

## XIV. Dissemination of Survey and Statistical Data

### Introduction

Dissemination of [survey](#) and [statistical data](#) requires careful consideration of several aspects of the process of making data and documentation files available to secondary analysts. More is involved in the dissemination process than merely sending files stored on removable media to interested researchers, or putting files up on a server for others to download. Data producers and archives must assure analysts that the data they provide accurately reflects the efforts of the data collection process and is trustworthy, fully documented, and securely preserved for future use. Many international organizations also embrace these objectives. Although focused on micro-economic data, The International Monetary Fund, for example, established a set of guidelines on macro-economic (data for member countries to follow in order to provide the public with “comprehensive, timely, accessible, and reliable economic, financial, and socio-demographic data” [\[3\]](#)).

### Guidelines

**Goal:** To ensure that survey and statistical research teams in all countries involved in a project follow accepted standards for the preservation and dissemination of data to members of the social science research community.

- 1. Preserve copies of all key data and documentation files produced at the end of the data collection process, as well as those made available for secondary analyses.**

#### *Rationale*

Preservation is an important part of the [data life cycle](#), a prerequisite for long-term access to valuable physical objects and digital materials. The materials that need to be preserved and kept available to members of the research community include such objects as public-use data and documentation files (including key files used in their construction), copies of the data collection instruments, user guides, information about the data collection process, and reports on field operations. Since dissemination policies may differ among countries, it is important that data producers take the necessary steps to make their collections as accessible as possible to members of the research community. This may include organizing dissemination themselves.

### ***Procedural steps***

- Consider converting into electronic format physical objects, commonly-used questionnaires, or other important administrative materials documenting the data collection and processing procedures.
- Protect digital materials through storage of multiple copies in multiple locations. An ideal preservation storage situation includes a minimum of several off-site copies of digital materials undergoing regularly scheduled back-ups. If it is not possible to store materials at multiple sites, preserve at least one copy in a different location.
- Make certain that digital materials remain retrievable through constant refreshment of the media on which they are stored. This is particularly important if removable media such as tapes are used for storage, since formats and the machines required to read these media change over time.
- Maintain older versions of important data and documentation files so users can follow the changes made from one version to the next.
- At a minimum, store a copy of all data and documentation files in software-independent formats such as [ASCII files](#) which, with proper accompanying documentation, can be read into all major statistical packages.
- Work if possible with a [trusted digital repository](#), such as a national or public social science data archive, to preserve all study materials. In doing so, data producers do their best to ensure that their data collections will remain available to the research community.
  - Such repositories make an explicit commitment to preserving digital information by:
    - Complying with the Open Archival Information System (OAIS) in the US and other similar standards in other countries which have their own digital preservation standards and practices [\[5\]](#) [\[7\]](#) [\[9\]](#).
    - 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 reliable, sustainable, and auditable digital preservation repository that has the flexibility to grow and expand

- Managing the hardware, software, and storage media components of the digital preservation function in accordance with environmental standards, quality control specifications, and security requirements
- Deposit collections with an archive in another country or investigate the possibility of doing so with a national statistical agency if no national or public social science data archives exist. Consider archiving collections in one repository to minimize the possibility of conflicting versions of data and documentation files.

### ***Lessons learned***

- Data producers should strongly consider a preservation strategy before putting files online for people to download. For example, many data and documentation files available on Web sites undergo frequent changes and updates. When updates are made, the older version of the files is often no longer available. This may make it difficult, if not impossible, to replicate previous analyses done by the user, or to test the assumptions and results of analyses done by others.

## **2. Conduct effective [disclosure analysis](#) to protect respondent [confidentiality](#).**

### ***Rationale***

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 or collection of data. The key goal of disclosure risk analysis and processing is to ensure that the data maintain the greatest potential usefulness while simultaneously offering the strongest possible protection to the confidentiality of the individual respondents.

### ***Procedural steps***

- Undertake both practical and statistical attempts to identify cases and variables. This allows the identification of areas or variables that need to be further masked in order to prevent identification of subjects, either through analysis or by matching study data with data from other external databases.
- Evaluate data files once those cases and variables are identified. 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.

- Use appropriate masking procedures to preserve respondent confidentiality while also trying to optimize the usefulness of the resultant data file for analysis. These procedures might include such steps as [top](#) or [bottom coding](#) of key demographic variables such as income, removing data for very sensitive variables, and swapping data values between similar cases [\[6\]](#).

### ***Lessons learned***

- With the enhanced emphasis on privacy in almost all countries, [confidentiality](#) reviews of [microdata](#) are increasingly important, if not indispensable, to assuring the future availability of public-use data.
  - The practice of reporting examples of privacy violations, particularly in the health care field in the United States, has increased awareness of this issue [\[2\]](#).
3. Consider the production of both [public-](#) and [restricted-use data files](#).

