Supporting Document
Checklist on Disclosure Potential of Data

Version 1.3

Issued: 11 Apr 07

Census Bureau Standard
Disclosure Review

Authored by:

Laura Zayatz
Chair, Disclosure Review Board
Statistical Research Division

USCENSUSBUREAU
Helping You Make Informed Decisions

Document Management & Control[1]

Version	Issue Date	Approval	Description
1.0	12 Mar 04	Associate Directors	Initial Release
1.1	14 Jan 05	Configuration Mgr.	Reformatted to comply with Census Bureau Identity Standard and Quality Program Document Management Plan
1.2	09 Mar 06	Configuration Mgr.	Inserted hyperlink for main standard.
1.3	11 Apr 07	M&S Council	Reformatted. Added sections for synthetic data.

[1] **The most current version of this document is maintained on the Census Bureau Intranet and may be accessed from the Methodology & Standards Council Website.**

CHECKLIST ON DISCLOSURE POTENTIAL OF DATA

Do you need to fill out this form?

To reduce your reporting burden, it is not necessary to complete this checklist for every issuance of a repetitive survey or census. You need only prepare a memorandum to the chair of the Disclosure Review Board if all of the following criteria are met: geographic information is not changed, no new subject matter is introduced, the disclosure avoidance measures approved for the first data release are implemented on all subsequent releases, and the checklist was completed for an earlier release. Note that even though the checklist is unnecessary in this case, the data will still be reviewed in order to take into account any change in the public availability of information that could be used for re-identification.

The form must be completed for all types of data release (for example, demographic and economic microdata, demographic and economic tabular data, audio tapes, etc.) except for survey demographic tabular data that do not identify geographic areas with less than 100,000 persons in the sampled area.

NOTE: After this form has been filled out, it contains information, **the release of which is prohibited by title 13 U.S.C. and is for Bureau of the Census Official Use Only**.

CHECKLIST ON DISCLOSURE POTENTIAL OF DATA

CENSUS/SURVEY TITLE:_____ DATE:_____

Project Mgr. Name:_____ Div._____ Br._____ Ph._____

Sponsoring Agency:

Age of Data at Proposed Time of Release:_____ (years)

Check the applicable category below:

[ ] This application is for a single data product.

[ ] This application is for a series of releases with substantially the same content.
(Specify the interval at which future products will be released.)

[ ] This application is for the re-release of an approved product, with the addition of supplemental or previously unreleased data.
(If marked, give the date the original product was submitted to the DRB/MRP)

(Only those checklist questions for which the answers are now different need to be completed)

This checklist is divided into three sections. Please answer all questions for the applicable section(s). If you need more space for an answer, please attach a continuation sheet and identify the number of the question.

- Section 1 (pages 1-12) asks questions about microdata. A microdata file consists of records at the respondent level. Each record contains values of variables for a person, household, establishment, or other unit. Most microdata files contain demographic information. Some questions in this section may not be applicable for establishment-based files.

- Section 2 (pages 12-13) deals with demographic tabular (frequency count) data. Frequency count data present the number of units (persons, households, etc.) in a cell. This checklist only needs to be completed if the tabulations are from a census or if the identified geographic areas contain less than 100,000 persons in sample.

- Section 3 (pages 13-17) concerns establishment tabular (magnitude) data. Magnitude data present the sum of a quantity of interest for all units in a cell.

Section 1. Microdata

1.1. Geographic Information on the File

General Rule: All geographic areas that are identified must have a minimum of 100,000 persons in the sampled area.

1.1.1. What level of geography will be shown on the file?

In addition to explicit geographic identifiers on the file, the data items, record identifiers, or file structure may provide additional geographic information by inference. Therefore, steps must be taken to avoid inadvertently identifying geographic areas that do not meet the 100,000 minimum population criteria. Potential problem areas are discussed below. For each area, please indicate the actions that have been or will be taken before the proposed file is released.

1.1.2. Primary sampling unit (PSU) or other geographic information usually is embedded in control numbers designed for internal use.

How will this problem be avoided on the released file?
[ ] Control numbers deleted or do not contain geographic information.

[ ] Control numbers scrambled; describe.

[ ] Other; describe.

1.1.3. Records in many databases are sequenced so that the first cases are in the lower numbered PSU or county that is first in alphabetic order.

Briefly, describe how the records on this file will be sequenced to avoid such geographic inferences.

