From data to information: the role of data analysis in the creation of information

Laine Ruus, University of Toronto. Data Library Service

2009-03

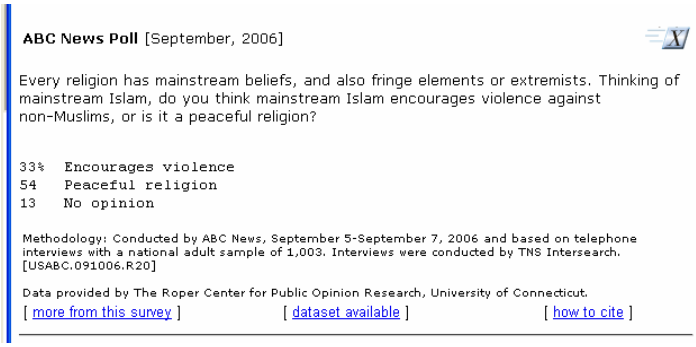
Vignette:

A student in religious studies was looking for statistics on US perceptions of Islam as a religion. By dint of searching the web, and the iPOLL public opinion poll question archive at the Roper Centre, we managed to identify among other things, an ABC News poll from September 2006, which contained the following question:

> *Every religion has mainstream beliefs, and also fringe elements or extremists. Thinking of mainstream Islam, do you think mainstream Islam encourages violence against non-Muslims, or is it a peaceful religion?*

The iPOLL record provided by the Roper Center contains the following information:

Figure 1: ABC News poll question, Sept. 2006



ABC News Poll [September, 2006]

Every religion has mainstream beliefs, and also fringe elements or extremists. Thinking of mainstream Islam, do you think mainstream Islam encourages violence against non-Muslims, or is it a peaceful religion?

33% Encourages violence
54 Peaceful religion
13 No opinion

Methodology: Conducted by ABC News, September 5-September 7, 2006 and based on telephone interviews with a national adult sample of 1,003. Interviews were conducted by TNS Intersearch. [USABC.091006.R20]

Data provided by The Roper Center for Public Opinion Research, University of Connecticut.

[more from this survey] [dataset available] [how to cite]

Source: iPOLL database <http://www.ropercenter.uconn.edu/data_access/ipoll/ipoll.html>

Essentially, what is being provided by iPOLL is the percentages, not the counts, of the respondents to the poll who felt that Islam encourages violence versus is a peaceful religion, respectively. The student then went on to ask, so who are these people who feel it is a peaceful versus violent religion. What distinguishes them? What makes the difference between one opinion and its diametric opposite? *[Research is after all an iterative process.]*

One possible answer could have been provided by doing additional searching of additional bibliographic databases, and the web, looking for newspaper articles, or other articles that had used the same or similar data; but a search on Google for “abc news islam september 2006” produced 606,000 hits. A very different, but much simpler and more direct alternative was to locate a copy of the original raw poll data, and produce the requested descriptive statistics reflecting the relationship (if any) between opinion on Islam and other characteristics of the respondents to the poll.

One of the hits that Google produced was a link to the ABC News poll in question in the ICPSR collection. Over the course of the past few years, ICPSR has been loading raw data files into SDA, an interactive interface to data developed at University of California, Berkeley, which allows users to interactively generate descriptive statistics (as well as inferential statistics) and even do simple modelling, without having to have access to and expertise in using special purpose statistical software such as SAS, SPSS or Stata. The ABC News poll was among those data files which have been loaded into SDA on the ICPSR server. With very little training, the student was able to produce and interpret tables reflecting the relationship of such characteristics as gender, age group, education, and religious beliefs, and opinion about Islam, able to judge whether those relationships were statistically significant, and able to document her interpretations in her paper. She found, for example, that gender (fig. 2) and age (fig. 3) have no appreciable (read statistically significant) influence, on average, on opinion about Islam,

Figure 2: Cross-tabulation: opinion of Islam by gender:

Cells contain: -Column percent -Weighted N		Q921		
		1 Male	2 Female	*ROW TOTAL*
Q18	1: Encourages violence	**38.5** 160	**36.8** 166	***37.6*** *326*
	2: Peaceful religion	**61.5** 255	**63.2** 285	***62.4*** *541*
	COL TOTAL	***100.0*** *415*	***100.0*** *452*	***100.0*** *867*
Means		1.61	1.63	1.62
Std Devs		.49	.48	.48
Unweighted N		420	462	882

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Summary Statistics						
Eta* =	.02	Gamma =	.04	Chisq(P) =	.26	(p= 0.61)
R =	.02	Tau-b =	.02	Chisq(LR) =	.26	(p= 0.61)
Somers' d* =	.02	Tau-c =	.02	df =	1	

*Row variable treated as the dependent variable.

