Democratizing Access to Data: The American Religion Data Archive

by Roger Finke, Jennifer McKinney and Matt Bahr *

Democratizing Access to Data: The American Religion Data Archive (www.TheARDA.com)
From its beginning, the American Religion Data Archive (ARDA) was developed to provide immediate access to the best data on American religion at no charge. Starting in 1997, the ARDA was created as an Internet-based archive and was designed to serve a highly diverse audience. But serving a diverse audience, including many with little or no background in the social sciences, required ARDA to meet the rigorous methodological standards of the social science community and still be easily used by those without a knowledge of statistics, research design or data management. Since its inception, the ARDA has attempted to democratize access to data, without compromising the integrity of the data being archived.

This essay will review our efforts. We begin by giving a brief overview of the data we archive and the audience we serve. Next, we will review the goals of the ARDA and how we attempt to achieve each goal. Although the goals are similar to many other archives, we will highlight how we have developed features that allow us to achieve these goals in new and creative ways.

Religion Data Sources and Users
When ARDA was initially conceived, the 1995-96 *ICPSR Guide to Resources and Services* reported on more than 40,000 data files from over 3,000 social research studies. Even a topic such as education, which had comparatively few entries, reported 119 data files from 65 studies, with 34 of these studies being conducted since 1980. By comparison, the subheading of religion reported only 9 data files from 9 studies, with only two of the studies being conducted after 1980. Yet, this paucity of archived data on religion does not mean that data are not being collected. Over the last 10 years alone, Lilly Endowment has funded over 150 grants with a data collection component, the Pew Charitable Trusts has funded several major national and international surveys, and many denominations support research divisions that collect large amounts of data each year. Unlike, education, health care and other substantive areas, where most studies are funded by government sources, nearly all of the data collections on religion are funded by private endowments or religious organizations. Most funding sources have either wanted the data to remain "in house" or have not required principal investigators to place the data files in a public archive.

In the mid-1990s, however, the Lilly Endowment began a major initiative for improving dissemination. One component of this initiative was the American Religion Data Archive. After awarding a planning grant to Roger Finke in 1996 to study the feasibility of starting a religion archive, Lilly Endowment funded the start up and operation of the ARDA from 1997-2000. Recently, they extended the support until 2003. Thus, the funding sources for the collection and archiving of data on American religion remain private sources.

The ARDA currently holds 120 data files and the number should approach 150 by the close of 1999. These studies include national samples of the United States and Canada, regional samples of selected communities or areas, and samples of selected religious groups or professionals. Although all surveys include the topic of religion, the survey items span a wide range of other topics (e.g., from involvement in small groups and politics to attitudes on race relations and professional development). In addition to the surveys, the ARDA also distributes data on American religion by ecological units, such as the Association of Statisticians of American Religious Bodies' data on churches and church membership by counties and states for 1980 and 1990. Both the size and the diversity of the collection will continue to grow.

Once established, the greatest challenge for the ARDA was appealing to the diverse audience interested in American religion. Initially, we were most aware of the social scientists from research universities who frequently conduct and report on the major data collections. For this group, the ARDA was a valued repository of past data collections and a source of new data for future research studies. But this audience, often sophisticated in research methods and statistics, represents only a small portion of the total audience. Many, and probably most, of our users have little background in the social sciences and are not located at research universities. Instead, many are faculty members and students located at small universities, colleges, and seminaries that previously had little access to data on American religion.   Based on our web site reports, seminaries have made more contacts and referrals to the ARDA site than any other type of educational institution. And, though we have no record of individual users, our most frequent e-mail and telephone inquiries come from journalists and students. Several instructors have informed us that they have incorporated ARDA data files and software into class assignments. Rather than limiting access to a small group of researchers, ARDA has democratized access to the data, and a very disparate audience is taking advantage of this access.   Below we review how we appeal to this disparate audience as we strive to achieve standard archiving goals.

Goals of the American Religion Data Archive

The goals of the ARDA are similar to those of many other archives. ARDA was established to:

1. Preserve Data
2. Improve Access to Data
3. Increase the Use of Data
4. Allow Comparison Across Data Files

To achieve these goals we combine proven archiving practices with new attempts to serve a diverse audience.

The first goal, ***preserving data,*** is the foundation of virtually all archives, and in the case of data on American religion, it was the most essential. Of the first 150 data files we received for the ARDA only three were previously held in a public archive. Preparing the data for the archive follows many of the same procedures developed by other scholarly archives. After we receive the data files, we verify the accuracy of the data by comparing our variable frequencies with those of the principal investigator and we begin collecting summary information, or metadata. For each of the files we offer a brief abstract of the study and we provide information on the number of cases, number of variables, the year it was conducted, sampling techniques, sources of funding, principal investigators, collection procedures, any related publications and additional information on the construction of indices or the use of weight variables when appropriate.

In our effort to "democratize" access to the data, however, we have gone beyond the standard procedures used to prepare data files for scholarly research. We have added a couple features that make the data files more accessible and easier to use. First, we recreate the original survey instrument within the data set.   Using the original questionnaire, we record the complete variable description and all response categories. Users are not forced to keep a codebook by their side to decipher variable names or truncated descriptions. Moreover, when the files are downloaded as MicroCase files the entire survey wording remains.[1] Second, we have designed the web site so users are forced to review the metadata before they download files, and they can easily link to the metadata whenever they are reviewing questions or data from the file. This is handy for experienced researchers and essential for those with less experience.

