When compared with the research laboratories of physicists and chemists, the "laboratory" of the social sciences is very inexpensive, but without the "ICPSR as archive," the future of social science research would be in jeopardy. This "laboratory" has certain fundamental requirements whose costs are not insignificant. Our membership fees support this archival mandate of the ICPSR which ensures that the data and documentation collected today will exist and be usable with the software and hardware of the

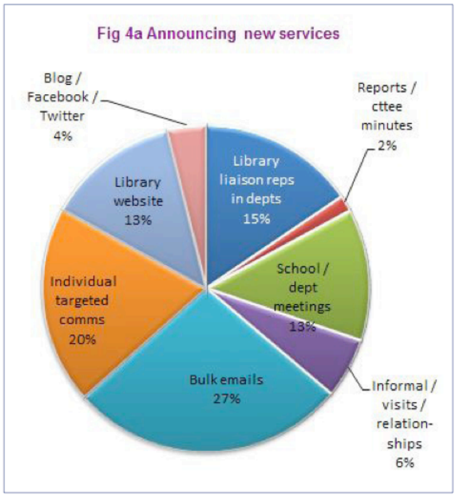
future. If, as discussed in chapter 11, your service plan relies on ICPSR to provide surrogate preservation functions, annual justification of the cost of ICPSR membership should call attention to those decisions and the alternative costs of local preservation. ICPSR also provides leadership and works with other key national and international organizations to promote standards for data and documentation. These intangibles are at the core of the ICPSR's mission and contributes to the essential value of the membership.

- **Your local collections and services plans.**
  If you have planned your collections and services around the availability of data and services provided by ICPSR membership, be sure to refer to those plans explicitly. Make sure that administrators understand the role that ICPSR membership plays in providing collections and services.

Summary

A recent study[5] showed the variety of effective methods of promoting existing and new services:





In the same study, librarians reported that the biggest barrier to effective communications with their users was time constraints. Promotion can be time consuming, but it can be a useful way to get to know your users and their needs as well as an additional way for your users to learn about your data and services. Finding a balance of promotional activities that are effective without taking up so much time that they get neglected is important. Starting

[5] Creaser, C., & Spezi, V. (2012). *Working together: evolving value for academic libraries*. Leicestershire, UK: Loughborough University, Library and Information Statistics Unit. Retrieved from https://libraryvalue.files.wordpress.com/2012/06/ndm-5709-lisu-final-report_web.pdf

a newsletter and never getting out a second issue, or creating a blog and never posting to it are may damage the reputation of a service.

Communication with your user communities is essential for having a robust, effective, trusted service. In the lifecycle of research data, all the stakeholders should have good working relationships and open lines of communication. Promotion of services is not a disagreeable, self-serving public relations chore. It is an essential component of your service.

19.8 – Promotion Strategies

Data Basics

Professional Development

Professional development, continuing education, training, and just keeping current are essential responsibilities of the data librarian. As noted in other chapters, the data library and its responsibilities and opportunities are greatly affected by events that it does not control. The essential tool that the data librarian has to deal with this is professional development.

In addition, data service professionals often feel isolated in their organizations because their work is so different from their colleagues. Keeping in contact with others data service professionals through associations and electronic discussions groups is an excellent way to keep current with new developments in the field, as well as to obtain support and advice from others doing similar work.

This chapter lists key organizations, and mailing lists that provide support to data professionals. These provide the tools for social networking, the first place to look for training and educational opportunities, and links to the community of people who can offer help and advice.

See also “Keeping Current In Social Science Data” by Joanne Juhnke, Special Librarian Data & Program Library Service (DPLS) University of Wisconsin—Madison: http://www.iassistdata.org/downloads/2006/a1_juhnke.pdf.

Informal Professional Development

While formal training opportunities exist and professional organizations provide professional development opportunities, the data services community provides many opportunities for informal professional development. As noted in Chapter 12, communicating with data service peers is an important part of providing services. Keeping in touch, knowing who to ask or where to post your question is one key to finding answers to your questions, whether generic or specific. The data services community is warm, friendly, and very helpful. Simply making connections with the right mailing lists is a good first step. “Lurking” on mailing lists is acceptable and useful way to pick up hints and find answers to questions you didn’t know to ask.

In addition, there are new opportunities for sharing. The IASSIST web site (http://www.iassistdata.org/) has many useful pages including a blog and a collection of issues of the *IASSIST Quarterly*. International in scope, the articles in *IQ* deal with social science information and data services and are written by individuals who manage, operate, and use data archives, data libraries, and data services.

The Canadian Data Interest Group for Reference Services (DIGRS) blog (http://blogs.library.ualberta.ca/digrs/) has postings for challenging data-reference questions, their answers, and how the answers were found. The postings are tagged to make it easy to browse or search for answers.

Canada boasts an extensive training program developed around the Canadian Data Liberation Initiative (DLI) and their training materials are available on the web at the *DLI Training Repository / Dépôt des documents de formation de l'IDD* (https://ospace.scholarsportal.info/handle/1873/69).