Source: ABC News poll, September 2006

What the table in figure 2 tells us is that almost 62% of male respondents to the poll consider Islam a peaceful religion, compared with a little over 63% of female respondents. In other words, both male and female respondents were half again more likely to consider Islam peaceful, than to consider it to be encouraging violence. Is this statistically significant. In one word, no. In the bottom right of figure 2 are two lines labelled ‘Chisq(…); beside each in brackets is a note (p=0.61). This ‘p-value’ is the probability of being wrong if one were to say that there is a relationship between gender and opinion on the peacefulness/violence of Islam. Translated into a percentage, that represents a 60% chance of being wrong.

Figure 3: Cross-tabulation: opinion of Islam by age:

Cells contain: -Column percent -Weighted N		AGEBREAK					
		1.00 18-29	2.00 30-39	3.00 40-49	4.00 50-64	5.00 65+	ROW TOTAL
Q18	1: Encourages violence	**29.5** 52	**37.8** 60	**36.2** 62	**38.4** 79	**41.4** 54	***36.5*** *308*
	2: Peaceful religion	**70.5** 124	**62.2** 99	**63.8** 110	**61.6** 126	**58.6** 77	***63.5*** *537*
	COL TOTAL	***100.0*** *176*	***100.0*** *160*	***100.0*** *172*	***100.0*** *205*	***100.0*** *132*	***100.0*** *845*
Means		1.71	1.62	1.64	1.62	1.59	1.64
Std Devs		.46	.49	.48	.49	.49	.48
Unweighted N		94	145	167	278	177	861

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Summary Statistics

Eta* =	.08	Gamma =	-.10	Chisq(P) =	5.62	(p= 0.23)
R =	-.07	Tau-b =	-.06	Chisq(LR) =	5.72	(p= 0.22)
Somers' d* =	-.05	Tau-c =	-.08	df =	4	

*Row variable treated as the dependent variable.

Source: ABC News poll, September 2006

The respondent's own religious beliefs, however, do influence opinion. both what religion the respondent professes (figure 5), as well as how important religion is in the respondent's own life (figure 4).

Figure 4: Cross-tabulation: opinion of Islam by importance of religion in respondent's own life:

Cells contain: -Column percent -Weighted N		Q38			
		1 Very important	2 Fairly important	3 Not very important	ROW TOTAL
Q18	1: Encourages violence	**42.9** 202	**31.3** 64	**30.6** 53	***37.6*** *320*
	2: Peaceful religion	**57.1** 270	**68.7** 141	**69.4** 120	***62.4*** *531*
	COL TOTAL	***100.0*** *472*	***100.0*** *205*	***100.0*** *174*	***100.0*** *851*
Means		1.57	1.69	1.69	1.62
Std Devs		.50	.46	.46	.48
Unweighted N		476	214	181	871

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Summary Statistics

Eta* =	.12	Gamma =	.21	Chisq(P) =	12.91	(p= 0.00)
R =	.11	Tau-b =	.11	Chisq(LR) =	13.02	(p= 0.00)
Somers' d* =	.10	Tau-c =	.12	df =	2	

*Row variable treated as the dependent variable.

Source: ABC News poll, September 2006

In the table in figure 4, the p-value of the Chisqs is 'p=0.00'; this, translates into a less than 1% chance of being wrong if you claim that there is a relationship between the importance of religion to the respondent and their opinion on the peacefulness/violence

of Islam. Any chance of being wrong that is less than 5% (ie, a p-value of 0.05) is considered an acceptable risk in the social sciences. So this relationship is considered to be statistically significant.

