Providing Access to a Data Library: SQL and Full-Text IR Methods of Automatically Generating Web Structure

Lynn Jacobsen, Senior Consultant
David Millman, Manager of Research & Development
Walter Bourne, Assistant Director
Academic Information Systems, Columbia University

Abstract

Social Science research in universities has traditionally been supported by a central *data archive,* holding both the raw data (e.g. the U.S. census) and the wide variety of support materials necessary to identify, understand and manipulate the raw data. Columbia University's Electronic Data Service (EDS) has built an online system using World-Wide-Web technology to offer these support materials, and ultimately the raw data itself, to our research community. We describe our historical motivations, our data format difficulties, our filing systems and our scalable technology solution. Our emphasis is on a new set of software, inter-connected by Web protocols, which is easy to use and is self-maintaining.

Keywords: Social Science; Statistical Analysis; Data Archive; Data Library; SAS; SPSS; World Wide Web; WWW; W3; SQL; Sybase; Full-text; IR; Pat; Web Gateway; CGI Script.

Introduction

Doing Social Science research is the same whether trying to predict political inclinations based on religious affiliations or assessing whether being an oldest child leads to greater success in life. A research project typically includes the following steps: defining the question, designing a methodology, selecting research tools, finding or collecting data, performing the analysis, interpreting results and, finally, writing up the findings.

If the researcher does not collect the data, he or she will need to identify an existing study relevant to the research question. At universities, data libraries have been created to meet these data needs. These archives have, or have access to, thousands of studies. But data libraries' traditional support methods have been labor intensive. With the recent, dramatic increase in the use of numerical data, one-on-one consultations are increasingly impractical. This paper describes a method of moving toward a situation in which the researcher can be more self-sufficient.

History & Background

Columbia University's data library is the Electronic Data Service (EDS). It is a joint operation of Academic Information Systems and the University Libraries and which officially opened its doors in September of 1992. EDS maintains a large library of data files and related documentation, and makes this material available to the research community.

The current inventory includes data files from over a thousand studies. Many of the studies are stored on computer tape, but Census and other data from governments and international agencies is available on CD-ROM as well. EDS is also a member of several consortia from which it can obtain studies. Because researchers may freely request any consortial study, EDS maintains information on many studies which are not in its own collection.

Researchers may come to EDS at any one of three stages in their work. First, they may need to identify an appropriate study through consultation with the staff or by searching through the catalog of study descriptions. After finding a promising study, the researcher needs to see other related documents to understand the study and the analytic possibilities in greater detail. Finally, the researcher will choose a study and begin working with the data itself.

For a given study, there are several major components a researcher needs to consult:

- Study Description - A brief description of the data in the study, the methodology, the format, related publications, etc.
- Codebook - A detailed description of the data, its variables and values, data problems, and the methods used.
- Sample SAS/SPSS Programs - The analysis tools of choice for most Social Science researchers are SAS and SPSS. For many studies, we have sample SAS or SPSS programs to help the researcher get started.
- Access Information - Including where and how the data is stored.
- Data - The study files themselves.

Up to now, one-on-one consultations have been absolutely necessary because of the complexity of gathering and understanding all of this information. Support materials (codebooks, study descriptions, etc.) and the data sets themselves appear in all possible forms, from paper to magnetic tape to online files. They have necessarily been organized in a series of different filing systems because of the variety of formats. The archive staff spend a lot of time simply steering clients through these resources. And the researcher often leaves with a combination of paper print-outs, ancient photocopies, pages of hand-written notes, lists of potential exceptions . . . and our phone number because they will certainly need to call us back for a random missing item or two.

This method clearly has a number of drawbacks, for both our staff and our researchers.

It is important to keep in mind that the researcher is most interested in her research project. Her primary goal is not to understand the intricate workings of either the Internet, or a computer operating system, or how to mount and read tapes, or the best way to access files in several different formats. Instead she wants to discover whether the data show any significant patterns with respect to her questions.

The goal of a data library, then, should be to make the process of selecting a study and using it as straightforward, as accessible, and as flexible as possible. So we have embarked on an ambitious redesign of our EDS services to simplify access to this very heterogeneous mix of data. We have focused on three key areas: putting these materials online; designing a consistent and scalable organization of these resources; and developing a few critical application programs.

The New System

Under the new EDS *DataGate* system, a researcher is presented initially with a Web document offering access to our study description library. They may either browse a subject category index or use a full-text search to obtain a relevant study description.

The study descriptions themselves are also Web documents. If the corresponding data set is held by EDS, links within the description give access to all of our support materials: codebooks, sample analysis programs, and location and data format information. If the study is not held by EDS but is available for ordering from one of our consortial partners, a link to an HTML order form appears instead.

Data sets held by EDS are not yet online and part of this Web tool. But we now have a structure through which to provide them.

Technical Details

In creating this system, we had to consolidate and simplify access to the variety of formats we had acquired or created over the years. For example, a commercial SQL relational data base (Sybase) contains bibliographic and access information; an unstructured text file contains descriptions of our, and our consortial partners', studies; and a structured text file assigns subject categories to studies.