Having a place to record your reference questions and answers can be a useful tool for sharing and learning. As noted in chapter 12, Stanford University is doing this and creating a search engine to aggregate across similar sites.[1] The University of California Berkeley is developing a custom search engine to search data-rich sites.[2]

Organizations

American Library Association (ALA)
http://www.ala.org/
50 East Huron
Chicago, IL 60611

Government Documents Roundtable (GODORT)
http://www.ala.org/godort/
http://wikis.ala.org/godort/index.php/Main_Page
GODORT provides a forum for discussion of problems and for exchange of ideas for librarians working with government documents. Of particular interest to data librarians is its *Government Information Technology Committee* and the *Education Committee*. The latter provides copies of instructional materials including instructional handouts for CD-ROMs.

American Statistical Association http://www.amstat.org/
Founded in 1839 to foster excellence in the use and application of statistics to the biological, physical, social and economic sciences, ASA is a leader in promoting statistical practice, applications, and research; publishing statistical journals; improving statistical education; and advancing the statistics profession. Relevant subunits include Committee on Professional Ethics, Section on Survey Research Methods, Social Statistics Section, and Section on Government Statistics.

Association of Public Data Users (APDU) http://apdu.org/
APDU is a national network of users, producers, and distributors of federal, state, and local governmental statistical data who are concerned about the availability, use, interpretation of public data in the United States. Its members include state data centers, governmental agencies, academic institutions, commercial vendors, non-profit organizations, and

[1] http://freegovinfo.info/node/1888

[2] http://snipurl.com/3bdr5 [sunsite3_berkeley_edu]

individuals. Membership benefits include the APDU Newsletter, annual conference, and membership directory. This group plays an advisory role in Census planning and other statistical policy and data delivery areas.

Canadian Library Association (CLA) http://www.cla.ca/

Access to Government Information Interest Group (AGIIG) AGIIG is open to individuals interested in access to government information issues. Members actively campaign for preserving and improving access to all formats of government information. Recent concerns include status of electronic products distributed to depository libraries and establishing an inclusive national information policy.

CAPDU (Canadian Association of Public Data Users) http://www.capdu.ca/

Formed in 1988, CAPDU is a member of the Learned Society of Canada and is concerned with the availability, use, and interpretation of public data in Canada. Membership is open to users, distributors, and producers of data within Canada. This group was successful in lobbying the Canadian Association of Research Libraries (CARL) to form a consortium to buy 1986 and 1991 Census of Canada data. More recently, CAPDU has partnered with other groups in Canada in support of the Data Liberation Initiative (DLI) to provide affordable access to Statistics Canada data files and databases for teaching and research. The main organ of CAPDU is a mailing list.

Digital Library Federation http://www.clir.org/dlf

DLF is a consortium of libraries and related agencies that are pioneering in the use of electronic information technologies to extend their collections and services. They promote standards and best practices in a wide range of areas including digital preservation, structural metadata and licensing.

DLM Forum

http://ec.europa.eu/transparency/archival_policy/dlm_forum/index_en.htm

The DLM Forum promotes best practices for access and preservation of information within the European community. DLM is an acronym for the French "Données lisibles par machine" (machine-readable data). The DLM-Forum is based on the conclusions of the European Council <documents/council.html> (94/C 235/03) of 17 June 1994 concerning greater cooperation in the field of archives.

Council of Professional Associations on Federal Statistics (COPAFS)

http://www.copafs.org/

Since 1980, COPAFS has represented individuals united to increase participation of professional associations in the development and improvement of federal statistics programs; establish communication with federal agency personnel, congressional committees, and others involved in federal statistical policy and programs; make information on federal statistics available to members; and encourage the discussion of issues of public concern. Member organizations include professional associations, businesses, research institutes, and others interested in Federal statistics.

International Association for Social Science Information Service and Technology (IASSIST)
http://www.iassistdata.org/
IASSIST is an international association of individuals who are engaged in the acquisition, processing, maintenance, and distribution of machine readable numeric social science data. Founded in 1974, the membership includes information system specialists, data librarians and administrators, archivists, researchers, and computer programmers and managers. IASSIST encourages and supports the establishment of information centers for data reference, maintenance, and dissemination at local and national levels. IASSIST membership benefits include an annual conference with workshops, a quarterly journal, and a Listserv for members only, and a blog.

International Council on Archives (ICA) http://www.ica.org/
Conseil international des Archives
60 rue des Francs-Bourgeois
75003 PARIS, France
ICA is the professional organization for the world archival community, dedicated to the preservation, development and use of the world's archival heritage. Its members include national archives, major institutions, professional associations, and individuals. They have an active Committee on Electronic Records.

International Federation of Data Organizations (IFDO) http://www.ifdo.org
Zentralarchiv, Bachemerstr. 40, D 50931 Koln, Germany
Founded in 1977, IFDO brings together change: 25 to 27 members worldwide who actively engage in providing the social science community with computerized numeric information, documentation and analysis. It promotes projects and procedures for enhancing exchange of data and technologies among data organizations. Every four years IFDO jointly sponsors a conference with IASSIST in Europe.

Society of American Archivists (SAA) http://www.archivists.org/
The Society of American Archivists provides leadership to help ensure the identification, preservation, and use of the historical record. SAA holds annual conferences with workshops on electronic records management, sponsors publications in the administration of electronic records, and holds several electronic forums including, ERECS-L which is operated jointly by the Electronics Records Section and SUNY at Albany. The list is dedicated to discussions about the preservation and management of records in electronic form.

Mailing Lists Discussion Lists, and Net News groups

Discussion Lists on particular topics can be particularly useful in keeping up with the varied aspects of data services.