Figure 5: Cross-tabulation: opinion on Islam by respondent's religion:

Cells contain: -Column percent -Weighted N		RELNET					
		1 Protestant	2 Catholic	3 Christian (Non-Protestant)	4 Other Non-Christian	5 None	*ROW TOTAL*
Q18	1: Encourages violence	**39.5** 151	**42.8** 71	**33.2** 33	**18.0** 12	**35.2** 42	***37.1*** *309*
	2: Peaceful religion	**60.5** 232	**57.2** 95	**66.8** 67	**82.0** 55	**64.8** 77	***62.9*** *525*
	COL TOTAL	***100.0*** *383*	***100.0*** *166*	***100.0*** *100*	***100.0*** *67*	***100.0*** *119*	***100.0*** *834*
Means		1.61	1.57	1.67	1.82	1.65	1.63
Std Devs		.49	.50	.47	.39	.48	.48
Unweighted N		388	188	87	76	119	858

Color coding:	<-2.0	<-1.0	<0.0	>0.0	>1.0	>2.0	Z
N in each cell:	Smaller than expected			Larger than expected			

Summary Statistics

Eta* =	.13	Gamma =	.11	Chisq(P) =	14.95	(p= 0.00)
R =	.08	Tau-b =	.07	Chisq(LR) =	16.14	(p= 0.00)
Somers' d* =	.05	Tau-c =	.07	df =	4	

*Row variable treated as the dependent variable.

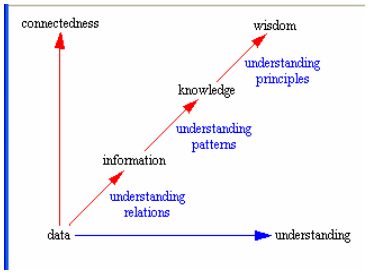
Source: ABC News poll, September 2006

New frontiers and the DIKW

What is the point of this vignette? The point is to illustrate the notion that as old frontiers for reference librarians have become successively tamed (by folks like Web of Science, LexisNexis, Google , etc.) new frontiers are opening. One of those new frontiers is the increased accessibility of computer-readable data, and the concomitant challenge to librarians to become actively involved in, not just locating existing information, but in the creative process of generating information (eg descriptive statistics) as users require it, and in training users to create information from raw data themselves, to evaluate it, and to interpret it, and in general, raising the level of numeracy in both the profession and the populations we serve.

A paradigm that I generally find useful in this context is the data – information – knowledge—wisdom model, or DIKW for short, formalized by Ackoff and diagrammed by Bellinger (2004) as.

Figure 6: DIKW model



Source: Bellinger, Durval and Mills 2004

Librarians traditionally deal with collections at level 2, the acquisition, management and location of information. Reference librarians in particular, have traditionally been specialized in understanding the structure of publishing, in locating books, articles, and abstruse facts by a variety of means. In the course of so doing, at least when I entered the ranks of academic reference librarians in the early 1970s, we learned our reference, as well as our general, collections. We perused carefully bibliographies compiled by others, and frequently, compiled bibliographies of our own. We chased citations, and poured over the (card) catalogues of other libraries, especially those of very specialized or very comprehensive collections, which were so popular that they were published in print form and sold to other libraries for their reference collection. What academic library in North America doesn't have a copy of the Peabody Museum catalogue and it's supplements? We collected handbooks, directories, and statistical compendia so as to have easy access to statistics compiled by others from data collected by others. We taught users how to locate and use the information in our collections, evaluate, manage and interpret it so as to enable them to produce knowledge, which once published, we added to our collections to aid in the production of yet more knowledge.

Those days are going, if not already gone. Frankly, Web of Science, Google Scholar, and other large indexing projects now do a far better job of generating the beginnings of a comprehensive bibliography of almost anything, than I could ever have done thumbing through dusty copies of the Peabody Museum catalogue. In the race for excellence in searching through millions of records looking for combinations of strings of characters, computers outperform us all. And the ubiquitous access to library catalogues, and aggregators of library catalogues (e.g. OCLC, LIBRIS, etc.) on the web has obviated the need to publish those large catalogues.

