Methods for Public-Use Microdata Files

One method of publishing the information collected in a census or survey is to release a public-use microdata file (see Section II.D). A microdata file consists of records at the respondent level where each record on the file represents one respondent. Each record consists of values of characteristic variables for that respondent. Typical variables for a demographic microdata file are age, race, and sex of the responding person. Typical variables for an establishment microdata file are Standard Industrial Classification (SIC) code, employment size, and value of shipments of the responding business or industry. Most public-use microdata files contain only demographic microdata. The disclosure risk for most kinds of establishment microdata is much higher than for demographic microdata. We explain the reasons for this in Section C.4 of this chapter.

This report concerns **public-use** microdata files that are available at a fee to anyone who wishes to purchase them. In addition to or instead of public-use files, some agencies offer **restricted-use** microdata files. Access to these files is restricted to certain users at certain locations and is governed by a restricted use agreement (Jabine, 1993a).

To protect the confidentiality of microdata, agencies remove all obvious identifiers of respondents, such as name and address, from microdata files. However, there is still a concern that the release of microdata files could lead to a disclosure. Some people and some businesses and industries in the country have characteristics or combinations of characteristics that would make them stand out from other respondents on a microdata file. A statistical agency releasing a microdata file containing confidential data must do its best to ensure that an outside data user cannot correctly link a respondent to a record on the file. Aside from not releasing any microdata, there is no way of removing all disclosure risk from a file; however, agencies must make reasonable efforts to minimize this risk and still release as much useful information as possible.

In 1962, the Social Security Administration's Office of Research and Statistics began releasing microdata files on tape from their Continuous Work History Sample to other Federal and State agencies. There were essentially no restrictions on these files, and they were later used extensively by non-government researchers. The first broad release of a public-use microdata file occurred in 1963 when the Census Bureau released a file consisting of a 1 in 1,000 sample from the 1960 Census of Population and Housing. A few years later, the Census Bureau publicly released a microdata file from the Current Population Survey. Currently, unrestricted microdata files are standard products of all Census Bureau demographic surveys. They are available to any purchaser, and researchers use them extensively (Greenberg and Zayatz, 1991). Several other Federal agencies including the National Center for Education Statistics, National Center for Health Statistics, Energy Information Administration, and Internal Revenue Service currently release microdata files.

This chapter describes the disclosure risk associated with microdata files, mathematical frameworks for addressing the problem, and necessary and stringent methods of limiting disclosure risk.

A. Disclosure Risk of Microdata

Statistical agencies are concerned with a specific type of disclosure, and there are several factors that play a role in the disclosure risk of a microdata file.

A.1. Disclosure Risk and Intruders

Most national statistical agencies collect data under a pledge of confidentiality. Any violation of this pledge is a disclosure. An outside user who attempts to link a respondent to a microdata record is called an **intruder**. The disclosure risk of a microdata file greatly depends on the motive of the intruder. If the intruder is hunting for the records of specific individuals or firms, chances are that those individuals or firms are not even represented on the file which possesses information about a small sample of the population. In this case, the disclosure risk of the file is very small. The risk is much greater, on the other hand, if the intruder is attempting to match *any* respondent with their record simply as a challenge or in order to discredit the agency that published the file. We can measure disclosure risk only against a specific compromising technique that we assume the intruder to be using (Keller-McNulty, McNulty, and Unger, 1989).

These issues as well as the contents of the file should be considered when an agency discusses the potential release of a proposed microdata file.

A.2. Factors Contributing to Risk

There are two main sources of the disclosure risk of a microdata file. One source of risk is the existence of high visibility records. Some records on the file may represent respondents with unique characteristics such as very unusual jobs (movie star, Federal judge) or very large incomes (over one million dollars). An agency must decrease the visibility of such records.

The second source of disclosure risk is the possibility of matching the microdata file with external files. There may be individuals or firms in the population that possess a unique combination of the characteristic variables on the microdata file. If some of those individuals or firms happen to be chosen in the sample of the population represented on that file, there is a disclosure risk. Intruders potentially could use outside files that possess the same characteristic variables and identifiers to link these unique respondents to their records on the microdata file.

Knowledge of which individuals participated in a survey, or even which areas were in sample, can greatly help an intruder to identify individuals on a microdata file from that survey. Warning survey respondents not to tell others about their participation in the survey might alleviate this problem, but it also might make respondents wary of participating in the survey.