OR-L. The List for ICPSR Official Representatives and others:

The University of Alberta hosts the list "orl" for ICPSR Official Representatives and others interested in exchanging information or sharing ideas about ICPSR, its data, and its services. Send email to orl-request@mailman.srv.ualberta.ca with the word "subscribe" (minus quotes) only in the body of the message.

ICPSR hosted lists:

The ICPSR offers several lists to promote discussion and information exchange among members and ICPSR staff. http://www.icpsr.umich.edu/icpsrweb/ICPSR/org/lists/index.jsp

- **icpsr-announce**
- **summprog-announce**
- **recent-updates-and-additions**

20.6 - Professional Development

Data Basics

Understanding Weight Variables

Weight Variables and Their Significance

Weight variables are used in a statistical analysis to adjust for a study's sampling design. If a sample is drawn in which all of the elements have the same chance of being selected, each case in the resulting data file will have the same weight, namely, 1 (think of this as multiplying 1 times the value of each variable, which doesn't change anything.) However, some studies employ disproportionate sampling to insure a sufficient number of cases for important subpopulations.

For example, in the 1982 and 1987 General Social Survey [US], African-Americans were oversampled to form a separate national probability sample for their racial group. In 1982, 156 African-Americans were selected as part of the regular cross-sectional sample. The oversample added another 354 African-Americans to increase the total to 510 respondents. The African-Americans and all other races in the regular 1982 sample can be analyzed as a national sample without using a weight variable (this is a total of 1,506 cases). If one wants to use all of the respondents from the 1982 sample (i.e., the 1,506 from the regular sample and the 354 from the oversample of African-Americans), a variable named OVERSAMP must be employed to correct for the oversample.[1]

Using a weight variable adjusts for the differing probabilities that cases have of being selected in a sample. In other words, not every case has a weight of 1. Using the weight variable permits making generalizations to the population from which the sample was drawn.

There is another reason to employ a weight variable. The researcher may want results based on an estimate of the population instead of the sample size. For example, the Individual Public Use Microdata File from the 1991 Canadian Census has a weight variable that when used will produce results based on a population estimate of 27 million instead of 809,654 (which is the sample size of the individual public use file.)

Table 1 contains the frequency distribution for marital status from the 1991 Canadian General Social Survey. This table is made up of three frequency distributions for the variable DVCURMS2: the unweighted frequencies from the raw data file [unweighted], the frequencies applying the official weight variable FWGHT (final weight) [weighted 1], and the frequencies applying a rescaled variable WT [weighted 2].[2]

[1] Appendix A of the GSS Cumulative Codebook describes this in full detail.
[2] The variable WT was created using the SPSS command: compute wt=fwght*(13495/20525561).

Table 1: Applying the Weight Variable in General Social Survey, Cycle 6 [Canada]

```
DVCURMS2   RESPONDENT'S CURRENT LEGAL MARITAL STATUS

                           [ unweighted ]   [ weighted 1 ]   [ weighted 2 ]

Value Label          Value  Freq.     %        Freq.     %      Freq.    %

MARRIED                  1   6759    50.1   11277210   54.9    7414   54.9
WIDOWED                  2   1500    11.1    1124533    5.5     739    5.5
MARRIED BUT SEPARATE     3    528     3.9     598178    2.9     393    2.9
DIVORCED                 4    975     7.2    1281271    6.2     842    6.2
SINGLE                   5   3622    26.8    6124539   29.8    4027   29.8
NOT STATED               9    111      .8     119830     .6      79     .6
                           -------  -----   --------  -----   -----  -----
                     Total  13495   100.0   20525561  100.0   13495  100.0
```

This General Social Survey, which employs a complicated sampling design, requires the use of a weight variable. However, the weight variable not only adjusts for the sampling method but also provides population estimates for Canada. This produces results with a count of 20+ million, which was the number of Canadians 18 years of age or older in 1991.

The total number of cases in the raw data file is 13,495 (shown in Table 1 under the column, [unweighted]). There are exactly 6,759 cases with the value 1 (married) for the variable DVCURMS2 in this file, 1,500 for widowed, 528 for married but separated, etc. But as raw frequencies, they do not adjust for the sampling method and cannot be used to generalize to the Canadian population.

The middle column of figures for WEIGHTED 1 are based on applying the weight variable contained in the public use microdata file, namely, FWGHT. Notice two things about applying the weight variable. First, the total N becomes 20,525,561 or the total number of Canadians in 1991 who were 18 years of age or older. Secondly, notice that the percentages differ between the unweighted and weighted distributions. This is due to the adjustment for the sampling methodology. When sampling, Statistics Canada oversampled widows and divorced to ensure capturing people in these categories in the study. The weight variable corrects for this bias. Thus, 54.9% of Canadians in 1991 were married (not 50.1% as reflected in the unweighted frequency distribution.)

Finally, the rescaled weight variable (shown in Table 1 under the column, [weighted 2]) returns the overall N to 13,495 (the size of the sample.) Notice however, the percentages of the rescaled weight variable (WT) match those of Statistics Canada's weight variable (FWGHT). In other words, the weighted sample using WT corrects for the sampling method but also allows working with an N equal to the original sample size.

The large counts produced by the population estimate cause problems for analysts who want to perform traditional inferential statistical tests. Large N's (anything over a few hundred) will generate significant test results by the very nature of inferential statistics. One way to compensate for the scale of the weight variables used by Statistics Canada is to rescale the weight variable to the sample size. This ensures that adjustments for sampling methods are retained and also the N is maintained at the sample size rather than the population estimate, which was what was accomplished with [weight 2] above.