One area which the big search engines have not yet mastered adequately is that of finding statistics published, not in traditional media such as books, periodicals, government documents, and the like, but rather in on-line dynamic statistical (numeric databases. Interfaces to statistical databases are approximately at the same level of development that bibliographic databases were in the late 1970s, early 1980s. Each database producer is experimenting with its own interface (compare SourceOECD, WDI, UNdata, and Datastream, to name only a few). The rise of large enterprises that concentrate on collecting many databases from many sources under one umbrella interface parallel the early rise of eg Dialog and its competitors back when reference librarians were intermediaries in searching bibliographic databases, and our users were backseat drivers. However, I am confident that in the not too distant future, there will also be comprehensive search capability across large quantitative databases, such that one will be able to comprehensively search most or all of the major numeric databases, such as Datastream, CRSP, Compustat, DRI basic economics, WDI, etc. all in interface.

LexisNexis is trying with it's Statistical database, but it has a long way to go before it becomes useful for anything other than finding tables in books or periodicals.

What I have been discussing in the past few paragraphs, however, is merely an extension of the kinds of resources that have already been developed for prose-based information, ie a way of searching already published textual information, and is in no way a new frontier. These search engines, however, are at a distinct disadvantage when attempting to locate information that does not yet exist, namely the almost unlimited information that can yet be produced from existing data.

My key argument, therefore, is that the next generation reference librarian needs to step back a stage in the DIKW model, and become as familiar with data resources as we were in my day with our print reference collections, and now with bibliographic databases. Working with data is no longer a matter of looking up pre-existing information, or finding existing statistics. Rather it is about creating descriptive or inferential statistics using existing data resources, of training users to create, evaluate, and interpret statistics. The tables presented in figures 2 through 5 do not exist other than in this paper; they were created on-the-fly for the information purposes of a user (and later of this paper), and then ceased to exist. However, the key here is that these tables and the bibliographic citations contain sufficient information (metadata) to make it possible for another researcher to both access the same raw data, and re-create the same tables, ie corroborative analysis. This process of working with data rather than the information derived from it is the new frontier.

Looked at from another point of view, data are a primary source, to the secondary and tertiary sources represented by published papers and monographs. Wikipedia defines a primary source as:

> **primary source**[1][2] is a term used in a number of disciplines. In historiography, a primary source (also called **original source**) is a document, recording or other source of information (paper, picture,....etc) that was created at the time being studied, by an authoritative source, usually one with direct personal knowledge of the events being described. It serves as an original source of information about the topic. Primary sources are distinguished from secondary sources, which often cite, comment on, or build upon primary sources.[3]
>
> *Source: Wikipedia, < http://en.wikipedia.org/wiki/Primary_source>, accessed 2009-03-02*

Wikipedia further goes on to inform us that:

> Although many documents that are primary sources remain in private hands, the usual location for them is an archive. These can be public or private. Documents relating to one area are usually spread over a large number of different archives. These can be distant from the original source of the document.
>
> *Source: Wikipedia, < http://en.wikipedia.org/wiki/Primary_source>, accessed 2009-03-02*

The above, however, is not true for data…data files can almost be described as the primary sources the traditional archives (with a few exceptions) forgot.

A brief history of data management in the social sciences

The collection of large amounts of raw data has it's roots in the movement to conduct population censuses in the late 1600s and into the 1700s, as an aid to state governance, the administration of taxes, and forecasting population change and growth. Demographers attempted to synthesize and analyze these vast amounts of data. all without benefit of computers. By the early 1900s, public opinion polling, initially an effort to predict the outcome of elections, was becoming a more scientific endeavour, and large national surveys of population on a variety of topics began to be more prevalent through the 1940s. This may partly have been because while a small sample survey of about 600 to 1,000 respondents, can, although painfully, be analyzed with a card sorter, by the 1940s the precursors to mainframe computers were beginning to be available at major academic and government institutions, making the analysis of large collections of records much more feasible.

The management and preservation of data has, at least in the social sciences, a history that stretches back to 1946, when the Roper Center for Public Opinion Research was created at Williams College as a repository of public opinion polls . It was followed by the establishment of data archives at e.g. the University of Wisconsin , Madison, the Universität zu Köln, the University of Michigan, the University of California, Berkeley, the University of Amsterdam, the University of North Carolina, Chapel Hill, and the University of Essex through the 1950s and 1960s. By the 1970s, managing collections of data was a sufficiently prevalent activity that the first associations began, including CSSDA, IASSIST, CESSDA, and IFO.

