Scientific Data and Social Science Data Libraries

by David Barber, Coordinator of Information Technology & Jan Zauha, Documents and Data Support Librarian Graduate Library University of Michigan*

There is a vast amount of quantitative information available in electronic form. Social science data makes up less than half that amount. The other, larger half is scientific data. While university libraries have made a considerable investment in so ci al science data, little has been done about scientific data. If administrators, librarians, or others believed that more attention should be paid to scientific data, one of the suggestions that might naturally arise is that social science data specialists should be involved. Though some common ground between the se areas should be acknowledged, the existance of very substantial differences must also be recognized. Those differences are especially significant because coping with them will require an investment of staff and financial resources by the data library.

What Is Scientific Data?

There are many ways to typify scientific data. It could be identified in terms of the disciplines that create and use it. Scientific data is then the data of chemists, geologists, physicists, and others. This type of data could also be identified by the type of phenomena that are measured. Scientific data measures physical, not social phenomena. For example, there are the kind of scientific measurements made in labs or in the field that yield a single n umber or a series of numbers describing physical phenomena like atomic weights, inches of precipitation, and voltage. There are measurements made of radiation reflected by or passing through objects such as are made by remote sensing or MRI scans. In ad dition, scientific data may be the result of a series of equations that model the behavior of a turbulent fluid and the numbers that are produced when that simulation is run.

Scientific data information is found in many locations. It is available from research institutes, commericial firms, and governmental data archives. It is often stored on single-user computer systems either on a scientist's desktop or in a science lib ra ry. The data are also often stored on departmental servers.

Social Science and Science Data: Is There a Connection?

There are things that social science data libraries share with the world of scientific data. The most basic, and perhaps important, similarity in the minds of non-data users is that both are numeric types of information. It is assumed that the numeric information resources of scientists must be like those of the social science librarian, and that both share a common understanding of the world of math and statistics.

Beyond these perceived similarities, there are some real links between science and social science data collections. Some scientific datasets are structured like rectangular social science data files. The activity of subsetting the data is then similar . Further, scientific datasets often are received as ASCII flat files which must be described by some kind of a programming language and converted into a standard file format prior to being used. Documentation describing the data and methodology usually a ccompanies the dataset. While the traditional tools of libraries designed for organizing text and citations into manipulable databases cannot do the same for numbers, certainly the tools of the social scientist can. The statistical packages used by the social scientist can re tr ieve and display numeric information from datasets. These tools do in fact link social scientists and some scientists: SAS and SPSS are used by both groups. In addition, scientific and social science data users are often linked because they make share d use of the computing center staff, and statistical consultants who support statistical software. When data librarians start to provide GIS data, they also bring themselves closer to the sciences. Geographers have always bridged the gap between the social sciences and sciences by their use of both social science and earth science data. Many GIS da ta collections which provide valuable boundary information also come with earth science data, and as a result become a common resource for both social scientists and earth scientists. Geographers also want access to images from remote sensing which once ac quired will also be used by other researchers from the physical sciences.

What Are the Differences?

These similarities and linkages between the worlds of science and social science data may spur data librarians to investigate scientific data more closely. This investigation should lead to the recognition of some significant differences between the tw o types of data. These differences fall into three main groups: first, there are many unique problems caused by the structure of scientific data; second, a different body of knowledge is required; and finally, there are unique needs for visualization of sc ientific data. Data Structures Scientific data are very often structured differently than social science data. The numbers in the files, the structures they constitute, and the types of the files can all be very different than most social science data files. Most social science dat a files are delivered in ASCII form with a series of lines each constituting a record or part of a record and are made up of numbers which are measures of different

attributes. These numbers are usually integers. This is in contrast with the numbers in a scientific data file. Those numbers are described in terms of a much more complex classification scheme. That scheme has descended from computer programming since it has always been very common for scientists to write their own analytical software. While in social science data, numbers may have a dollar or time format, or a certain number of decimal points, in scientific data numbers are described as floating point numbers, short integers, long integers, IEEE f ormat floating point numbers, etc. This categorization reflects the common use of binary files to store scientific data, and the many different ways that computers can store numbers in binary form. Choosing the right number type is important to insure c orrect results, quick processing, and efficient storage of data. In the files in which scientific data is supplied, these individual numbers are also not always structured as a series of records made up of a number of measured variables each stored as one number. Instead, the numbers may describe a point, scalar, or v ector measurement and so an individual variable may include one or several numbers. In addition, lines of numbers in the file may not be series of records, but may instead represent measurements over a grid, and as a result, the number at each row and co lumn position represents a measurement at a different place in the grid. In effect, data for all of the lines in a grid constitute one record. In the terminology of scientists, most social science data would be described as one dimensional. Each measurement is usually made once for each respondent.

In the sciences, where measurements are repeatedly taken for some entity which has length and wi dth, as is usually the case for physical phenomena, data can become multidimensional. Three dimensional data is common for physical objects. Four dimensional data is not uncommon where measurements occur over time. Further, data of these varying forms are supplied to users in a number of file formats unique to scientific data. There are flat ASCII files available, but other file types are as or more common.

Crystallographers use data in the Cambridge Crystallographic File Format. Atmospheric scientists often use data in the National Center for Atmospheric Research's NetCDF format. And, NASA has its own file format, the Common Data Format, CDF. With these kind of specialized file types, specialized software packages are needed. SAS and SPSS do well enough with flat ASCII files, but they can't read many specialized scientific file types. Typical social science statistical packages also do not do well at representing the specialized structural properties of scientific data, e.g. 3 dimensional collections of vector values. The types of files created by these packages also do not often have sufficient mechanisms for storing the metadata which needs to accompany scientific files. CDF, NetCDF, and other file types can store information about the measurement units of the dat a, the names of the data's dimensions, the length of those dimensions, and calibration factors, among other forms of metadata.