1.1.4. Data items that imply specific geography of residence may reveal more than the explicit identifiers displayed on the population table prepared for the Board. Examples: inclusion of Spanish surname (coded only in five southwestern states) when the explicit identification of that group of states will not be on the file; a migration code specifying movement from a metro area to a nonmetro area when metro-nonmetro will not be included as part of the geographic identifiers; residence within X miles of a nuclear

reactor or an airport when there is only one in an identified geographic area; telephone area code; or latitude and longitude coordinates.

List all items that will be deleted for this reason:

List all other items that you think might have geographic significance, but could not decide if they should be deleted.

1.1.5. Sampling information also may provide some geographic indicators. For example, certain weights may distinguish between self-representing and nonself-representing PSUs or identify types of areas intentionally oversampled. Also, codes for "Durbin type," "Hit number," etc., may be related to geography.

List all sampling information including that for variance estimation that will be deleted for confidentiality reasons or subsampling plans to make weights less identifying:

List all other sampling information that you think might have geographic significance, but could not decide if it should be deleted:

1.1.6. Based on available information, will any data item on the file identify residence in a particular type of institution (such as a prison or nursing home) of which there may be only one in an identified area; or for which a system of records could be obtained?

[ ] Yes – Identify the type of institution _____.

[ ] No

1.2. File Contents Presenting an Unusual Risk of Individual Disclosure

The disclosure criteria for public-use microdata require a review of each file to determine if any of the proposed contents present an unusual risk of individual disclosure. The DRB has identified several disclosure avoidance measures that can be taken to protect the confidentiality of individual respondents. The measures are discussed below and relevant information pertaining to the proposed file is requested to assist the Board in its review.

1.2.1. Names, addresses, and other unique numeric identifiers such as Social Security, Medicare or Medicaid numbers must be removed from the file.

1.2.2. High income is a visible characteristic of individuals or households and is considered to be a sensitive item of information. Therefore, each income figure on the file, whether for households, persons, or families, including total income and its individual components should be topcoded. Topcodes for income variables that apply to the total universe (person/households) should include at least ½ of 1 percent of all cases. For income variables that apply to subpopulations, topcodes should include either 3 percent of the appropriate cases or ½ of 1 percent of all cases, whichever is the higher topcode. Exceptions to this rule are possible under certain circumstances; for example, if there is very little geographic detail. Variances from these topcode rules should be discussed with the Board well in advance of the final submission for approval to release a file.

Do all income topcodes satisfy the appropriate rule:

[ ] Yes

[ ] No – Specify percent topcoded and topcode amount and briefly summarize discussions with the Board.

1.2.3. In addition to income, certain other characteristics may make an individual more visible than others; for example, very high age, value or purchase price of own property, rent, mortgage amount. Depending on the geographic detail shown on the file, consideration should be given to topcoding (and/or collapsing) these items when they are represented as interval or ordinal variables. The Board suggests that these topcode categories include at least ½ of 1 percent of the total universe (persons/households) represented on the file (weighted counts). In a few cases, where variables apply only to very small populations, the Board may consider topcode categories including approximately 3 to 5 percent of the appropriate subpopulation. Examples of approved topcodes:

Age – 90 (Approximately ½ percent of all persons in Census 2000)

Rent for housing units paying cash rent – $1,700 (Approximately ½ percent of all housing units in Census 2000)

Heating fuel cost – $2,100 (Approximately 3 percent of all occupied housing units where heating fuel is used and is paid separately from rent or condominium fees in Census 2000)

List all items that will be topcoded (or collapsed) and the corresponding topcodes:

List all other items about which you have questions regarding the need to topcode:

1.2.4. Describe any proposed information to be released for the topcoded data items (for example, means or medians of the topcoded values).

1.2.5. There are other characteristics that may make a person highly visible, depending upon the geography, that are represented as nonordinal variables, and therefore cannot be topcoded; for example, codes indicating Foreign or Indian Tribal language spoken; detailed racial identification such as Eskimo, Aleut, Guamian, or Samoan; codes for place of prior residence, etc. In these cases, the amount of detail on the file may have to be collapsed into larger categories.

List all items that will be collapsed (or deleted) for confidentiality reasons:

List any other items about which you have questions regarding the need to collapse the detail:

1.2.6. Contextual Variables (variables describing the area in which a person or household resides)

Identify any contextual variables and the level at which they are coded.