Notice that none of these developments happened in libraries. A Ford Foundation funded report by Lucci, Rokkan and Meyerhoff, for the Columbia University School of Library Service published 1957 (Lucci, Rokkan, Meyerhoff, 1957) , did propose the notion of library-based management of data collections, but fell largely on deaf ears in the realms of library administration Twenty years later, in a PhD dissertation for University of California, Berkeley, in 1974, Howard White (White, 1957) argued that libraries should become at least tangentially involved in data management, to the point of including the documentation about data files (ie metadata), aka codebooks, in library collections in order to improve their accessibility. In Canada, the Data Clearinghouse for the Social Sciences project collected just metadata, and then was terminated; as a result, we have the catalogue of descriptions of data files that the DCH published, but most of the data files described therein have been lost.

The last 3-4 decades have seen enormous advances in technology related to data analysis and data management. When I entered the data management field in the early 1970s, my learning curve included figuring out how to use a card punch machine , a card reader, and 7-track magnetic tapes. We did have early versions of statistical software (SPSS), command-line driven from punched cards, and a very kind mainframe operating system (MTS, from the University of Michigan). By the late 1970s we were using a dialup modem (into which one plugged the telephone receiver), and 9-track magnetic tapes at 1600 bpi were becoming more common. By the 1980s, early IP-based protocols were beginning to allow us to use e-mail and ftp, which eventually replaced magnetic tapes as a transportation medium. Also, during the 1980s, we got our first pcs. Early 1990s it was the time for gopher servers, and cd-roms, and by the mid-1990s, "the Web" began. Also, in 1995 a project to define a DTD (document type definition) for metadata, known as the

Data Documentation Initiative (DDI) was begun. In the scant 14 years since the beginning of the DDI, there have now been developed three major projects, which support the emerging DDI standard, and which provide an interactive, web-based interface to raw data: Nesstar, SDA (Survey Data & Analysis), and the Virtual Data Center (part of the Dataverse project).

In the almost 40 years that I have been involved with data management, we have gone from analyzing data including producing descriptive statistics, using punched cards and overnight tape runs, and the interpretation of statistics an arcane science, to personal computers on every desktop and in every backpack, and interactive, remote, on-line interfaces that are easier to use than MS Excel, but deliver much the same analytic capabilities as those pc desktops. The development of these interfaces represents a radical change in the accessibility of data, rather than information, to researchers, and to the population at large.

Models of organizing data management

There have emerged two major models of data management. In Canada and the United States, large academic institutions with an emphasis on quantitative methods in key disciplines, tend to 'grow' a local data service. In addition, in the United States, there is a network of state data centres, as well as a few large data archives, that collect and manage data from multiple sources: the Roper Center for Public Opinion Research (now at the University of Connecticut), the Inter-University Consortium for Political and Social Research at the University of Michigan, and the National Archives and Records Administration's Electronic and Special Media Records. Canada has no equivalent institutions to any of these data archives with national stature. In Canada, all but one of the local data services is located organizationally in a library,. In the United States, I estimate (based on the institutional affiliation of IASSIST members) that now about 40% of data services are administratively located in libraries. Outside Canada and the United States, it is important to note, the data management model is very different. The tendency is to the establishment of a central national data archive, usually funded by a social science research council. As far as I am aware none of those data archives is located administratively or physically in a library, although one is located administratively in the national archives (the Dansk Data Arkiv in Denmark).

The European model has an advantage in the synergies that arise from a number of persons with expertise in related fields working in close physical proximity. The North American model has the advantage of putting user services close to the users, rather than at some remote national archive. However, those who 'do data' at academic institutions in North America tend to be specialized, and relatively isolated – services to which one refers users, when all else fails. I am suggesting that, now that it is becoming possible for every reference librarian (public and school, as well as academic) to provide at least some reference services involving data, that previously arcane set of skills should become as common-place as the ability to search bibliographic databases and chase citations.

Additional synergies arise from those providing data-based reference services also knowing the published literature.Increasingly, researchers find current statistics, or statistics in computer-readable formats (the web, statistical databases, etc.) but want to also locate earlier statistics, which are often available in print form only.