Calculating statistics without using one of the two weight variables produces biased results that prevents one from making generalizations to the full population. In other words, a weight variable needs to be applied when doing statistical analysis with the 1991 Canadian General Social Survey. Now, which weight variable? It doesn't really matter if FWGHT or WT is used. Either corrects for the sampling methodology. The choice becomes one more of ease in working with statistical tests.

Because of the complexity in some sample designs, a study may have more than one weight variable. To generalize for all employed workers, one weight variable may be used, while another weight variable may adjust the number of cases to represent all families. A good codebook will have a section describing the sampling methodology and will note the variables required to make generalizations about populations.

A.4 - Appendix on Weight Variables

Data Basics

Data Alphabet Soup
An alphabetized list of data-related acronyms.

AAPOR = American Association of Public Opinion Research (U.S.)

ACSPRI = Australian Consortium for Social and Political Research Incorporated

AHEAD = Asset and Health Dynamics Among the Oldest Old (a nationally representative longitudinal data collection, U.S.)

AIP = Archival Information Package (OAIS)

APDU = Association of Public Data Users. (U.S.)

ANES = American National Election Studies (national surveys carried out by the Survey Research Center (SRC) and by the Center for Political Studies (CPS) of the Institute for Social Research at the University of Michigan)

ASM = Annual Survey of Manufactures (U.S.)

BAS = Boundary and Annexation Survey (U.S. Bureau of the Census)

BEA = Bureau of Economic Analysis (U.S. Department of Commerce)

BIRON = Bibliographic Information Retrieval Online (The Data Archive, Essex, UK)

BJS = Bureau of Justice Statistics (U.S. Department of Justice)

BLS = Bureau of Labor Statistics (U.S.)

BST = Basic Summary Tabulations (Canadian aggregate census data)

CAI = Computer-Assisted Interviewing

CAPDU = Canadian Association of Public Data Users

CAPI = Computer-Assisted Personal Interviewing

CATI = Computer-Assisted Telephone Interviewing

CDC = Center for Disease Control (U.S. Department of Health and Human Services)

CDNet = Consortium Data Network (ICPSR's order facility for Official Representatives)

CIESIN = Consortium for International Earth Science Information Network

CISER = Cornell Institute for Social and Economic Research

CESSDA = Council of European Social Science Data Archives.

COPAFS = Council of Professional Associations on Federal Statistics

COPDAB = Conflict and peace data bank (COPDAB) [computer file] / Azar, Edward E.[principal investigator(s)].

CPS = Current Population Survey (U.S. Bureau of the Census) [computer file] / U.S. Department of Commerce. Bureau of the Census [principal investigator(s)].

CPS = Center for Political Studies (located in the Institute of Social Research at University of Michigan)

CPS = Centre for Population Studies (London School of Hygiene and Tropical Medicine, UK)

CRSP = Center for Research in Security Prices

CSES = Comparative Study of Electoral Systems

CTSI = Census Tract Street Index (U.S. Bureau of the Census)

CWBH = Continuous Wage and Benefit History

DAS = Data Analysis Systems (ICPSR's online analysis system using the SDA software developed at U. California Berkeley)

Data-PASS = The Data Preservation Alliance for the Social Sciences

DDA = Danish Data Archives

DDI = Data Documentation Initiative

DIP = Dissemination Information Package (OAIS)

DLI = Data Liberation Initiative (Canada)

DOT = Direction of Trade (IMF)

DPLS = Data and Program Library Service (University of Wisconsin)

Duraspace = A not-for-profit organization that provides leadership and innovation in open source and cloud-based technologies primarily for libraries, universities, research centers, and cultural heritage organizations.

EA = Enumeration Area (Canadian census geography)

ECPR = European Consortium for Political Research

ESRC = Economic and Social Research Council (UK)

ESRI = Environmental Systems Research Institute

GFS = Government Financial Statistics (IMF)

GIS = Geographic Information Systems

GSS = General Social Survey [computer file] / National Opinion Research Center [principal investigator(s)]

GSS = General Social Survey [computer file] / Statistics Canada [principal investigator]

HANES = Health and Nutrition Examination Surveys [computer file] / U.S. Department of Health and Human Services. National Center for Health Statistics [principal investigator(s)]

HRS = Health and Retirement Study (a nationally representative longitudinal data collection, U.S.)

IASSIST = International Association of Social Science Information Services and Technology

ICPR = Inter-university for Political Research (the original name of the ICPSR)

ICPSR = Inter-university Consortium for Political and Social Research

IFDO = International Federation of Data Organizations

IFS = International Financial Statistics

IMF = International Monetary Fund

IAED = International Archive of Education Data (an ICPSR topical archive)

IHSN = International Household Survey Network

ILSES = Integrated Library and Survey-data Extraction Service (project of Dutch, German, French, and Irish institutes)

IPEDS = Integrated Postsecondary Education Data System (computer file) / U.S. Department of Education, National Center for Education Statistics [principal investigator(s)]

IPUMS = Integrated Public Use Microdata Series (Historical Census Projects, University of Minnesota)

IRSS = Institute for Research in Social Science (University of North Carolina)

ISR = Institute of Social Research (the home of ICPSR at the University of Michigan)

ISSP = International Social Survey Program [computer file] / International Social Survey Program [principal investigator(s)]. Koln, Germany.