The disclosure risk of a microdata file is greatly increased if it contains administrative data or any other type of data from an outside source linked to survey data. Those providing the administrative data could use that data to link respondents to their records on the file. This is not to imply that providers of administrative data would attempt to link files, however, it is a theoretical possibility and precautions should be taken.

The potential for linking files (and thus the disclosure risk) is increased as the number of variables common to both files increases, as the accuracy or resolution of the data increases, and as the number of outside files, not all of which may be known to the agency releasing the microdata file, increases. Also, as computer technology advances, it becomes quicker, easier, and less costly to link files.

The disclosure risk of a microdata file increases if some records on the file are released on another file with more detailed or overlapping recodes (categorizations) of the same variables. Likewise, risk increases if some records on the file are released on another file containing some of the same variables and some additional variables.

As a corollary, there is greater risk when the statistical agency explicitly links the microdata file to another published microdata file. This happens in the case of longitudinal surveys, such as the Census Bureau's Survey of Income and Program Participation, where the same respondents are surveyed several times. Risk is increased when the data from the different time periods can be linked for each respondent. Changes that an intruder may or may not see in a respondent's record (such as a change in occupation or marital status or a large change in income) over time could lead to the disclosure of the respondent's identity.

The disclosure risk of a file increases as the structure of the data becomes more complex. If two records are known to belong to the same cluster (for example, household), there is a greater risk that either one may be identified (even if no information about the cluster per se is provided).

A.3. Factors that Naturally Decrease Risk

Probably the factor with the biggest role in decreasing risk is the fact that almost all microdata files contain records that represent only a sample of the population. As we stated previously, if an intruder possesses such a microdata file and is looking for the record of a specific individual or firm, chances are that that individual or firm is not even represented on the file. Also, records on such a file that are unique compared with all other records on the file need not represent respondents with unique characteristics in the population. There may be several other individuals or firms in the population with those same characteristics that did not get chosen in the sample. This creates a problem for an intruder attempting to link files.

The disclosure risk of the file can be decreased even further if only a subsample of the sampled population is represented on the file. Then, even if an intruder knew that an individual or firm participated in the survey, he or she still would not know if that respondent appeared on the file. Data users, however, generally want the whole sample.

Another naturally occurring factor that decreases the risk of disclosure is the age of the data on microdata files. When an agency publishes a microdata file, the data on the file are usually at least one to two years old. The characteristics of individuals and firms can change considerably in this length of time. Also, the age of data on potentially matchable files is probably different from the age of the data on the microdata file. This difference in age complicates the job of linking files.

The naturally occurring noise in the microdata file and in potentially matchable files decreases the ability to link files (Mugge, 1983b). All such data files will reflect reporting variability, non-response, and various edit and imputation techniques.

Many potentially matchable files have few variables in common. Even if two files possess the "same" characteristic variables, often the variables are defined slightly differently depending on the purpose for collecting the data, and often the variables on different files will be recoded differently. These differences in variable definitions and recodes make an intruder's job more difficult.

The final factors that decrease risk are the time, effort, and money needed to link files, although as mentioned previously, as computer technology advances, these factors are diminished.

B. Mathematical Methods of Addressing the Problem

Although several mathematical measures of risk have been proposed, none has been widely accepted. Techniques for reducing the disclosure risk of microdata include methods that reduce the amount of information provided to data users and methods that slightly distort the information provided to data users. Several mathematical measures of the usefulness of disclosure-limited data sets have been proposed to evaluate the trade off between protection and usefulness. However, none has been widely accepted. More research is necessary to identify the best disclosure limitation methodology sufficient for both users and suppliers of confidential microdata.

Before describing these mathematical methods of addressing the problem of disclosure risk, we must mention several mathematical and computer science problems that in some way relate to this problem. For example, various mathematical methods of matching a microdata file to an outside file can be found in literature concerning record linkage methodology. Record Linkage Techniques, 1985 -- Proceedings of the Workshop on Exact Matching Methodologies presents reprints of the major background papers in record linkage as well as discussions of current work. Another related problem concerns computer science methods of addressing disclosure risk that involve storing the confidential data in a sequential database and monitoring and restricting access to the data (Lunt, 1990). The danger that this method seeks to avoid is that a data user could gain information about an individual respondent through multiple queries of the database. The National Center for Education Statistics has released compressed data with access controls along with software that allows users to obtain weighted cross tabulations of the data without being able to examine individual data records. Due to constraints of time and space, we will not describe these techniques in detail, though we will discuss some suggested research concerning

databases in Section VII.C.1. Readers interested in these techniques are encouraged to consult the references.