New skills for reference librarians

Conquering the new frontier of the creative use of data to both inform and train our users will not be without costs. For reference librarians to achieve a level of comfort will require the development of numeracy skills above those currently common in the population at large.

Knowing a collection is crucial, as will any reference activity. Currently, there are metadata standards emerging for data, and eventually, finding data sets with the requisite mix of variables for a table or analysis will become much simpler. But locating the right data is not merely a matter of locating the right variables. It is also a matter of identifying from whom data could be collected, and what agencies would have the authority and wherewithal to do so, whether the population from which the data are collected is appropriate to the user's needs. Were the data collected in the right time period, for the right level of geography? What weight variables are available, and do they adjust the sample to the demographic proportions in the population at large, or do they adjust the number of responses to the population size? Are the variables coded appropriately, ie is income a continuous variable, or a categorical one? What about age, height and weight, etc.?

Next, the numerate reference librarian remembers enough from his/her library school statistics courses to know the difference between a categorical and continuous or numeric variable, and the types of descriptive statistics that are appropriate for each type. For example, if income is coded exactly as reported to the IRS (ie the actual annual income) it would be utterly meaningless to attempt to produce a cross-tabulation. Instead, the numerate librarian will know to either do a comparison of means, or to collapse income to income ranges. Further the numerate librarian may know (or be able to suggest where to find out) how to interpret eg a chi-square statistic and it's associated probability, as well as other common measures of statistical significance, magnitude, direction, etc. And finally, the numerate librarian should be able to make suggestions and raise cautions about appropriate versus inappropriate methods of data visualization (bar chart, pie chart, line graph, thematic map).

I am not suggesting that the next generation of reference librarian needs to become a statistician or master the minutia of statistical interpretation. But I am suggesting that we now have an opportunity to become conversant with common types of descriptive statistics, their creation, the definition and analysis and interpretation of relationships among variables, or in general, more numerate. Statistics are, by and large, taught in our library schools, but general numeracy is not.

The rewards, I suggest, will lie in improved and expanded reference services, both for users without access to a local data service as well as to those with, in fewer but more appropriate referrals. Conquering this new frontier should result in better exposure to researchers and other users to data resources when they are appropriate, and better referrals to local data services; those codebooks that Howard White suggested be included in library collections are now almost ubiquitously available freely on the web. And for reference librarians themselves, there will be increased personal satisfaction at no longer having to refer the statistics questions, as well as a practical use for those mandatory statistics courses.

Conclusion:

Technological and software developments of the past 15 years have resulted in the development of tools which now make it practicable for non-specialists to interactively engage in the iterative creation of descriptive and inferential statistics, on an as required basis. For reference librarians, this means becoming comfortable with collections at the level of data in the DIKW model, rather than as previously, at the level only of information. It means increasing ones numeracy levels, increasing one's comfort with how data are collected ,how statistics are derived from them, and how relationships among variables are defined, measured, interpreted and displayed. This opens a whole new frontier for reference librarians, quite beyond the almost infinite variations in searching for pre-existing information in its myriad sources, and turns the reference librarian from a passive consumer of information to an active participant in the creative process of information creation.

Bibliography

ABC News. ABC News 9/11 anniversary poll, September 2006 [computer file]. ICPSR04665-v2. Horsham, PA: Taylor Nelson Sofres Intersearch [producer], 2006. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2008-01-24. doi:10.3886/ICPSR04665
[accessed via ICPSR SDA server, 2009-03]

Bellinger, Gene, Durval Castro, Anthony Mills. Data, Information, Knowledge, and Wisdom [computer file], 2004 < http://www.systems-thinking.org/dikw/dikw.htm> Accessed 2009-03-05

Data Documentation Initiative: http://www.ddialliance.org/

Lucci, York, Stein Rokkan and Eric Meyerhoff. A library center of survey research data: a report of an inquiry and a proposal. New York: Columbia University. School of Library Service, June 1957.

Survey by ABC News, September 5-September 7, 2006. Retrieved March 9, 2009 from the iPOLL Databank, The Roper Center for Public Opinion Research, University of Connecticut. <http://www.ropercenter.uconn.edu/ipoll.html>.

White, Howard Dalby. Social science data sets: a study for librarian. PhD thesis, University of California,Berkely, 1974.