ISSR = Institute for Social Science Research (University of California, Los Angeles)

JPMS = Joint Program in Survey Methodology

LUCA = Local Update of Census Addresses (U.S. Bureau of the Census)

MIDAS = Manchester Information Datasets and Associated Services (UK)

MTF = Monitoring the Future (a nationwide survey of high school seniors, U.S.).

NACDA = National Archive of Computerized Data on Aging (an ICPSR topical archive)

NACJD = National Archive of Criminal Justice Data (an ICPSR topical archive)

NAPA = New Acquisitions Preservation Archive (at ICPSR)

NARA = National Archives and Records Administration (U.S.)

NBER = National Bureau of Economic Research (U.S.)

NCHS = National Center for Health Statistics (U.S.)

NCES = National Center of Education Statistics (U.S.)

NES = National Election Studies (located within the Center for Political Studies at The University of Michigan's Institute for Social Research)

NESSTAR = Networked European Social Science Tools and Resources

NHANES = National Health and Nutrition Examination Surveys [computer file] / U.S. Department of Health and Human Services. National Center for Health Statistics [principal investigator(s)].

NHDA = Netherlands Historical Data Archive

NIJ = National Institute of Justice (U.S.)

NIPA = National Income and Product Accounts

NIWI = Netherlands Institute for Scientific Services

NLS = National Longitudinal Survey of Labor Market Experience [computer file] Ohio State University. Center for Human Resource Research and Parnes, Herbert S. [principal investigator(s)].

NLS = National Longitudinal Study of the High School Class of 1972 [computer file] / U.S. Department of Education. National Center for Education Statistics [principal investigator(s)].

NLSY = National Longitudinal Survey [of Labor Market Experience] Youth

NNSP = National Network of State Polls (U.S.)

NORC = National Opinion Research Center (at University of Chicago)

NSD = Norsk samfunnsvitenskapelig datatjeneste (Norwegian Social Science Data Services)

OAIS = Open Archival Information System

OECD = Organization for Economic Cooperation and Development

OR = Official Representative (at ICPSR member institutions)

PAIMAS = Producer-Archive Interface Methodology Abstract Standard (OAIS)

PDI = Preservation Description Information (OAIS)

PI = Principal Investigator

PRA = Publication-related Archive (ICPSR's archive for replication studies)

PSID = Panel Study of Income Dynamics [computer file] / Morgan, James N., et al [principal investigator(s)].

PUMA = Public Use Microdata Area (U.S. census geographic unit)

PUMF = Public Use Microdata File (Canadian census)

PUMS = Public Use Microdata Sample (U.S. census)

PUS = Public Use Sample (U.S. census)

r-cade = Resource Centre for Access to Data on Europe

REIS = Regional Economic Information System (U.S. Dept. of Commerce, Economics and Statistics Administration, Bureau of Economic Analysis)

ROAD = Record of American Democracy [computer file] / King, Gary, et al [principal investigator(s)], Cambridge, MA: Harvard University [producer]

SADA = South African Data Archive

SAMHDA = Substance Abuse and Mental Health Data Archive (an ICPSR topical archive)

SAR = Sample of Anonymised Records (U.K. census microdata)

SDA = Survey Documentation and Analysis (software for online analysis used at ICPSR)

SIP = Submission Information Package (OAIS)

SLID = Survey of Labour and Income Dynamics (Statistics Canada panel survey)

SRC = Survey Research Center (located at the Institute of Social Research at the University of Michigan)

SIDOS = Swiss Information and Data Archive Service

SIPP = Survey of Income and Program Participation [computer file] / U.S. Department of Commerce. Bureau of the Census [principal investigator(s)].

SOSIG = Social Science Information Gateway (University of Bristol, UK)

SPSS = Statistical Package for the Social Sciences

SSRC = Social Science Research Council (U.S.)

STF = Summary Tape File (U.S. census aggregate data)

SWIDOC = Social Science Information and Documentation Centre (Netherlands)

TDR = Trusted Digital Repository

TRAC = Trustworthy Repositories Audit and Certification

UN = United Nations

UNIDO = United Nations Industrial Development Organization

WIStat = Women's Indicators and Statistics Database (U.N.)

XML = Extensible Mark-up Language

ZA = Zentralarchiv für Empirische Sozialforschung (Universität zu Köln)

B.8 - Appendix on Data Speak

Data Basics

Glossary

Scope: This glossary includes terms commonly used in managing data collections and providing basic data services. It does not attempt to cover all social science research terms or all computer terms.

Aggregate Data

Data that have been aggregated (summarized). Contrast with Microdata.

Anonymized Microdata

Data that records information about the unit of analysis (e.g., individual survey respondents) with certain information removed, altered, or otherwise changed so that that the confidentiality of the respondent will not be disclosed. A variety of methods can be used to anonymize data. For instance, telephone numbers, and names can be removed; geographic locations such as addresses and postal codes can be altered to indicate only a much larger level of geography such as a city or state.

Branching

(See Skip Pattern.)

Card

(See Card Image.)

Card Image

A format used for storing raw data. This format is a remnant of the time when data were input on punch cards. The cards had a physical limit of eighty characters per card and so the card image format uses exactly eighty characters of data, no more and no less, in each physical record. Usually a case or all the variables of a single respondent are stored on several of these physical records. Each such record, or "card," is numbered and stored in numerical sequence. Cards with the same sequence number (i.e., having a common format for the layout and contents of variables) are called a "deck;" thus cards are often referred to in documentation by their "deck number." Example: "The variable for age is stored in deck 01 in columns 10-11 and the variable for race is stored in deck 02 in column 10."

