Review of Web-Based Dissemination of the General Social Survey

Steven Ruggles
University of Minnesota

The General Social Survey has evolved a decentralized model of web dissemination strategies, with widely varying websites maintained at different locations. The following sections summarize and evaluate each website in turn. I then comment on the dissemination as a whole, and propose some recommendations for future development.

1. NORC Websites

http://www.norc.org/projects/General+Social+Survey.htm
The project website at NORC provides a summary of the GSS project, contact information, and links to GSS dissemination websites. No data or documentation is distributed from this website.

http://gss.norc.org/
This is described as "the main GSS website." It has no content except for a page of "Frequently Asked Questions" and links to other GSS websites.

2. Berkeley Website

http://sda.berkeley.edu/archive.htm
This is the data archive page for the Survey Documentation and Analysis (SDA) software developed and maintained by the Computer-assisted Survey Methods Program at the University of California, Berkeley. This is the top dissemination source recommended by the main GSS website, and it provides online tabulation and subsetting of the 1972-2004 cumulative file. The system is usable, but it has a lot of limitations. As the SDA documentation site acknowledges at the outset, "it does not provide full documentation of the dataset." The documentation component of the website is not integrated with the analysis and extraction facility. The documentation is sparse and difficult to navigate; there is no system for variable search and retrieval, and no facility for adding variables to a basket for downloading. There is no convenient way to determine the available years for any particular question; essentially, finding variables that have a long run of responses is a matter of trial and error. Apparently, not all questions that were asked are available from the SDA website. The Stata command files provided with extracts require extensive massaging before they will run. There is no explanation why the data stop in 2004, since 2006 data should be available.

3. GSSDIRS Website

http://www.icpsr.umich.edu/GSS/
This is the General Social Survey Data and Information Retrieval System (GSSDIRS) website hosted by ISPSR. The design of the website is extremely dated; the front page prominently warns that it is best viewed with Netscape 4.0 or higher, a browser released over a decade ago. Nevertheless, the website is surprisingly functional. Identifying variables and their availability is far easier than on the SDA website. Users pick variables as they browse the documentation, and then can extract their selections. The documentation clearly indicates chronological availability within each variable description, although there is no easy way to compare availability for broad groups of variables. Although there is no variable search capability, there is a useful subject

classification and it is possible to locate variables without too much difficulty. GSSDIRS does not do on-line data analysis; when one clicks on the analyze button, one is sent to the SDA website. The biggest problem with GSSDIRS is that the data are almost a decade out of date. Although the front page indicates that the system covers the 1972-2000 period, the data do not actually go past 1998.

4. Regular ICPSR Website

http://www.icpsr.umich.edu/cocoon/ICPSR/SERIES/00028.xml

This is the ICPSR web page where you can actually retrieve the full codebook and data for the 1972-2004 cumulative file. The codebook is a 70MB PDF file with 2,390 pages. Access to the codebook or data file requires ICPSR membership and login. Obviously, this format is unwieldy and would pose a serious obstacle for many potential users of the data.

ICPSR also runs its own version of the SDA analysis and subsetting program from this site. It is branded with the ICPSR logo, runs on ICPSR servers, and has a slightly different interface than the SDA website at Berkeley. The ICPSR SDA version of the documentation has no subject classification whatsoever; thus, unless the user already knows the mnemonics of all the variables they are interested in, I believe that the ICPSR SDA version of the documentation would be effectively unusable. As in the case of the Berkeley site, access to documentation is completely separated from access to data. The ICPSR SDA analysis system could be used in conjunction with another source of documentation, such as the massive PDF codebook, but I suspect most users would prefer to simply go to the Berkeley website.

Unlike Berkeley SDA and GSSDIRS, the ICPSR SDA documentation does at least provide basic information about the samples—information that all users should be aware of—in a prominent location on the main documentation page. On the other websites, it is necessary to dig deeply for this information.

5. Roper Center

http://www.ropercenter.uconn.edu/data_access/data/datasets/general_social_survey.html
The only way to obtain the 2006 GSS at this writing is through the Roper Center. Persons at member institutions may obtain a CD by emailing the center; non-members must pay $400. Since the University of Minnesota is not a member, I did not attempt to obtain a CD. Apparently, Roper has some sort of agreement with NORC that gives them exclusive rights to disseminate the data for a certain period of time. This practice is inappropriate because it impedes data access, and should be discontinued.