Improving access to data, the second ARDA goal, was primarily achieved by adding an easy download feature to the site. Thanks to the support of the Lilly Endowment, anyone with access to the Internet can download the data free of charge. Once users find a data file they want to use, they can easily download it to their own PCs as an SPSS, ASCII or MicroCase file. They also have the option of downloading a codebook without the data.

Once again we have added a feature that allows the data to be used by non-specialists. MicroCase Corporation's statistical software, ExplorIt, can be downloaded free of charge and is fully compatible with the MicroCase data files available from our site. The ExplorIt software is used by thousands of social science students each year throughout the United States and Canada and is remarkably easy to use. The ExplorIt version offered from the ARDA site holds fewer statistical options than the version typically distributed for classroom use, but it offers an important option for non-specialists who do not have a statistical package readily available.[2] Many professors have found this to be an especially attractive option for their students.

For the third goal, *increasing the use of the data*, we wanted to allow users to conduct basic analyses of the data files on-line. Yet, from our own classroom experiences with undergraduates, we knew how confusing bivariate cross-tabular analysis can be for those not familiar with statistics. First, constructing the table requires students (or any user) to fill in boxes that ask for an independent and dependent variable — unfamiliar and unfriendly words for most. Second, they must select variables with an appropriate number of categories. For example, when a student tries to set up a table with age by income, the resulting table might offer an incomprehensible 70 columns and 20 rows. And, even if they are successful in constructing an appropriate table, they need to know which way to percentage the table.   Choosing to percentage in the wrong direction leads to meaningless or often misleading results.

We have avoided this quagmire by working with MicroCase Corporation to develop a simplified version of their *auto-analyzer* for our web site. When users find a question of interest, they can click on a button called "Analyze" and tables are constructed using preset variables. The tables are percentaged correctly and typically include standard demographic variables such as age, gender, income, marital status and education. This avoids the potential problems of choosing an independent and dependent variable or deciding which way to percentage.[3] For example, if a question is selected that asks "which party's candidate would you be most likely to support if a federal election were held tomorrow?," the user would first see a table summarizing the number and the percentage of respondents who would vote for each candidate. Then a series of tables would follow, showing how these percentages and numbers vary by age, gender, income and so forth. The user has received a series of meaningful tables on the question of interest, without struggling through a series of commands.

The fourth goal, *allowing comparisons across data files and over time*, is achieved through standard searches. The user can search for a topic of interest within a single data file, a selected group of data files, or all ARDA data files. After locating questions of interest, the user can quickly compare the results for each question by conducting on-line analysis or they can compare the data files from which the questions were selected. Thus, users can quickly compare similar questions to see if they offer equivalent results, and they can review information about the data files to better understand why the results might differ (e.g., the samples might vary by location, time or religion).

Once users receive questions from their searches, they can also place the questions in their own question bank. In other words, they can start saving questions for their own survey. During the planning phase of the ARDA, we were encouraged by prospective users to establish an archive of questions as well as an archive of data. Because the complete survey questions are entered and stored in the data file, however, the data file represents a complete record of the survey instrument. Hence, when data collections are submitted, ARDA serves as an archive for the questions used and the data received. By combining the question bank feature with the search feature, the ARDA becomes a rich resource for constructing a new survey as well as using a previous one.

Summary

We recognize, of course, that ARDA's initial efforts to democratize access to data are simply that: initial efforts. Still, we are encouraged. The support of the Lilly Endowment has made the archive possible and has eliminated the barrier of financial cost for using the data. The availability of downloading MicroCase's ExplorIt software and using their on-line analysis tool has greatly reduced the barrier of data analysis for a larger audience. And, providing data files that offer complete question wording, detailed metadata, verified data, and muliple download formats, renders a rich resource to the experienced and inexperienced user alike. Reducing each of these barriers, and extending the services offered, has helped to increase the use of the data and expand the diversity of the audience using the ARDA.

Finally, we want to close with a gentle reminder to ourselves and others. Democratizing access to data and metadata are noble goals made possible by recent advances in technology. Yet, we should remember that metadata are often an empty promise unless the data are available; and, easily accessible data can still be useless (and misleading) unless they are carefully conducted and prepared data collections. A data archive will still be judged by the quality of the data it provides. Hence, just as evangelists close each revival with an invitation to submit to the message just heard, we end each essay and presentation with an invitation for submitting data. If you have data on American religion to submit, or you know of data that should be submitted, contact the ARDA (archive@sri.soc.purdue.edu) or download a submission form from our web site (www.TheARDA.com).

[1] Due to the character limitations of SPSS for variable descriptions, some of the questions will be truncated when SPSS portable files are downloaded.

[2] The simplified version of the ExplorIt software, downloaded from the ARDA web site, provides univariate statistics with the appropriate bar graphs and pie charts, crosstabs with the appropriate statistics, and a complete list of survey questions that can searched for a topic of interest.

[3] If the variable has too many categories for constructing a table, the user receives a message with this information.

* Paper presented at the IASSIST Conference, May 17, 1999, Ryerson Polytechnic University, Toronto, Ontario. Roger Finke, Jennifer McKinney and Matt Bahr, The American Religion Data Archive, Department of Sociology, 1365 Stone Hall, Purdue University, West Lafayette, IN 47907-1365, 765-494-0081, archive@sri.soc.purdue.edu, www.TheARDA.com