B.1. Proposed Measures of Risk

Several researchers have proposed mathematical measures of the disclosure risk of a microdata file (Spruill, 1982; Duncan and Lambert, 1987; Paass, 1988; Mokken, Pannekoek, and Willenborg, 1990; Skinner, Marsh, Openshaw, and Wymer, 1990; Cox and Kim, 1991). Most include calculations of:

- the probability that the respondent for whom an intruder is looking is represented on both the microdata file and some matchable file,
- the probability that the matching variables are recorded identically on the microdata file and on the matchable file,
- the probability that the respondent for whom the intruder is looking is unique in the population for the matchable variables, and
- the degree of confidence of the intruder that he or she has correctly identified a unique respondent.

More research into defining a computable measure of risk is necessary (see Section VII.A.1).

The percent of records representing respondents who are unique in the population plays a major role in the disclosure risk of a microdata file. These records are often called **population uniques**. The records that represent respondents who are unique compared with everyone else in sample are called **sample uniques**. Every population unique is a sample unique, however, not every sample unique is a population unique. There may be other persons in the population who were not chosen in the sample and who have the same characteristics as a person represented by a sample unique. Working Paper 2 states that "uniqueness in the population is the real question, and this cannot be determined without a census or administrative file exhausting the population." This remains true for each individual record on a sample microdata file. However, since then, researchers have developed and tested several methods of estimating the percent of population uniques on a sample microdata file (Skinner and Holmes, 1992). These methods are based on subsampling techniques, the equivalence class structure of the sample together with the hypergeometric distribution, and modeling the distribution of equivalence class sizes (Bethlehem, Keller, and Pannekoek, 1990; Greenberg and Zayatz, 1991).

A measure of relative risk for two versions of the same microdata file has been developed using the classic entropy function on the distribution of equivalence class sizes (Greenberg and Zayatz, 1991). For example, one version of a microdata file may have few variables with a lot of detail on those variables while another version may have many variables with little detail on those variables. Entropy, used as a measure of relative risk, can point out which of the two versions of the file has a higher risk of disclosure.

B.2. Methods of Reducing Risk by Reducing the Amount of Information Released

Recoding variables into categories is one commonly used way of reducing the disclosure risk of a microdata file (Skinner, 1992). The resulting information in the file is no less accurate, but it is less precise. This reduction in precision reduces the ability of an intruder to correctly link a respondent to a record because it decreases the percent of population uniques on the file. If an agency is particularly worried about an outside, potentially matchable file, the agency may recode the variables common to both files so that there are no unique variable combinations on the microdata file, thus preventing one-to-one matches. For example, rather than release the complete date of birth, an agency might publish month and year of birth or only year of birth. Rounding values, such as rounding income to the nearest one thousand dollars, is a form of recoding. Recoding variables can also reduce the high visibility of some records. For example, if occupation is on the file in great detail, a record showing an occupation of United States Senator in combination with a geographic identifier of Delaware points to one of two people. Other variables on the file would probably lead to the identification of that respondent. Occupation could be recoded into fewer, less discriminatory categories to alleviate this problem.

Another commonly used way of reducing the disclosure risk of a file is through setting top-codes and/or bottom-codes on continuous variables (see Section II.D.2). A **top-code** for a variable is an upper limit on all published values of that variable. Any value greater than this upper limit is not published on the microdata file. In its place is some type of flag that tells the user what the top-code is and that this value exceeds it. For example, rather than publishing a record showing an income of $2,000,000, the record may only show that the income is > $150,000. Similarly, a **bottom-code** is a lower limit on all published values for a variable. Top- and bottom-coding reduce the high visibility of some records. Examples of top-coded variables might be income and age for demographic microdata files and value of shipments for establishment microdata files. If an agency published these variables on a microdata file with no top-coding, there would probably be a disclosure of confidential information. Examples of bottom-coded variables might be year of birth or year built for some particular structure.

Recoding and top-coding obviously reduce the usefulness of the data. However, agencies could provide means, medians, and variances of the values in each category and of all top-coded values to data users to compensate somewhat for the loss of information. Also, recoding and top-coding can cause problems for users of time series data when top-codes or interval boundaries are changed from one period to the next.