### ***Rationale***

In order to ensure that researchers have access to the greatest amount of data without compromising respondent [confidentiality](#), data producers, when appropriate, must make every effort to create both public- and restricted-data documentation files, and make these files available to the research community through secure and predictable channels.

### ***Procedural steps***

- Make data files as fully available to the research community as possible within the confines of how the project is organized and financed. If general distribution is not feasible, establish clear rules under which researchers can obtain the data.
- Remain cognizant of the fact that data files, however they are disseminated, are always 'owned' by the principal investigator(s) who maintain permanent copyright privileges over their products.
- Provide access directly by the data producer if resources permit, but also always send copies to a [trusted digital repository](#) for permanent preservation, in case the data producer should cease to provide access at some time in the future.
- Consider the creation of less-thoroughly masked versions that can be distributed under restricted-use contracts, or made available within a

research data center or “enclave,” i.e., a secure environment in which the user has access to restricted data and analytic outputs under controlled conditions.

- Establish clear policies for how researchers may access [restricted data files](#) by creating a set of application materials and restricted-use data agreements that specify how researchers can obtain and use such data [4].
- Distribute restricted files through signed data use agreements. These may incorporate data protection plans, formal licenses, and travel to a special facility at which researchers can access the data in a very controlled environment.
- In order to provide optimal utility for researchers, produce a variety of products for varied constituencies.
  - Produce setup files and ready-to-use [‘portable’ files](#) in SAS, SPSS, and Stata to address the needs of those who seek to do intensive statistical analyses with particular software packages.
  - Consider disseminating data on removable media, e.g., CD\_ROM or DVD if appropriate.
- Address the needs of policymakers and those who are browsing for new data sources, seeking summary analytic information, or wanting to quickly download specific variables by creating tools within the Web-based system to permit online analysis, subsetting, and access to full documentation.

### ***Lessons learned***

- Despite general agreement about the advantages of making data accessible to other researchers, as well as strong data-sharing cultures in many nations, too few social science data collections are effectively preserved. Data archives should do as much as possible to facilitate the deposit process by contacting principal investigators and data producers as they prepare data and documentation files.
- More than ten years ago the International Monetary Fund (IMF) began to develop a set of dissemination standards “to guide countries in the provision to the public of comprehensive, timely, accessible, and reliable economic, financial, and socio-demographic data” [3]. These standards were considered best practices but their implementation was completely voluntary depending on the policies and wishes of each nation. The Fund recently published a report about the success of this initiative over the last decade. It concluded that more accurate and reliable statistical information is now being produced by many nations

than ever before but also recognized that dissemination mechanisms are not fully developed in many locations. Nations also have internal challenges and constraints in addressing dissemination goals from resource constraints, shifting priorities, and in their ability to generate periodic and timely statistical data.

#### **4. Produce data files that are easy for researchers to use.**

##### ***Rationale***

An effective data processing strategy focuses on the production of data files that will provide optimal utility for researchers. Such files have been thoroughly checked and cleaned, possess uniform and consistent coding strategies, and address the potential research needs of secondary analysts.

##### ***Procedural steps***

Processors should perform a series of steps to ensure the integrity and maximum utility of public-use files. Such steps include:

- Make a thorough investigation of any [undocumented codes](#) or [inconsistent responses](#).
- Standardize all [missing data](#) values, unless it is not possible to do so because of different cultural understandings.
- Reformat any variables to maximize storage capacity.
- Create complete and concise variable and value labels which will provide researchers with clear descriptions of their analytic results.
- Format the data files in a way that permits access through a wide variety of statistical packages, all of which will produce the same results no matter how complicated the analysis requested, particularly with any variable where decimal precision is an important consideration.
- Consider producing ancillary files for those data collection efforts which cover multiple waves of respondents or several geographic areas. Such files may include recoded variables to summarize information contained in many questions or special [constructed variables](#) that producers feel will aid researchers in their analyses.
- Create special subsets of data which take advantage of the longitudinal richness of long-term collections and 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.

- Whenever possible and expedient, make individual country datasets available in cross-national surveys.

### ***Lessons learned***

- Users increasingly expect data files to come in a variety of formats that will work easily with their statistical package of choice. In some settings this may be just an SPSS portable file, but in others data producers and/or archives might need to create the same file in a variety of formats, particularly if a standard database conversion package, such as STAT-TRANSFER, is not available.
- “Don’t know” responses may have different meaning in different countries based on different response styles. Treating all of these responses as missing data may lead to unwarranted conclusions about the attitudes of whole populations [8].