Case

In survey research, an individual respondent. Contrast with unit of analysis.

CATI

(See Computer Assisted Telephone Interviewing.)

Cleaning

To "clean" a data file is to check for wild codes and inconsistent responses (see Consistency Check); to verify that the file has the correct and expected number of records, cases, and cards or records per case; and to correct errors found.

Code

In most numeric data files, answers to questions are recorded with numbers rather than text and often even numeric answers are recorded with numbers other than the actual response. The numbers used in the data file are called "codes." Thus, for instance, when a respondent identifies herself as a member of a particular religion, a "code" of 1 might be used for Catholic, a 2 for Jewish, etc. Likewise, a person's age of 18 might be coded as a 2 indicating "18 or over." The codes that are used and their correspondence to the actual responses are listed in a codebook.

Codebook

Generically, any information on the structure, contents, and layout of a datafile. Typically, a codebook includes: column locations and widths for each variable; definitions of different record types ; response codes for each variable; codes use to indicate non-response and missing data; exact questions and skip patterns used in a survey; and other indications of the content of each variable. Many codebooks also include frequencies of response. Codebooks vary widely in quality and amount of information included. They may be machine-readable or paper copy or microfiche.

Column

In a data file, a single vertical column, each being one byte in length. Fixed format data files are traditionally described as being arranged in rows (i.e., lines) and columns. In a fixed format file, column locations describe the locations of variables.

Column Location

The precise location in a datafile of a variable expressed in column numbers, beginning with the first column in a physical record as column number 1.

Computer Assisted Telephone Interviewing (CATI)

A method of coding information from telephone interviews directly into a computer during the interview. CATI software usually has built in consistency checks, will not allow wild codes to be entered, and automatically prompts the interviewer for correct skip pattern questions.

Consistency Check

A process of data cleaning which looks for wild codes (q.v.) and other inappropriate values for the variables in a datafile. In a survey with branched questions a consistency check would look for response values in variables that should have been skipped.

Control Cards

A file of programming commands describing a datafile and written in the language of a particular statistical software. Useful because it provides variable locations, names and labels. Additional programming code must be added to perform analysis.

Cross Sectional Study

In survey research, a study in which data are obtained only once. Contrast with longitudinal studies in which a panel of individuals is interviewed repeatedly over a period of time. Note that a cross sectional study can ask questions about previous periods of time, though.

Data

Social science data are the raw material out of which social and economic statistics are produced. Social science data originate from social research methodologies or administrative records, while statistics are produced from data. Data are the information collected and stored at the level at which the unit of analysis was observed. Summaries of these data are usually “statistics.” Data must be processed to be of practical use. This compilation is accomplished with statistical software, which reads the raw data from a computer file.

Database

A collection of data, that have been processed and stored in the internal representation of a database software. The database software provides facilities for creating queries of the data to build tables and subsets, and, perhaps, facilities for doing statistical analysis of the data. Contrast with “datafile, and “raw data.”

Datafile

A computer file that contains raw data.

Dataset

A file or group of files associated with one part of a study. Files associated with a dataset might include a datafile, a machine-readable codebook, SPSS control cards, and other files related to the datafile.

Datatype

Data values in social science data files are often defined to be of a particular type such as integer, floating point, date, boolean, character, and so forth. The assignment of a datatype to a variable helps ensure consistency, accuracy, precision, scope, and range of the variable. In raw datafiles, the datatype of variables is defined separately in the codebook or other documentation. Statistical and database software store values in an internal representation that reflects the specific datatype of the value.

Derived Variable

Variables created by transforming other variables – that is, variables derived from other variables. The process of deriving new variables from existing variables is called "recoding." Example: an age variable containing a respondent's actual age in years is recoded to produce a derived variable, "eligible voter," with a code of "1" for all those 18 and over and a code of "2" for all those under 18.

Fixed Format

A file structure consisting of physical records of a constant size within which the precise location of each variable is based on the column location and width of the variable. Contrast with Free Format.

Flat File

(See Rectangular File.)

Free Format

A physical file structure that specifies the order of variables in a file and that they are delimited from each other by a special character or characters (usually a blank or other white-space). Free format files may have variable physical record lengths; when they do, they are typically delimited by an end-of-line character. Contrast with Fixed Format.

Frequencies

In survey research, the number of respondents who responded to each of the possible answers to a question. Often codebooks list the frequencies of response for each question. Also called "marginals."

Frequency File

A file that contains the frequencies for each question in a survey.

Hierarchical File

A hierarchical file is one that contains information collected on multiple units of analysis where each unit of analysis is subordinate to another unit. For example, if the physical housing structure is one unit, and individual persons within the structure is another unit, the person records are subordinate (e.g. related to) the housing unit. Such studies are sometimes referred to as having a relational structure.

Hierarchical File Structure

A format for storing hierarchical files. Each unit of analysis has its own record structure or record type. Different units of analysis do not necessarily have the same number of bytes or characters as the records for other units of analysis. In order to give such a file a common physical record length, short logical records are typically "padded" with blanks so that they will all be the same physical record length. A hierarchical file can be also be stored in a rectangular file. Typically, the hierarchical file structure is more space-efficient but more difficult to use.