List all contextual variables that will be collapsed (or deleted) for confidentiality reasons:

List any other contextual items about which you have questions regarding the need to collapse the detail:

1.3. Disclosure Risks Associated with the Ability to Match to External Files

Efforts must be made to avoid the potential for matching microdata on this file to data on external files, because external files usually contain names and addresses, and thus

can be used to identify survey respondents. Such matching may be possible if the survey contains highly specific characteristics that are also found on mailing lists or administrative records maintained by other agencies or organizations. For example, the inclusion of vehicle make, model, and year in conjunction with specific geographic identifiers is unacceptable because these items can be matched to automobile registration lists that contain name and address. These items probably could be left on the file if they were recoded into broad categories. Some examples are: manufacturer's list of purchasers of particular major durable goods (for example, airplanes); voter registration lists in some states; Federal, state, or local tax records; criminal justice system records; state hunting and fishing license registers; and membership rosters of certain trade associations.

Matching is also highly possible if the sampling frame for a survey comes from a source outside the Census Bureau. The agency that provided the sampling frame may be able to match survey records to its original records, particularly if the survey records include data from the originating agency's files; e.g., amount of program benefit received, date of entry into program.

1.3.1. Outside files

1.3.1.1. Are you aware of administrative records, a mailing list, or any other outside file that contains data also included in this proposed file?

[ ] Yes – Identify the list(s)

[ ] No

1.3.1.2. Were any of the sample cases contained in the proposed file selected from a list provided by a source outside the Census Bureau?

[ ] Yes – Identify the source and describe how and by whom sample cases were selected from the list:

[ ] No

1.3.2. Matching

When an external file exists, several steps may be taken to avoid the possibility of matching survey data to this file; for example, selected items may be deleted or recoded, or "noise" (i.e., small amounts of random variation) may be introduced into these items. The Board cannot specify in advance exactly which steps must be taken to avoid the potential for matching. However, it does consider several factors in

determining the risk associated with releasing a file when the possibility of matching to external data bases exists; 1) the number of variables available for matching purposes, 2) the resources needed to perform the match, 3) the age of the data, 4) the accessibility, reliability, and completeness of the external file, and 5) the sensitivity or uniqueness of the data. Some factors that make matching easier are listed below and information is requested on steps that will be taken before the file is released to avoid the matching potential. (NOTE: This information is necessary even if you are not aware of any external files that could be used in matching.)

Matching is easier –

1.3.2.1. ...if any data item or combination of items isolates any small and readily identifiable population. The inclusion of codes that identify very small population segments should be avoided; for example, Indian tribes or detailed occupation in combination with highly specific geography. Normally one does not have to consider more than one variable at a time unless that group of variables is likely to appear together on a file or list. For example, age and sex are likely to appear together on external files but not country of birth and occupation; thus, it should not be necessary to protect against rare occurrences like Russian-born architects.

List all data item(s) proposed for inclusion on the file that isolate a small, readily identifiable population.

List all data item(s) that will be altered (i.e., deleted, recoded, noise added) for this reason.

1.3.2.2. ...if the file includes a substantial fraction of a population (say $p > 0.5$). Examples: large employers, high-income individuals, doctors, scientists of a specified type, or inmates of certain types of institutions. Additional subsampling frequently is required within certain strata prior to data release.

Identify these populations, if any are on the file, and how they will be subsampled.

1.3.2.3. ...if the file contains any information obtained from records or other sources where that information could serve as a link to an external file that has individual identifiers or detailed geographic information. Examples include fuel consumption or cost records from a utility company; neighborhood, tract, or RD summary characteristics from a decennial census; welfare or social security data from a government agency; arrest record from a police department; benefits provided to employees such as pensions and health insurance.

List all data item(s) proposed for the file that were not obtained from an interview with the respondent.

List all data item(s) altered or deleted for this reason.

1.3.2.4. ...if the file includes data items frequently used for matching, such as exact date of birth, sex, and race, or if it includes other items that should be identical on both files, such as an exact income amount, real estate taxes or other taxes, or date of entry or termination from a government-sponsored program.

List these data items, if any.

List all data item(s) altered or deleted for this reason.

1.3.2.5. ...if longitudinal data are being collected; i.e., if the data for the same respondents/units will be collected for several different reference periods. Primary concern relates to time series of data items potentially matchable to outside records; e.g., income tax or employment records. If data are collected from the same respondents more than once, indicate the frequency of interview, length of time any one unit may remain in sample, and factors affecting the likelihood of matching a sample unit from one time period to the next.