## **5. Develop finding aids to guide users in their quest to locate data collections they want to use.**

### ***Rationale***

Finding aids are critical to all data dissemination systems, from individual data producers, with only a few data collections, to social science archives with thousands of such collections.

### ***Procedural steps***

- Create a robust search engine to query the fielded [metadata](#) so that the user can find variables of interest efficiently.
- Allow the search engine to run against a study’s bibliography to enable two-way linking between variables and publications based on analyses of those variables.
- Display the abstracts of the publications with links to the full text whenever possible, in order to realize the full potential of the online research environment.
- Dedicate staff time to continuously search journals and online databases to discover new citations where the data have been used.

- Encourage data archives to create metadata records for surveys they do not preserve and distribute these records to facilitate their discovery and use.

### ***Lessons learned***

- Use of data increases when the data are easy to find, and when users know which publications previous scholars have generated from such data. There are many datasets that would be of interest to secondary analysts if the analysts only knew about them. For example, many surveys were conducted in Latin America and Africa in the 1960s and 1970s which might offer opportunities for interesting comparative analyses with the more recent and much more popular Latino and Afrobarometer surveys. These are not always as visible to researchers, however, as they might not possess immediately obvious substantive or methodological interest.

## **6. Create comprehensive training, outreach, and user support programs to inform the research community about the dataset.**

### ***Rationale***

Training and support of users will increase their usage and enhance results. It is very important that major survey research producers or archives reach out to the user community effectively, in order to explain the structure of new datasets and to encourage the greatest possible use. The most straightforward way to reach out is to develop an effective on-line presence, ensuring that the data are easily located and acquired, and that [metadata](#) and bibliographical citations are also available. Good user support will prevent obvious misuse or possible misunderstanding of the structure and content of the dataset.

### ***Procedural steps***

- Organize workshops at relevant professional organizations or plan conferences soon after the data are released, in order to bring early users together to discuss important preliminary results, as well as ensure that the data are used effectively and that any problems with the data are recognized and corrected.
- Hold training workshops to ensure that novice users have a chance to learn about the data from experts and, if possible, from the data production team itself.
  - Without specialized instruction and training, analyses of cross-cultural, longitudinal data and repeated cross-sectional data are particularly challenging.

- These training courses can be brief half-day or one-day sessions at the time of professional meetings, or they can continue for longer periods, e.g., three- or five-day sessions in the summer (or during the academic year) with a more detailed focus.
- 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.
- Provide easy access to user support through phone, email, online chat, user forums, and tutorials.
- Track all user questions in a database that creates an accumulating knowledge base and that can also serve to generate Frequently Asked Questions.
- Create tutorials, some of which may be offered in video format, to provide help in using the data, the online analysis system, and the major statistical software packages.
- Establish moderated user forums to provide the foundation for an online community of researchers and students who can discuss their experiences using data and learn from each other.

### ***Lessons learned***

- Training programs must be well-planned, with a high level of substantive, methodological, and technical expertise, in order for participants to benefit from the experience. While data producers are usually those who best understand their data, they may not have the resources or desire to provide ongoing user support for the research community. Some may delegate this task to a data archive, but a joint approach, with data archives providing basic user support and data producers addressing more complicated substantive questions, often works best.
- Complex data sets often require specialized training in their proper use. If, for example, data collection methods or sampling frames change between different waves or in different countries, or researchers need to choose between a variety of weights in their analyses, there is no real substitute for intensive training and ongoing user support.

## 7. Produce comprehensive documentation for all public-use data files.

### *Rationale*

High-quality documentation is essential for effective data use. Data producers must strive to provide documentation, commonly referred to as [metadata](#), on all aspects of the survey or statistical life cycle, from initial planning through final data production and its release to the research community.

### *Procedural steps*

- Keep good records from the very beginning of the project and make every attempt to record important project events at the time they occurred. This will assist secondary analysts in understanding the goals and purpose of each survey.
- Update documentation continually during the entire life cycle of the project and preserve old versions of key files.
- For cross-national surveys, provide complete information about how the survey was conducted in each country, and describe specific procedures and practices involving data collection and data processing activities.
- Consider adopting the [Data Documentation Initiative \[1\]](#) standard for producing metadata. The use of this emerging standard, which is based on the use of [XML \(eXtensible Markup Language\)](#), allows specification of each metadata element (e.g., title of the survey, name of the principal investigators, type of sampling) for storage and future searching.