Logical Record

All the data for a given unit of analysis . It is distinguished from a physical record because it may take several physical records to store all the data for a given unit of analysis. For instance, in Card Image data, a "card" is a physical record and it usually takes several "cards" to store all the information for a single case or unit of analysis.

Logical Record Length

A file storage format in which the length of a logical record is equal to the length of a physical record, which is constant. Thus, when the data for each case or unit of analysis is stored in a single physical record, the file structure is called "logical record length."

Longitudinal Study

In survey research, a study in which the same group of individuals is interviewed at intervals over a period of time. See also: panel study. Note that some cross sectional studies are done regularly (for instance, the *General Social Survey* and the *Current Population Survey* (Annual Demographic File) are conducted once a year), but different individuals are surveyed each time. Such a study is not a true longitudinal study. An Example of a longitudinal study is the *National Longitudinal Survey of Labor Market Experience*.

Margin of error

A measurement of the accuracy of the results of a survey. Example: A margin of error of plus or minus 3.5% means that there is a 95% chance that the responses of the target population as a whole would fall somewhere between 3.5% more or 3.5% less than the responses of the sample (a 7% spread). However, for any specific question, the margin of error could be greater or less than plus or minus 3.5%.

Marginals

(See Frequencies.)

Microdata

Microdata files are those that contain information on the individual unit of analysis rather than aggregate data. The U.S. *Census of Population and Housing* "Summary Files" contain aggregate data and consist of totals of individuals with various specified attributes in a particular geographic area. They are, in a sense, tables of totals. The Census PUMS (Public Use Microdata Sample) files, however, contain the data from the original census survey instrument with certain information removed to protect the confidentiality of the respondent.

Panel

A group of individuals who are interviewed more than once over time in a longitudinal survey.

Panel Study

A longitudinal study in which a panel of individuals is interviewed at intervals over a period of time. In general usage, the definitions of longitudinal study and panel study overlap. At least one author says that the term "panel study" is sometimes used for studies that are restricted to a short period of time or are limited to two or three interviews and "longitudinal study" is used for studies that last longer or include more interviews; but there are significant examples where this distinction is not accurate. In general, longitudinal studies involve panels of respondents and panel studies are longitudinal studies. Examples of panel studies include the *Survey of Income and Program Participation* (SIPP) and the *Panel Study of Income and Dynamics* (PSID).

Physical Record

A chunk of data that has a specified and constant size in bytes or that is clearly delimited from other records by and end-of-line character or sector of a disk or other means identifiable to a computer program reading the file.

Physical Record Length

The length, in bytes, of a physical record.

PI

An abbreviation for Principal Investigator.

Population

See Universe.

Portable

In computer usage, a file or program is "portable" if it can be used by a variety of software on a variety of hardware platforms using any operating system. Numeric data files written as plain character format files are fairly portable.

Principal Investigator.

The person or organization responsible for a study; equivalent to "author" in bibliographic citations.

Raw data

Data that reside in a datafile, unprocessed by statistical software. The term raw data is used to refer to a portable, plain-text format of a collection of data, suitable, along with documentation, for use on any computing platform with any statistical software.

Record

Depending on the context, "record" may refer to a physical record or a logical record of a raw datafile or a case in a study.

Record Length

Depending on the context, the length in bytes (i.e., columns) of a physical record or a logical record of a raw datafile

Record Type

A record that has a consistent logical structure. In files that include different units of analysis, for instance, different record types are needed to hold the different variables. For example, one record type might have a variable for income in one column and another record type might have a variable for household size in that same column. The codebook will describe these different structures and how to determine which is which so that you can tell your statistical software how to interpret that particular column as income or household size.

Rectangular File

A physical file structure. A rectangular file is one that contains the same number of card-images or the same physical record length for each respondent or unit of analysis.

Relational Structure

A study that includes different units of analysis, particularly when those units are not arranged in a strict hierarchy as they are in a hierarchical file, has a relational structure. Note that the data could be arranged in several different physical structures to handle such a data structure. For instance, each unit of analysis might be stored in a separate rectangular file with identification numbers linking each case to the other units; or, the different units of analysis might be stored in one large file with a hierarchical file structure; or the different units could be stored in a special database structure used by a relational data base management system such as Oracle. An example of a study with a relational structure is the Survey of Income and Program Participation which has eight or more record types ; these record types are related to each other but are not all members of a hierarchy of membership. For instance, there are record types for household, family, person, wage and salary job, and general income amounts.

Respondent

In survey research, the person responding to the survey questions.

Response codes.

Typically responses to questions are "coded" by assigning numeric codes to each possible response. Thus a "yes" might be coded "1" and a "no" "2"; female respondents might be indicated by a "1" and male respondents by a "2"; each state or county might be assigned a numeric code.

Sample

A selected subset of a “population.”

Sampling Method

The methodology used to select cases from a universe. Some common sampling methodologies include random, probability, stratified, cluster.

Secondary data analysis

Analysis of data collected by another researcher.

Skip Pattern

In survey research, the sequence of questions asked and skipped. For instance, persons who answer one question that indicates they did not vote in the last election would trigger a "skip" so that the interviewer would not ask those respondents questions about how they voted in the last election.