B.3. Methods of Reducing Risk by Disturbing Microdata

Since Working Paper 2 was published, researchers have proposed and evaluated several methods for disturbing microdata in order to limit disclosure risk. These techniques, described in Chapter II, slightly alter the data in a manner that hinders an intruder who is trying to match files.

Probably the most basic form of disturbing continuous variables is the addition of, or multiplication by, random numbers with a given distribution (McGuckin and Nguyen, 1988;

Sullivan and Fuller, 1989; Kim, 1990a; Skinner, 1992). This **noise** may be added to the data records in their original form or to some transformation of the data depending on the intended use of the file (Kim, 1986). Probability distributions can be used to add error to a small percent of categorical values. An agency must decide whether or not to publish the distribution(s) used to add noise to the data. Publishing the distribution(s) could aid data users in their statistical analyses of the data but might also increase disclosure risk of the data. See (Kim, 1990a) for a description of one process that involved the addition of random noise to a microdata file.

Swapping (or **switching**) and **rank swapping** are two proposed methods of disturbing microdata. In the swapping procedure, a small percent of records are matched with other records in the same file, perhaps in different geographic regions, on a set of predetermined variables (Dalenius and Reiss, 1982; Dalenius, 1988; Griffin, Navarro, and Flores-Baez, 1989). The values of all other variables on the file are then swapped between the two records. In the rank swapping procedure, values of continuous variables are sorted and values that are close in rank are then swapped between pairs of records.

Another proposed method of disturbing microdata is to randomly choose a small percent of records and blank out a few of the values on the records (see Section II.D.5). Imputation techniques are then used to impute for the values that were blanked (Griffin, Navarro, and Flores-Baez, 1989).

Blurring involves aggregating values across small sets of respondents for selected variables and replacing a reported value (or values) by the aggregate (Spruill, 1983). Different groups of respondents may be formed for different data variables by matching on other variables or by sorting the variable of interest (see Section II.D.6). Data may be aggregated across a fixed number of records, a randomly chosen number of records, or a number determined by (n,k) or p-percent type rules as used for aggregate data. For a definition of the (n,k) and p-percent rules, see Chapter IV. The aggregate associated with a group may be assigned to all members of the group or to the "middle" member (as in a moving average). See (Strudler, Oh, and Scheuren, 1986) for an application of blurring. In **microaggregation**, records are grouped based on a proximity measure of all variables of interest, and the same groups of records are used in calculating aggregates for those variables (Govoni and Waite, 1985; Wolf, 1988). Blurring and microaggregation may be done in a way to preserve variable means.

Another proposed disturbance technique involves super and subsampling (Cox and Kim, 1991). The original data are sampled with replacement to create a file larger than the intended microdata file. Differential probabilities of selection are used for the unique records in the original data set, and record weights are adjusted. This larger file is then subsampled to create the final microdata file. This procedure confuses the idea of sample uniqueness. Some unique records are eliminated through nonselection, and some no longer appear to be unique due to duplication. Some non-unique records appear to be unique due to nonselection of their clones (records with the same combination of values). Biases introduced by this method could be computed and perhaps released to users as a file adjunct.

Two other procedures have been suggested that have similar objectives, but differ from the disturbance procedures described above in that they are not applied to a set of true data records before their release. Randomized response is a technique used to collect sensitive information from individuals in such a way that survey interviewers and those who process the data do not know which of two alternative questions the respondent has answered (Kim, 1986; Dalenius, 1988). Rubin has proposed the use of multiple imputation techniques to produce a set of pseudo-data with the same specified statistical properties as the true microdata (Rubin, 1993).

B.4. Methods of Analyzing Disturbed Microdata to Determine Usefulness

There are several statistical tests that can be performed to determine the effects of disturbance on the statistical properties of the data. These include the Kolmogorov-Smirnov 2-sample test, Fischer's z-transformation of the Pearson Correlations, and the Chi-Square approximation statistic to the likelihood ratio test for the homogeneity of the covariance matrices (Wolf, 1988).