### *Lessons learned*

- XML metadata markup offers new opportunities for data producers to create their documentation, as well as several advantages to users of the documentation:
  - All information that the analyst needs is available in a core document, from which other products, such as text files that contain the necessary information to run statistical analyses in software programs, can be produced.
  - The XML file can be viewed with Web browsers and lends itself to Web display and navigation.
  - 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 software agents written to assist the research process.
- Preparing documentation in DDI format at the outset of a project means that the documentation will also be suitable for archival deposit and preservation, because it will contain all of the information necessary to describe all of aspects of the corresponding data files. DDI XML should ideally be generated by the CAI system used to collect data, but can also be collected from paper and pencil surveys through access to the information in the original questionnaire.
- Although few principal investigators of survey data have yet produced full DDI-compliant metadata, the few examples [1] that exist illustrate the importance of using this developing standard at the variable level. New use cases, currently under preparation, will demonstrate additional features, such as:
    - The presentation of instrument documentation, so that users can track the logic of the questionnaire.
    - The creation of questions banks, comprising everything asked in multi-year studies, years they were asked, differences in question wording, etc.
    - The establishment of links to the documentation of related surveys – e.g., 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.

## **8. Make quality control an integral part of all dissemination steps.**

### ***Rationale***

Dissemination requires the ongoing availability of data and documentation files though constantly new versions of hardware, software, and possible changes in management and staff. Clear procedures must be in place to make certain all files are readable as statistical and word processing software systems change over time.

### ***Procedural steps***

- Test archived files periodically to verify user accessibility.
- Establish procedures early in the survey life cycle to insure that all important files are preserved.
- Create electronic versions of all project materials, whenever feasible.

- Develop specific procedures for assessing disclosure risk to respondents and execute these procedures whenever public-use files are produced.
- Produce and implement procedures to distribute restricted-use files if applicable.
- Provide data files in all the major statistical software packages and test all thoroughly before they are made available for dissemination.
- Designate resources to provide user support and training for secondary researchers.

### ***Lessons learned***

- Data producers should strongly consider a preservation strategy before putting files online for people to download. For example, many data and documentation files available on Web sites undergo frequent changes and updates. When updates are made, the older version of the files is often no longer available. This may make it difficult, if not impossible, to replicate previous analyses done by the user, or to test the assumptions and results of analyses done by others.

## Glossary

<b>ASCII files</b>	Data files in American Standard Code for Information Interchange (ASCII) format.
<b>Bottom coding</b>	A type of coding in which values that exceed the predetermined minimum value are reassigned to that minimum value or are recoded as missing data.
<b>Confidentiality</b>	Securing the identity of, as well as any information provided by, the respondent, in order to ensure to the greatest extent possible that public identification of an individual participating in the study and/or his individual responses does not occur.
<b>Constructed variable</b>	A recoded variable, one created by data producers or archives based on the data originally collected. An example might be the creation of a variable called POVERTY from information collected on the income of respondents.
<b>Data Documentation Initiative (DDI)</b>	An international effort to establish a standard for technical documentation describing social science data. A membership-based Alliance is developing the DDI specification, which is written in XML.
<b>Data life cycle</b>	The history of a data collection from initial proposal planning and writing to final dissemination of the data, research findings, and preservation strategies.
<b>Disclosure analysis</b>	The process of protecting the <a href="#">confidentiality</a> of data. It involves limiting the amount of detailed information disseminated and/or masking data via noise addition, data swapping, generation of simulated or synthetic data, etc.
<b>Inconsistent responses</b>	Inappropriate responses to branched questions. For instance, one question might ask if the respondent attended church last week; a response of "no" should skip the questions about church attendance and code the answers to those questions as "inapplicable." If those questions were coded any other way than "inapplicable," this would be inconsistent with the skip patterns of the survey instrument.

<b>Metadata</b>	Data that describes other data. The term encompasses a broad spectrum of information about the survey, from study title to sample design, details such as interviewer briefing notes, contextual data and/or information such as legal regulations, customs, and economic indicators.
<b>Microdata</b>	Data about variables within a behavioral unit, such as an individual or a corporation. Micro-data is often contrasted with aggregate data, which is about groups of behavioral units, such as individuals grouped by race, sex, or class, or corporations grouped by economic sector.
<b>Missing data</b>	The lack of information on individual data items for a <a href="#">sample element</a> where other data items were successfully obtained.
<b>'Portable' file</b>	A file that can be used by a variety of software on a variety of hardware platforms.
<b>Public use data files</b>	A data file, stripped of respondent identifiers, that is distributed for the public to analyze.
<b>Restricted-use data files</b>	A file that includes individually identifiable information that is confidential and protected by law. Restricted-use data files are not required to include variables that have undergone coarsening disclosure risk edits. These files are available to researchers under controlled conditions.
<b>Statistical data</b>	Data from a survey or administrative source used to produce statistics.
<b>Survey data</b>	Information collected by researchers which encompasses any measurement procedures that involve asking questions of respondents.
<b>Top coding</b>	A type of coding in which values that exceed the predetermined maximum value are reassigned to that maximal value or are recoded as missing data.
<b>Trusted digital repository</b>	A repository whose mission is to provide reliable, long-term access to managed digital resources to its designated community, both now and in the future.