Some practical principles guided our design. We needed to deliver these materials consistently and continuously; thus our choice of Web technology, for which we already had an infrastructure of well-organized and reliably supported servers, and a wide penetration of clients on the campus. We believed that the most natural access method for researchers would be through the study descriptions and, in particular, by doing full-text searches on them. And we wanted to make maintenance as simple as possible, by having a system which would both integrate transparently with our existing procedures and also automatically adjust its access paths into our continually growing online collection.

A certain proportion of our work has been relatively mundane: we have shuffled and reorganized our online files and moved a number of files from tape.

But most "documents" the system delivers are generated by automatic processes. A number of "filters" have been created to translate material from its original form into fully-linked HTML. These programs are executed on a regular basis ("cron jobs") to keep the information up-to-date.

We also pre-process and combine our full-text study description file, our SQL relational database of holdings, and the subject catalog file. The result is a marked-up

text file containing enriched study descriptions of our current and potential holdings. This file forms the foundation and the entry point of the DataGate.

Access to the study descriptions is through either a subject index or by full-text search. The subject index is periodically produced in HTML by a filter which combines the subject-category file with our holdings database. Full-text retrieval is performed by a commercial engine (Pat) running under a home-grown Pat-to-HTML gateway. We have chosen this route, rather than the more common WAIS method, because it offers much more flexibility: searches can be performed on specific fields and on "hidden" fields; and our gateway program manipulates the tags, thus offering plain-text Gopher access as well. In fact, because the Pat engine is designed to work with SGML, the more generic superset of HTML, our documents are somewhat more richly tagged than would have been otherwise possible.

For those studies held at EDS, the supporting documents are available via the study description through hypertext links. These links do not point to static files but rather to a program we call the *CGI Menu Builder*.

The *CGI Menu Builder* constructs an HTML document based on the information available in EDS at the moment: it dynamically generates links to the codebooks, sample programs, holdings and, eventually, to the data set itself. And the system is automatically self-documenting. For example, when a staff member creates a new sample SAS or SPSS program, they simply place it in the appropriate area and it immediately becomes part of the DataGate, linked-to from the appropriate study descriptions.

In order to provide up-to-the-minute information on our holdings, we use another CGI script to retrieve this information from our SQL database. This database contains not only the title, author and other bibliographic information but also the access information for the data set, such as the storage medium and format of the files. Again, the DataGate acquires new information from the SQL database automatically: our CGI script performs the SQL query on-the-fly, and so it always reflects the current state of the holdings database.

Conclusions, Plans

This assembly of filters, scripts, gateways and engines constitutes the entire EDS DataGate. After investing relatively little labor, we are pleased with its capabilities and flexibility. Our maintenance procedures are easier than they have ever been. We anticipate that our researchers will be much happier with its consistent interface and its availability around the clock and around the world. As we continue to add materials to the DataGate, researchers will be increasingly able to focus on their research, not on locating support materials.

Our ability to create this system so quickly depended on the underlying Web infrastructure. Using WWW capabilities, we have a delivery mechanism which is complex enough to incorporate our broad technological needs and yet which presents

a simple picture of our resources to our customers. And, within six months, we plan to add actual study data to the DataGate.

The widespread availability of Web tools will perhaps accomplish our next goal too. With several of our consortial partners and other data libraries, we hope to establish a comprehensive, distributed, digital library of Social Science materials for researchers across the country. We have little doubt that Web technology will be, in the near term, the reason we can become seamlessly available and yet also substantive partners in these national efforts.

About the Authors

Lynn Jacobsen, Senior Consultant at Columbia University's Academic Information Systems (AcIS). After spending six years working for an economic research firm as a Senior Economic Analyst, Lynn joined AcIS six years ago. She has taken her experience as a researcher and drawn on it in her current position supporting Social Scientists within the University. In addition to being the local SAS expert, Lynn was an integral part in the formation of the Electronic Data Service two years ago. She continues to be part of its growth and development today. Email: lynn@columbia.edu

David Millman, Manager of Research and Development, AcIS. David is the coordinator of Columbia's Digital Library Initiative, responsible for its technology development. He is the author of the *ColumbiaNet* campus-wide information system and continues to deploy and coordinate strategic technologies for the University, including WWW, security, and policy. Email: dsm@columbia.edu

Walter Bourne, Assistant Director, AcIS. Walter has supported Social Science computing at Columbia since 1976 -- through the Data Archive, Research, and Training Service (DARTS) until 1988, and then at AcIS where he participated in the reconstruction of University data support through EDS when DARTS lapsed in 1991. He was also an early promoter of the use of WWW at Columbia and participated in the design of ColumbiaNet -- Columbia's windowed, ASCII-terminal network navigator. Email: walter@columbia.edu

For further information, contact Lynn Jacobsen at lynn@columbia.edu.