These procedures are mainly conducted to see if the means and the variance-covariance and correlational structure of the data remain the same after disturbance (Voshell, 1990). Even if these tests come out favorably, disturbance can still have adverse effects on statistical properties such as means and correlational structure of subsets and on time series analyses of longitudinal data. If an agency knows how the file will be used, it can disturb the data in such a way that the statistical properties pertinent to that application are maintained. However, public-use files are available to the entire public, and they are used in many ways. Levels of disturbance needed to protect the data from disclosure may render the final product useless for many applications. For this reason, almost no public-use microdata files are disturbed, and agencies, instead, attempt to limit disclosure risk by limiting the amount of information in the microdata files. Disturbance may be necessary, however, when potentially linkable files are available to users, and recoding efforts do not eliminate population uniques.

C. Necessary Procedures for Releasing Microdata Files

Before publicly releasing a microdata file, a statistical agency must attempt to preserve the usefulness of the data, reduce the visibility of respondents with unique characteristics, and ensure that the file cannot be linked to any outside files with identifiers. While there is no method of completely eliminating the disclosure risk of a microdata file, agencies should perform the following procedures before releasing a microdata file to limit the file's potential for disclosure. Statistical agencies have used most of these methods for many years. They continue to be important.

C.1. Removal of Identifiers

Obviously, an agency must purge a microdata file of all direct personal and institutional identifiers such as name, address, Social Security number, and Employer Identification number.

C.2. Limiting Geographic Detail

Geographic location is a characteristic that appears on all microdata files. Agencies should give geographic detail special consideration before releasing a microdata file because it is much easier for an intruder to link a respondent to the respondent's record if the intruder knows the respondent's city, for example, rather than if he or she only knows the respondent's state.

In Working Paper 2, it was stated that the Census Bureau will not identify on a microdata file any geographic region with less than 250,000 persons in the sampling frame. After Working Paper 2 was published, however, the Census Bureau determined that this geographic cut-off size was excessive for most surveys. Currently, the Census Bureau will not identify any geographic region with less than 100,000 persons in the sampling frame. A higher cut-off is used for surveys with a presumed higher disclosure risk. Microdata files from the Survey of Income and Program Participation, for example, still have a geographic cut-off of 250,000 persons per identified region. Agencies releasing microdata files should set geographic cut-offs that are simply lower bounds on the size of the sampled population of each geographic region identified on microdata files (Greenberg and Voshell, 1990). This is easier said than done. Decisions of this kind are often based on precedents and judgement calls. More research is needed to provide a scientific basis for such decisions (Zayatz, 1992a).

Some microdata files contain contextual variables. Contextual variables are variables that describe the area in which a respondent or establishment resides but do not identify that area. In general, the areas described are smaller than areas normally identified on microdata files. Care must be taken to ensure that the contextual variables do not identify areas that do not meet the desired geographic cut-off. An example of a contextual variable that could lead to disclosure is average temperature of an area. The Energy Information Administration adds random noise to temperature data (because temperature data are widely available) and provides an equation so the user can calculate approximate heating degree days and cooling degree days (important for regression analysis of energy consumption).

C.3. Top-coding of Continuous High Visibility Variables

The variables on microdata files that contribute to the high visibility of certain respondents are called **high visibility variables**. Examples of continuous high visibility variables are income and age for demographic microdata files and value of shipments for establishment microdata files. As stated previously, if an agency published these variables on a microdata file with no top-coding, there would probably be a disclosure of confidential information. For example, intruders could probably correctly identify respondents who are over the age of 100 or who have incomes of over one million dollars.

For 10 years following the 1980 Census of Population and Housing, the Census Bureau used a top-code of $100,000 on all types of income variables. Following the 1990 Census of Population and Housing, the Bureau's Microdata Review Panel members raised the top-code for some types of income that are usually high and lowered the top-code for income variables that are usually low. The Panel often requests that a given percentage of values be top-coded for a variable.

The percentage may depend on the sensitivity of the variable. The Bureau will be providing the medians of top-coded values on microdata files from the 1990 Census of Population and Housing. Appropriate top-codes (and/or bottom-codes in some cases) should be set for all of the continuous high visibility variables on a microdata file. Top-coded records should then only show a representative value for the upper tail of the distribution, such as the cut-off value for the tail or the mean or median value for the tail, depending on user preference.