Overview

The decentralized GSS dissemination system is highly inefficient and provides suboptimal dissemination. The greatest problem is that data and metadata are replicated in several locations in widely varying formats. This is a maintenance nightmare; if an error is uncovered, it must be fixed in many different files by staff members at multiple institutions. Needless to say, it is unlikely that most errors are ever corrected. In addition, with four different dissemination pathways doing overlapping work, there is considerable duplication of effort.

Not only does this approach waste scarce resources for social science infrastructure, it also results in inferior service. Of the websites I evaluated only GSSDIRS provided an adequately user friendly—if old-fashioned—interface. Unfortunately, the data and documentation at GSSDIRS is eight years out of date, making it unusable for a majority of research projects. The only way to obtain up-to-date data is through Roper, and except for the small minority at institutions which subscribe to Roper, the cost is high.

Recommendations

GSS is expensive, and the costs can only be justified if it is widely used. Accordingly, NSF should explicitly fund web-based dissemination of GSS. I have three specific recommendations for the call for proposals.

1. GSS dissemination efforts should be centralized. This would mean that only one version of the data and metadata would have to be maintained, greatly reducing the costs of corrections. It would avoid costly duplication of effort for software development and maintenance.

2. A new integrated web dissemination system for data and documentation should be developed using modern software tools driven by standardized XML metadata. Because GSS is a simple rectangular survey, such a system would be comparatively inexpensive. The chief data access challenge posed by GSS is the sheer number of survey questions, many of which appear only once or twice. Tools to cut through the clutter and select variables of interest should be the highest priority of the data access system. The data access system should allow easy browsing of variable availability and incorporate a sophisticated system for variable search and retrieval.

The current GSS data access systems present all available variables as a simple pick list, but this approach does not work well for a dataset that has so many variables. Users should be able to narrow the variable list according to keyword or subject area; reduce the list to only those variables appearing in every survey year of interest or to expand it to include all variables in any selected survey year; and view simplified pick lists focusing on the most commonly requested variables, as determined through analysis of extract logs. At any point, users should be able to select variables and add them to a data basket. When they are ready, they should then be able to view the basket and either extract data for download or carry out on-line analysis.

3. GSS dissemination should be separated from survey design and administration. There are two reasons for this. First, when unexpected expenses arise or budgets are reduced, survey developers are notorious for cannibalizing their dissemination budgets. From the perspective of the agency, however, such rebudgeting is highly inefficient. Second, those with the greatest expertise in survey development are not the same as those with greatest expertise in social science dissemination cyberinfrastructure. Therefore, it would make the most sense to have a separate call for proposals for the dissemination component of the project.

Other Ideas for GSS Cyberinfrastructure

In addition to web-based data access tools, GSS could benefit from web-based training and user support. On-line tutorials could cover both basic issues, such as how to get GSS data into a statistical software package, and more complex analytic issues, such as variance estimation.

GSS dissemination could also benefit from the tools and technologies of the Social Web, known also as Web 2.0, which stress collaboration and sharing among users of web-based services. Such tools are built on the observation that the collective knowledge of users in a community is substantial, and if leveraged properly can benefit the entire community. GSS has thousands of experienced users. Tools and systems that allow users to support one other would mean that less individualized user support would be needed. By promoting interaction among users researching similar topics, research communities can provide intellectual support as well as purely technical assistance.

Here are some ideas for tools that would exploit the social web concept:

- **Wiki-enabled documentation** that would allow users to suggest corrections and improvements to the extensive documentation of the datasets. The user community contains many experts with deep knowledge of specific subject areas, and many are quite willing to share their expertise to help others.

- **Expert Q&A system** where users can pose specific queries. Volunteer experts can answer these questions by starting discussion threads; other users can comment on or clarify an answer, which generates better quality answers. These threads can then be archived and indexed by keywords, allowing users to search old queries before submitting a new one.

- **Specialized research forums** which can bring together smaller groups of users with detailed knowledge on a problem to share their latest developments. These forums would encourage research collaborations among scholars from diverse disciplines who otherwise might not interact.

- **Tools for sharing SAS, Stata, and SPSS code for data manipulation** developed by individual users that could also benefit others. Currently sharing among users is *ad hoc*, with no systematic match-making. A shared repository with a searchable directory would maximize the efficiency of researchers.

- **Tools for sharing curricular materials** based on the same principles as code sharing. The software developed for code sharing can be substantially reused for this purpose.

- **Expert recommendation system** for problems frequently encountered by users. The idea of this tool is to infer interests and requirements of users from their data requests and other activities, and then to recommend datasets, research forums, discussion threads from the expert Q&A, and code based on a 'match-making' algorithm. This approach has been very successful in many domains and has been shown to improve user experience and effectiveness.