1.3.2.6. ...if highly specific geography is included on the file; for example, states, SMSAs, etc. (This geography should be presented in the Population Table.)

1.3.2.7. Describe any considerations not previously mentioned that protect against matching this file to external data; e.g., unreliability or natural noise in the data.

1.3.3. Cross Tabulations to Identify Unique Sets of Characteristics

1.3.3.1. Were any cross tabulations performed to identify sets of unique characteristics?_____.
If no, skip to 1.4.

1.3.3.2. What were the results?

1.3.3.3. Will any additional steps be taken to reduce disclosure risk based on these results?

1.4. Noise

1.4.1. Was any noise added to the data? _____.
If no, skip to 1.5.

1.4.2. What procedure(s) was used to add noise to the data? Please give specifics for that procedure (i.e. percent of records affected, distribution of noise, etc.).

Some possibilities:
- random noise
- record swapping
- rank swapping
- blanking and imputation

1.4.3. Was any attempt made to match back the noise-added data to the original file?_____.
If no, skip to 1.5.

1.4.4. How was it done and what was the rate of success in matching?

1.5. Synthetic Data

If none of these data are synthetic, skip to 1.6

1.5.1. Describe the method used to generate the synthetic data, including which variables were synthesized, which were not, which were used in the model, and the percent of records that were synthesized. Include references to more detailed documents about the procedure if available

1.6. Edited data (data values provided by respondents that we have altered) and imputed data (data values that we have created due to non-response) have their own "noise" built in. The processes of editing and imputation protect against disclosure. Please answer the questions in this section *if the values are known.*

1.6.1. What percent of records contain at least one edited data item?_____.

1.6.2. What percent of all data items were edited? _____.

1.6.3. What percent of records contain at least one imputed data item?_____.

1.6.4. What percent of all data items were imputed? _____.

1.7. Other Issues

1.7.1. Files that include every sample case or cases in strata that are sampled at high rates ($p > 0.5$) are more likely to lead to disclosure than files containing only a subsample of cases. For example, if it were known that a certain individual participated in a particular survey, one could infer that the person's record could be found in the corresponding microdata file, assuming all sample cases were available on that file.

Does this file contain

[ ] Every case

[ ] A subsample of cases (if so, specify the range of sampling rates)

1.7.2. Project managers should be aware that confidentiality problems may arise if special tabulations are made from an internal version of file, which includes detail omitted from the public use file. For example, the tabulation might provide specific geography not included on the public use file, cross-tabulated by multiple data items on the file. The Board has prepared guidelines outlining procedures for reviewing these tabulations. Please refer to these guidelines ("Disclosure Potential of Survey Tabulations Given the Availability of Public-Use Microdata") and consult with the Board if you are planning to release tabulations that make use of detail not available on the public-use file.

1.7.3. Briefly describe the sample design

1) Include a description of any stratification, clustering, and stages, including the identification of the kinds of units sampled at any stage with probability > 0.5.
2) Include a comparison and contrast of the proposed sampling units, units of enumeration, and units of analysis in the study.
3) Identify the information of the sample design (sampling plan and estimators) that will and will not be put in the public domain.

4) Describe how users will estimate sampling variances potentially identifying any proposed "nesting variables" on the proposed file layout or the design of any weights used for replication approaches.

1.7.4. Supplements

Was this information gathered as a supplement to another survey?_____.

If no, you are finished with this section of the checklist.

Can this microdata file be linked to the file produced from the main survey?_____.
If yes, what geographic information is on the main file?

Section 2. Demographic Tabular Data

2.1. The Data

2.1.1. What makes this product "non-standard" (i.e. census data or identified geographic areas with sampled populations of fewer than 100,000 persons)?

2.1.2. Is this sample or census data?_____.
If census, skip to 2.1.5.

2.1.3. Briefly describe the sample design, including sampling rates.

2.1.4. Are weights common knowledge (or could easily be inferred) so that a cell showing 10, for example, implies that only one person in the survey possessed that cell's given characteristics?