Statistical software

Software designed specifically for statistical analysis. Typical features of statistical software include: built-in mathematical and statistical formulas and routines; a programming language for reading raw data, analyzing data, and creating graphs, charts and other output; and the ability to handle very large quantities of data (often with no built-in limits as to the number of cases, number of variables, or size of datafile). Statistical software is typically designed so that the structure of the data reflects the unit of observation or unit of analysis.

Statistics

Statistics are produced from data. *Statistics are frequently used to condense a large amount of information into a few numbers and in this context, provide a concise, descriptive summary.* The dictionary definition of "statistics" refers to numeric indicators of nations. Popular usage of the term points to numeric summaries that condense information, or numbers that are used to make comparisons, or numbers that portray relationships or associations. The term statistics also refers a formal discipline of study. The field of statistics is the science of generalization. Built upon theories of probability and inference, statistics support the making of broad generalizations from a smaller number of specific observations.

Study

All the information collected at a single time or for a single purpose or by a single principal investigator. A study may consist of one or more datasets and one or more files.

Text File

In computer usage, any file written in pure character format. Sometimes called a "plain text file."

Time Series

Observations of a variable made over time. Many economic studies consist of are time series data. Time series, of a sort, can also be constructed from a cross sectional study if the same questions are asked more than once over time. See also longitudinal study.

Undocumented Code

(See Wild Code.)

Unit of analysis

The basic observable entity being analyzed by a study and for which data are collected in the form of variables. Although a unit of analysis is sometimes referred to as the case or "observation," these are not always synonymous. For instance, in public opinion polls, the unit of analysis is usually a single person and the answers to the survey questions by one person constitute a "case." In a census, however, a "case" could be considered the household because all the data for one household is collected on one survey instrument; the household "case" may contain different variables for the different units of analysis: a physical housing structure, a family within the structure, a person within the family.

Unit of observation

When social science methodology is used to collect data, the entity which is observed or about which information is collected is the unit of observation. The unit of observation is the same as the unit of analysis when the generalizations being made from a statistical analysis are attributed to the unit of observation (i.e., the objects about which data were collected and organized for statistical analysis). While the units of observation and analysis are often the same, the wealth of secondary data sources creates opportunities to conduct analyses with data from multiple units of observation. This is probably most recognizable in GIS research.

Example: A major national study uses a form that collects information about each person in a dwelling and information about the housing structure. Therefore, this study collects data for two units of observation: persons and housing structures. From these data, different units of analysis may be constructed: Household could be examined as a unit of analysis by combining data from people living in the same dwelling. Family could be treated as the unit of analysis by combining data from all members in a dwelling sharing a familial relationship. This expresses how the unit of analysis can be constructed from units of observation consisting of some type of relationship constructed by time, space or social properties.

Universe

The entire collection of items or the entire group that a researcher is interested in analyzing or about which the researcher wishes to draw conclusions. Normally, the universe is so large that it cannot be easily measured comprehensively, so a sample of the population is used as surrogate for the whole. Sometimes also called “population.”

Variable

In social science research, for each unit of analysis, each item of data (e.g., age of person, income of family, consumer price index) is called a variable.

Wave

In a panel study, a wave is the interviewing period during which the entire panel is questioned and asked the same questions. Typically, a panel study consists of several waves. Waves are important because each wave typically covers a different time period and, often, different topics.

Weight

In survey research, a number associated with a case or unit of analysis; the weight is used as a measure of the relative significance of the variables of that case when making estimates for the entire population. When a probability sample is used, there is often a chance that some elements of the population are under or over represented in the sample. In order to allow more accurate estimates of a complete population, therefore, "weights" are assigned to each case and used to adjust the overall results to more closely conform to the total population.

Wild Code.

In survey research, "wild" codes are codes that are not authorized for a particular question. For instance, if a question that records the sex of the respondent has documented codes of "1" for female and "2" for male and "9" for "missing data," a code of "3" would be a "wild" code, sometimes called an "undocumented code."

Bibliography

The literature surrounding issues related to research data is extensive, varied, and growing rapidly. We provide links to bibliographies for those who wish to explore this literature at

http://3stages.org/class/2012/bibliographies.shtml

- A few classic articles and collections
  http://3stages.org/class/2012/classics.shtml

  Data collection, preservation, service, and management have a long history in the Social Scienes. We collect here a few of the early classics of this literature.

- Bibliography of selected works Compiled by Margaret O. Adams. (IASSISST)
  http://www.iassistdata.org/publications/bibliography.html

  This bibliography is an introduction to literature from the mid-1970s to the mid-1990s that represents the work of the international social science data community. It is by no means inclusive. The bibliography was prepared for a session at the annual conference of the Society of American Archivists (SAA), August 30, 1996, that focused on the potential for partnerships between traditional and data archivists. Includes a selection of articles from the IASSIST Quarterly*, Vol. 1 (1977) - Vol. 18 (1994)*

- Research Data (Zotero)
  https://www.zotero.org/groups/research_data/items

  A Zotero group bibliography managed by your instructors. Includes Information related to the preservation and management of research data, and the provision of data services, particularly in libraries.

- Data Preservation
  http://3stages.org/class/2012/datapreservation.shtml

  Selected articles on Data Preservation.

- Big Data
  http://3stages.org/class/2012/bigdata.shtml

  Selected articles on "Big Data."

- Publications tagged by class (Bibsonomy)
  http://www.bibsonomy.org/tag/icpsr2012?resourcetype=publication

  Publications tagged by 2012 class.