C.4. Precautions for Certain Types of Microdata

C.4.a. Establishment Microdata

Almost all microdata files now publicly released contain demographic microdata. It is presumed that the disclosure risk for establishment microdata is higher than that for demographic microdata. Establishment data are typically very skew, the size of the establishment universe is small, and there are many high visibility variables on potential establishment microdata files. Also, there are a large number of subject matter experts and many possible motives for attempting to identify respondents on establishment microdata files. For example, there may be financial incentives associated with learning something about the competition. Agencies should take into account all of these factors when considering the release of an establishment microdata file.

C.4.b. Longitudinal Microdata

There is greater risk when the microdata on a file are from a longitudinal survey where the same respondents are surveyed several times. Risk is increased when the data from the different time periods can be linked for each respondent because there are much more data for each respondent and because changes that may or may not occur in a respondent's record over time could lead to the disclosure of the respondent's identity. Agencies should take this into account when considering the release of such a file. One piece of advice is to plan ahead. Releasing a first cross-sectional file without giving any thought to future plans for longitudinal files can cause unnecessary problems when it comes to releasing the latter. One needs to consider the entire data collection program in making judgments on the release of public use microdata.

C.4.c. Microdata Containing Administrative Data

The disclosure risk of a microdata file is increased if it contains administrative data or any other type of data from an outside source linked to the survey data. Those providing the administrative data could use that data to link respondents to their records. This is not to imply that providers of administrative data would attempt to link files, however, it is a theoretical possibility and precautions should be taken. At the very least, some type of disturbance should be performed on the administrative data or the administrative data should be categorized so there exists no unique combination of administrative variables. This will reduce the possibility that an intruder can link the microdata file to the administrative file. Many feel that agencies should not release such microdata at all or should release it only under a restricted access agreement.

C.4.d. Consideration of Potentially Matchable Files and Population Uniques

Statistical agencies must attempt to identify outside files that are potentially matchable to the microdata file in question. Comparability of all such files with the file in question must be examined. The Census Bureau's Microdata Review Panel recently began using methods that estimate the number of population uniques on a microdata file to determine if that file was matchable to an outside file on a certain set of key variables. The National Center for Education Statistics has matched microdata files under consideration for release to commercially available school files looking for unique matches.

D. Stringent Methods of Limiting Disclosure Risk

There are a few procedures that can be performed on microdata files prior to release that severely limit the disclosure risk of the files. One must keep in mind, however, that the usefulness of the resulting published data will also be extremely limited. The resulting files will contain either much less information or information that is inaccurate to a degree that depends on the file and its contents.

D.1. Do Not Release the Microdata

One obvious way of eliminating the disclosure risk of microdata is to not release the microdata records. The statistical agency could release only the variance-covariance matrix of the data or perhaps a specified set of low-order finite moments of the data (Dalenius and Denning, 1982). This greatly reduces the usefulness of the data because the user receives much less information and data analyses are restricted.

D.2. Recode Data to Eliminate Uniques

Recoding the data in such a way that no sample uniques remain in the microdata file is generally considered a sufficient method of limiting the disclosure risk of the file. A milder procedure allowing for broader categorization--recoding such that there are no population uniques--would suffice. Recoding the data to eliminate either sample or population uniques would likely result in very limited published information.

D.3. Disturb Data to Prevent Matching to External Files

Showing that a file containing disturbed microdata cannot be successfully matched to the original data file or to another file with comparable variables is generally considered sufficient evidence of adequate protection. Several proximity measures should be used when attempting to link the two files (Spruill, 1982). An alternative demonstration of adequate protection is that no exact match is correct or that the correct match for each record on a comparable file is not among the K closest matches (Burton, 1990).

Microaggregation could be used to protect data, perhaps using (n,k) or p-percent type rules as used for tables. In this way, no individual data are provided, and intruders would be prevented

from matching the data to external files. For a definition of the (n,k) and p-percent rules, see Chapter IV.

Microaggregation and other methods of disturbance that hinder file matching, however, result in inaccurate published data. Taken to a degree that would absolutely prevent matching, the methods would usually result in greatly distorted published information.

E. Conclusion

Public-use microdata files are used for a variety of purposes. Any disclosure of confidential data on microdata files may constitute a violation of the law or of an agency's policy and could hinder an agency's ability to collect data in the future. Short of releasing no information at all, there is no way to completely eliminate disclosure risk. However, there are techniques which, if performed on the data prior to release, should sufficiently limit the disclosure risk of the microdata file. Research is needed to understand better the effects of those techniques on the disclosure risk and on the usefulness of resulting data files (see Section VI.A.2).