Best Practices in the Dissemination of Survey Data

Peter Granda
Inter-university Consortium
for Political and Social Research (ICPSR)
University of Michigan

Preservation

Preservation is an important part of the data life cycle, allowing for long-term access to valuable digital materials. Digital materials are best protected by having multiple copies stored at off-site locations. An ideal preservation storage situation includes a minimum of six off-site copies of digital materials undergoing regularly scheduled back-ups. In addition to this redundancy, the media on which the digital materials are stored require ongoing refreshment. Data and documentation formats should also be software independent. An organization with an effective preservation strategy makes an explicit commitment to preserving digital information by:

- Complying with the Open Archival Information System (OAIS) and other digital preservation standards and practice
- Ensuring that digital content can be provided to users and exchanged with archives so that it remains readable, meaningful, and understandable
- Participating in the development and promulgation of digital preservation community standards, practice, and research-based solutions
- Developing a scalable, reliable, sustainable, and auditable digital preservation repository
- Managing the hardware, software, and storage media components of the digital preservation function in accordance with environmental standards, quality control specifications, and security requirements

Disclosure Analysis

Any plan to disseminate survey data must include very specific procedures for understanding and minimizing the risk of breaching the promise of confidentiality that is made to respondents at the time of the survey. Appropriate disclosure risk analysis involves both practical and statistical steps that attempt to identify cases and variables that might be recognizable to an intruder, or matched with external databases. Once those cases and variables are identified, the survey can be evaluated. In virtually every case, the data can be masked in various ways that make it possible for public use data to be distributed, usually through a Web-based system. Sometimes these masking procedures reduce the usefulness of the data for analysis, in which case it is appropriate to create less-thoroughly masked versions that can be distributed under restricted use contracts or made available within a research data center or “enclave.” The key goal of disclosure risk analysis and processing is to ensure that the data have the greatest potential usefulness while simultaneously offering the strongest possible protection to the confidentiality of the individual respondents.

Documentation Processing

High quality metadata is essential to effective data use, and adopting the Data Documentation Initiative (DDI) XML standard for metadata offers several advantages. First, all information that the analyst needs is available in a core document, from which other products, such as setup files, can be produced. Second, the XML file can be viewed with Web browsers and lends itself to Web display and navigation. Third, because the content of each field of the documentation is tagged, the documentation can serve as the foundation for extract and analysis programs, search engines, and other intelligent agents. Finally, preparing documentation in DDI format at the outset of a project means that the documentation will also be suitable for archival deposit and preservation. DDI XML should ideally be generated by the CAI system used to collect data.

With rich DDI markup, instrument documentation can be presented so that users can track the logic of the questionnaire. Also enabled is a bank of all questions ever asked in multi-year studies, years they were asked, differences in question wording, etc. This approach permits linking to the documentation of related surveys -- for example, those conducted in other countries -- with variable text viewable in the native languages, so that analysts can study relationships among all of the survey items.

Data Processing

An effective data processing strategy focuses on the production of data files which will provide optimal utility for researchers. Processors must perform a series of steps to ensure the integrity of public-use files. Such steps include: a thorough investigation of any wildcode or inconsistent responses, the standardization of all missing data values, reformatting any variables to maximize storage capacity, and the creation of complete and concise variable and value labels which will provide researchers with clear descriptions of their analytic results. The format of the data files should permit access through a wide variety of statistical packages all of which will produce the same results no matter how complicated the analysis requested.

New Data Products

Data producers and archives should consider producing ancillary files for those data collection efforts which cover multiple waves of respondents or several geographic areas. Special subsets of data which take advantage of the longitudinal richness of long-term collections provide unique opportunities to study important social, political, and economic issues from different perspectives particularly with regard to the changing characteristics of the sampled respondents.

Finding Aids

Finding aids are critical to a Web-based system. A robust search engine is needed to query the fielded metadata so that the user can find variables of interest efficiently. The search should also run against a study’s bibliography so that there is two-way linking enabled between variables and publications based on analyses of those variables. Displaying full text of the publications whenever possible is also essential to realize the full potential of the online research environment. Dedicated staff should be continuously searching journals and online databases to discover new citations.

Types of Dissemination

The data producer must make every effort to make all public and restricted data and documentation files available to the research community through secure and predictable channels. The producer may decide to provide their own access but should also send copies to a trusted digital repository for permanent preservation should the producer decide to cease such services in the future.

Providing optimal utility for researchers means that data producers and archives produce a variety of products for their varied constituencies. To address the needs of those who seek to do intensive statistical analyses with particular software packages, processors should produce setup files and ready-to-use 'portable' files in SAS, SPSS, and Stata. To address the needs of policymakers and those who are browsing for new data sources, seeking summary analytic information, or may want to download specific variables quickly, producers and archives can create tools within the Web-based system to permit online analysis, subsetting, and access to full documentation.

Training and Outreach

It is very important that major survey research products reach out to the user community effectively in order to ensure that they receive the greatest possible use. The most straightforward way to reach out is to develop an effective on-line presence and to ensure that the data are easily located and acquired, and that metadata and bibliographical citations are also available. Beyond that, effective outreach usually includes three activities. First, data producers often organize workshops or conferences soon after the data are released to bring early users together to discuss important preliminary results and to ensure both that the data are used effectively and that any problems with the data are recognized and corrected. Second, data producers often hold training workshops to ensure that novice users have a chance to learn about the data from experts and especially from the data production team itself. Longitudinal data and repeated cross-sectional data are particularly challenging to analyze without specialized instruction and training. These training courses can be brief half-day or one-day sessions at the time of professional meetings, or they can be three- or five-day sessions in the summer (or during the academic year) with a more detailed focus. Finally, data producers send representatives to important professional meetings with a display "booth" where staff from the project can describe the data, distribute documentation and sample data, and encourage researchers to make use of the data.

User Support

Rounding out such a Web-based system is easy access to user support through phone, email, online chat, user forums, and tutorials. All user questions should go into a database that tracks them and creates an accumulating knowledge base, which can also serve to generate Frequently

Asked Questions. Tutorials, some of which may be offered in video format, can be used to provide help in using the data, the online analysis system, and the major statistical software packages. The user forums provide the foundation for an online community of researchers and students who can discuss their experiences using data and learn from each other.