**Undocumented codes** Codes that are not authorized for a particular question. For instance, if a question that records the sex of the respondent has documented codes of "1" for female and "2" for male and "9" for "missing data," a code of "3" would be an "undocumented code."

**XML (eXtensible Markup Language)** The eXtensible Markup Language (XML) is a simple dialect of SGML. Its goal is to enable generic SGML to be served, received, and processed on the Web in the way that is now possible with HTML. XML was designed for ease of implementation and for interoperability with both SGML and HTML.

## References

- [1] Data Documentation Initiative (DDI). Retrieved Sept. 15, 2008 from <http://www.ddialliance.org/>
- [2] HIPAAs. Examples of Privacy Violations. Retrieved Sept. 15, 2008 from <http://www.hipaaps.com/examples.html>.
- [3] International Monetary Fund's Dissemination Standards Bulletin Board. Retrieved Sept. 15, 2008 from <http://dsbb.imf.org/Applications/web/dsbbhome/>.
- [4] Inter-University Consortium for Political and Social Research (ICPSR). Data Sharing for Demographic Research. Retrieved Sept. 15, 2008 from <http://www.icpsr.umich.edu/DSDR/rduc/>
- [5] National Digital Archive of Datasets (NDAD). Retrieved Sept. 15, 2008 from <http://www.ndad.nationalarchives.gov.uk/>
- [6] O'Rourke, J. M., Roehrig, S., Heeringa, S. G., Reed, B. G., Birdsall, W.C., Overcashier, M., et al. (2006). Solving problems of disclosure risk while retaining key analytic uses of publicly released microdata. *Journal of Empirical Research on Human Research Ethics*, 1(3), 63-84.
- [7] Royal Statistical Society & the UK Data Archive. (2002). *Preserving & sharing statistical material*. UK Data Archive: Essex. Retrieved Sept. 15, 2008 from <http://www.data-archive.ac.uk/news/publications/PreservingSharing.pdf>
- [8] Sicinski, A. (1970). "Don't Know" answers in cross-national surveys, *The Public Opinion Quarterly*, 34(1), 126-129.
- [9] Van Diessen, R. & Steenbergen, J. (2002). *Long Term Preservation Study of the DNEP Project – an Overview of the Results*. Amsterdam: IBM Netherlands. Retrieved Sept. 15, 2008 from <http://www-05.ibm.com/nl/dias/resource/overview.pdf>

## Further Reading

Allum, P. & and Mehmet A. Economic Data Dissemination What Influences Country Performance On Frequency and Timeliness? November 2001IMF Working Paper No. 01/173. Retrieved Sept. 15, 2008 from [http://papers.ssrn.com/sol3/papers.cfm?abstract\\_id=880222](http://papers.ssrn.com/sol3/papers.cfm?abstract_id=880222)

*Handbook on Civil Registration and Vital Statistics Systems. Policies and Protocols for the Release and Archiving of Individual Records.* Department of Economic and Social Affairs, United Nations Statistics Division. *Handbooks on Civil Registration and Vital Statistics Systems. Studies in Methods Series F, No. 70, 1998.* Retrieved Sept. 15, 2008 from [http://unstats.un.org/unsd/publication/SeriesF/SeriesF\\_70E.pdf](http://unstats.un.org/unsd/publication/SeriesF/SeriesF_70E.pdf)

International Federation of Data Organizations Data Access and Conditions. Retrieved Sept. 15, 2008 from [http://www.ifdo.org/data/data\\_access\\_conditions.html](http://www.ifdo.org/data/data_access_conditions.html)

Inter-university Consortium for Political and Social Research (ICPSR). *The Guide to Social Science Data Preparation and Archiving: Best Practice Throughout the Data Life Cycle, Version 3, 2005.* Retrieved Sept. 15, 2008 from <http://www.icpsr.com/ICPSR/access/dataprep.pdf>

The Dataverse Network Project. Retrieved Sept. 15, 2008 from <http://thedata.org/>

United Nations Statistics Division. Retrieved Sept. 15, 2008 from <http://unstats.un.org/unsd/default.htm>