2.1.5. Give a description of the tables to be released (i.e. dimensions, variables and their categories).

2.1.6. What is the level of geography released?

2.2. Were any administrative data used to create these tables?_____.
If yes, please describe.

2.3. Edited data (data values provided by respondents that we have altered) and imputed data (data values that we have created due to non-response) have their own "noise" built in. The processes of editing and imputation protect against disclosure. Please answer the questions in this section *if the values are known.*

2.3.1. What percent of records contain at least one edited data item?_____.

2.3.2. What percent of all data items were edited?_____.

2.3.3. What percent of records contain at least one imputed data item?_____.

2.3.4. What percent of all data items were imputed?_____.

2.4. Disclosure Avoidance

What disclosure avoidance technique(s) (if any) were used for this data and why? Please provide details. Some possible techniques:

- record swapping
- blanking and imputation
- rank swapping
- random noise
- cell suppression
- controlled rounding
- generation of synthetic data (include information on which variables were synthesized, which were not, which were used in the model, and the percent of records that were synthesized)

Section 3. Establishment Tabular Data

3.1. The Data

3.1.1. Is this sample or census data?_____.

3.1.2. Are establishment counts released? _____.
If census, skip to 3.1.5.

3.1.3. Were some types of establishments selected with certainty?_____.

3.1.4. Briefly, describe the sample design, including sampling rates.

3.1.5. What data will be released and in what formats (i.e. table dimensions, variables and their detail)?

3.1.6. What is the level of geography released?

3.2. Were any administrative data used to create these tables? _____.

If yes, please describe.

3.3. Edited data (data values items provided by respondents that we have altered) and imputed data (data values that we have created due to non-response) have their own "noise" built in. The processes of editing and imputation protect against disclosure. Please answer the questions in this section *if the values are known.*

3.3.1. What percent of records contain at least one edited data item? _____.

3.3.2. What percent of all data items were edited? _____.

3.3.3. What percent of records contain at least one imputed data item? _____.

3.3.4. What percent of all data items were imputed? _____.

3.4. Disclosure Avoidance

What disclosure avoidance technique(s) (if any) were used for this data and why?
Some possible techniques:

- cell suppression
- noise
- synthetic data

If cell suppression was not used, skip to 3.4.2.

3.4.1. Cell Suppression

3.4.1.2. What rule (and with what parameters) was used to determine primary suppressions?

If census data, skip to 3.4.1.4.

3.4.1.3. How was the rule adapted to fit the survey?

3.4.1.4. What amount of protection was given to primary suppressions?

3.4.1.5. Were establishments combined by company (or farms by owner) prior to determining primary suppressions and the amount of protection needed for each primary?

3.4.1.6. Was a key item chosen in performing cell suppression? _____.

If so, what was it and why?

3.4.1.7. Was cell suppression performed by Census suppression software or by hand?

If by suppression software, skip to 3.4.1.9.

3.4.1.8. Were the suppression patterns in the tables audited? _____.

If suppression was done by hand, skip to 3.4.2.

3.4.1.9. What shortcuts (if any) were used and with what parameters?

3.4.1.10. Were any suppressions removed by hand?_____.

If so, why?

If so, how were others chosen to replace them?

3.4.1.11. Were all 3-dimensional tables audited?_____.

3.4.1.12. Will any additional information be released for values that were suppressed (i.e. ranges, medians, estimates, rounded values, values with noise, etc.)?

If so, please give details.

If noise was not used as a disclosure avoidance technique, skip to 3.4.3.

3.4.2. Noise

3.4.2.1. Which items received noise?

3.4.2.2. How was noise added to the data?

3.4.2.3. How much noise was added to the data?

3.4.3. Synthetic Data

If none of these data are synthetic, skip to 3.4.4.

3.4.3.1. Describe the method used to generate the synthetic data, including which variables were synthesized, which were not, which were used in the model, and the percent of records that were synthesized. Include references to more detailed documents about the procedure if available.

3.4.4. Were any other disclosure avoidance techniques used for this data?_____.

If so, please describe them in detail.

3.5. Treatment of Special Types of Data

3.5.1. Some data require special treatment in terms of applying disclosure avoidance techniques. Some possibilities:

- Negative valued data
- Percent/net change data
- Difference between positive values data
- Weighted average data

Did any data require special treatment?_____.

If so, which data and what was done?

3.6. Coordination of Disclosure Avoidance

3.6.1. Is this a special tabulation?_____.
If no, skip to 3.6.3.

3.6.2. All suppressions must be coordinated among all tables generated from the same data set in order to ensure that suppression patterns do not unravel each other.

Were disclosure avoidance techniques (such as cell suppression patterns) coordinated with those used for previously released standard tables?_____.

3.6.3. Has the same (or very similar) data also been released by another division or branch?

If no, you have completed the checklist.

3.6.4. Were disclosure avoidance techniques (such as cell suppression patterns) coordinated with those used by the other division/branch? _____.