Methodology/Subject Expertise

A social science data librarian can have skills, tools, and data resources that are used across a number of social science disciplines. In the sciences these vary from one discipline to the next: a knowledge of chemical data doesn't help much when deal in g with high energy physics data, or models of the magneto-sphere.

The social sciences seem more like one of the subdisciplines of the sciences than an equal to the sciences as a whole. Like chemistry and its subdivisions, there is a substantial body o f common methods. Scientific data is also not as accessible to the lay person as a social survey. It is easy to understand the questions asked in a survey and the range of responses. It is not as easy to understand scientific tests and their outcomes. Such tests are o ft en referred to by their technical name and assume a knowledge of the relevant field.

Graphics and Visualization Graphical representation or visualization of data is also different in the scientific world. Scientific data is much more frequently given graphic form as part of the research process than has been true for social science data. Models of fluids can on ly be appreciated via graphic representation. Molecules need to have their structure drawn to be easily differentiated. With most forms of scientific data, it is important to know how to give the data graphic form. As part of exploratory data analysis, and other analytical methods, social scientists do create graphic representations of their data. However, very often it is possible for analysts to use frequency tables, statistical calculations, and cross-tabulati on s which never produce graphical output. With many forms of scientific data, the object being studied can be unintelligible without graphic representation.

Images are very important to scientific research in additional ways. Much scientific data consists of images produced by cameras, or pseudo-images produced by other forms of sensing like x-rays. These images require specialized tools which can facilit at e their presentation and conversion between different file formats. These tools also must provide image processing capabilities for filtering, smoothing, and otherwise enhancing the image. Both photographic images and images produced by other means wi ll need to receive this treatment. Images are not something used very extensively in the social sciences. Graphical rendering of scientific information not only requires an appreciation of the techniques used in a particular discipline. It also requires attention to be paid to the factor of human perception. Different colors can affect the perception of t he size of objects, or the attention paid to parts of a graphic. The way that images are drawn or processed is a methodological issue for the sciences.

How Do You Cope With the Differences?

Motivated by user need, or intrigued by the challenge presented by these differences, a data librarian may choose to determine what is required to provide access to scientific data similar to that provided for social science data. The differences betwe en scientific data and social science data are significant but not insurmountable. The required subject expertise can be borrowed from a number of sources and encapsulated in data management systems. Scientific visualization software can also be obtaine d to handle data management and graphical rendering tasks.

Scientific Visualization Software

It will be necessary to invest in new software tools. Software must be purchased which can read and write the required file formats. The best software tools available for reading scientific data files are scientific visualization packages. Scientific v isualization tools also incorporate image processing and rendering routines. These allow the creation of graphical representation of data, the display of images, and analytical processing of both data and images. To cope with the problem of investing in new software tools, cooperative efforts should be made with the computing center, scientific departments, or other universities. Software can be cheaper if it has been site licensed, a department already has a se rver with the software on which data can be placed, or if the costs can be shared among a number of institutions. Consulting assistance, and graduate student workers will also more likely be available for software packages already in use locally. Subject Expertise Disciplinary experts need to be found to assess scientific data collections, and software tools, and to respond to public queries about the data collected. These may be library staff, like the chemistry librarian, or they can be staff of a science depa rt ment, such as the science department's computer guru. Existing data library staff could be used if they spent the time to obtain the necessary disciplinary expertise, but this would be a tremendous cost to be paid by the data library. Without addition al outside assistance, no extensive scientific data service program is possible. Once outside expertise is obtained, the role of the data librarian must be to serve as the expert on data management in its most general sense. Resolution to the issues already well understood by the data librarian, such as what file formats are used a nd what software is required need to be answered. The data librarian should supply the list of data management questions which need answers and ensure that the disciplinary expert provides the needed information. Because subject expertise oftens comes from outside the data library it may not be available on a regular basis. To compensate for this, it is necessary to encapsulate that expertise in the form of software. The software commands and procedures for fi le handling, visualization, and data extraction which are most commonly used by experts need to be incorporated into a menu-driven data management system. The resulting program would resemble recently developed programs for social science data extraction using a WWW interface. With these extraction programs, typical steps and statistical software package commands are automatically executed for the user o f a WWW browser. The user is only required to make a selection from a limited number of actions appropriate for the dataset, and who need not understand the syntax of that software package or the physical format of the data. Thus, automation makes it po ss ible to give inexperienced staff or librarians, including data librarians, an easy to use tool which doesn't require them to confront the complete set of program manuals, or the issue of the proper syntax, every time they want to access the dataset. Th es e tools can then also be made remotely available in the science library.

Conclusion

The social science data librarian has the knowledge necessary to understand in general terms what it takes to provide access to scientific data, if not the specific skills or tools needed to implement this access. We know the questions that need to be as ked and the kind of experts who can answer them. Issues of data structures, software tools, relative importance of datasets, and disciplinary practice are common to both areas. We can act as managers of data services for both forms of data. There are many potential beneficiaries of such an effort to create broader numeric data services. The library may gain additional support and respect from the scientific community. Data libraries, many of which are now under tremendous financial press ur es, may gain new sources of funding. The user of scientific data will gain in many of the same ways that social scientists have gained from centralized provision of data. Data will not always be locked up in one faculty members office, or in the hands o f another unapproachable department. Data collections that might not have been acquired because their user communities were divided can be purchased via cooperative cross department efforts. Producers of scientific data will have expanded facilities t hr ough which to share data with their colleagues and students.

*. Paper presented at IASSIST95 May 1995 Quebec City, Quebec, Canada.