National/Homeland Security and Privacy/Confidentiality Checklist and Guidance

Background

As the Federal government increases the availability of government data to the public in more easily accessible and interoperable formats, great care must be taken to protect national/homeland security and safeguard privacy and confidentiality. As required by law, agencies have in place processes for screening datasets for national/homeland security and privacy/confidentiality risks.

- The Open Government Directive requires agencies to re-visit their past decisions not to release datasets to the public. To expand access to information while complying with valid restrictions, agencies must collaborate to establish appropriate risk analysis and mitigation strategies to limit the likelihood of privacy/confidentiality and/or national/homeland security breaches as a result of Open Government efforts.
- The content of any individual dataset may not pose a threat to national/homeland security or a risk of a breach in promises of confidentiality or privacy. However, the collection of many datasets disseminated jointly, especially within the context of an environment that encourages 'mash-ups,'[1] may increase the risk. Specifically, there are concerns that the cross-government cataloging and searching of data made available through Data.gov may facilitate identification of relationships that may not otherwise be apparent.
- Therefore, there is a need for processes that protect national/homeland security and safeguard privacy and confidentiality while advancing the President's direction "to disclose information rapidly in forms that the public can readily find and use"[2] and consistent with "the presumption of openness that the President has endorsed."[3]

The Open Government Initiative Privacy and Security Working Group ("Working Group") is an interagency group led by the National Security Staff. The Working Group is composed of Executive Branch agencies with specialization in the security and privacy realms. It developed the screening procedures outlined in this document to help reduce the risk associated with the mosaic effect, in which datasets that pose no disclosure threat by themselves

[1] A mash-up combines the data from two different sources to provide useful new information.

[2] Memorandum for the Heads of Executive Departments and Agencies: Transparency and Open Government, January 21, 2009.

[3] Open Government Directive, December 8, 2009.

can create a national/homeland security concern or produce identifiable information when combined with other datasets. This is of particular concern for datasets available in formats that are conducive to *mash-ups*, as are the datasets at sites such as Data.gov.

The Working Group will continue to evaluate, and where appropriate, enhance Federal data dissemination guidelines to guard against intentional and unintentional unmasking of sensitive or personally identifiable information and/or national/homeland security-sensitive information. As additional opportunities to enhance policies and procedures for evaluating datasets for mosaic effect concerns are developed, agency training will be provided.

Introduction

The Data.gov National/Homeland Security and Privacy/Confidentiality Checklist and Guidance (referred to here as the Checklist) should be used by departments and agencies (referred to here as agencies) submitting datasets for publication on Data.gov. The Checklist augments and does not replace or supersede the processes the agencies are using to meet their existing statutory, regulatory or policy requirements for protecting national/homeland security and privacy/confidentiality interests. The Checklist is intended to ensure that national or homeland security and privacy and confidentiality concerns are addressed as datasets from across the government are published on Data.gov. The Checklist should be submitted along with the proposed dataset for publication. (This may be in the form of completing a metadata template with a set of standard questions derived from this Checklist in the Data Management System (DMS), the on-line tool for review and approval of datasets for Data.gov, or through a successor system.)

The Checklist has two parts. Part A addresses potential national or homeland security issues. Part B addresses privacy and confidentiality issues associated with releasing a dataset. Agencies should review all their datasets for the criteria listed in both Part A and Part B, unless otherwise noted. Datasets need to be reviewed not only for national/homeland security concerns and to protect the privacy and confidentiality of individuals and businesses, but also for the disclosure risk that arises from the mosaic effect. If the datasets do not contain any of the types of data described, agencies can submit the dataset and the completed Checklist (which may be in the form of answering standard questions in a metadata template associated with the DMS, or a successor system).

If a dataset contains any of the types of data listed in Part A, the submitting agency will be requested to provide a brief description of the reasoning or methods used to determine that the dataset is appropriate for publication, along with contact information for the person making the determination.

While there are no set criteria, typically datasets containing variables listed in Part A are acceptable for release through Data.gov, when:

- Historically the dataset has been available to the public; or
- The dataset is publicly available in similar forms on other, non-government web sites.

In addition, agencies should note if there are statutory or other authorities requiring or permitting the publication of the dataset.

For datasets that contain any of the types of data described in Part B, the submitting agency should provide a completed Checklist (in the form of a DMS-associated metadata template or successor system) that includes information describing the corrective actions taken to prevent the identifiability of the respondents in the dataset. Again, contact information for the Agency person responsible for the decision needs to be provided.

For all datasets, list any statutory or other legal authorities requiring or permitting publication of this dataset: _____

Part A
National/Homeland Security Flags

Agencies should already have processes in place to prevent the release of classified national security information, information marked Controlled Unclassified Information, Law Enforcement Sensitive, Sensitive Security Information, or other information not releasable under statute, Executive Order, regulation, or agency policy. The following is a list of the types of information that are often restricted from release, and that, even if not marked as restricted, could potentially raise national/homeland security concerns. The list is not exhaustive and is intended to be a guide. The fact that a dataset contains one of these types of information does not mean that it cannot be published on Data.gov, only that the agency must provide a description of why it deems publication does not pose a risk to national/homeland security.

1. The names of U.S. Government (USG) employees not publicly disclosed in an agency or commercial directory.[4]
2. Locations of sensitive Department of Defense, Intelligence Community, or other USG facilities.
3. Information regarding physical security of USG facilities (owned or leased).
4. Detailed spending information on particular security equipment or measures used by the USG (e.g., number/type of scanners, explosives detectors at a particular facility).
5. Information regarding the security of USG information or communications systems.
6. Information on physical security/protection of USG officials.
7. Information concerning USG continuity of operations/continuity of government (COOP/COG) plans.
8. Previously non-public information on USG plans for responding to disasters and emergencies.
9. Previously non-public investigative information—pertaining to a law enforcement, administrative, or inspector general investigations, government inspections, etc., or revealing law enforcement or investigative techniques, including investigative

[4] For purposes of this checklist, USG "employee" or "official" should be read to include contractors, volunteers, and other individuals with an affiliation with an agency or department of the USG. Although employee data is about persons, in this context the concern is primarily security rather than privacy.

information that alone would not reveal law enforcement or investigative techniques but that in aggregation could establish a pattern revealing such techniques.

10. Previously non-public information concerning border security or immigration enforcement such as criteria or algorithms for additional screening of goods or individuals, and security and immigration enforcement that alone or in combination could establish patrol patterns or schedules (e.g., dates/locations of border drug arrests or border control lane assignments).
11. Information relating to USG contracts for national or homeland security capabilities, or other non-public information concerning government procurement including pre-award vendor information.
12. Foreign government information—unclassified information provided to the USG in confidence by a foreign government or entity or other non-public information concerning communications or negotiations between the USG and a foreign government.
13. Detailed critical infrastructure information (dams, bridges, water, power, and nuclear facilities, telecommunications, banking/finance, cyber, ports, air travel, mass transit, food supply) that, alone or in combination with other information, could be advantageous to an adversary by revealing key nodes or other vulnerabilities.
14. Unclassified research with potential use for Weapons of Mass Destruction (biological, chemical, radiological, nuclear).
15. Previously non-public information on government planning on public health issues such as infectious disease, pandemics, quarantines, etc.

Question for Part A:

Does dataset contain national/homeland security-related information?

___ No, dataset does <u>not</u> contain national/homeland security related information.
___ Yes, dataset <u>does</u> contain national/homeland security related information.
If Yes, indicate the following:
Type of information: _____

Rationale why publication is acceptable: _____

Agency Point of Contact for this dataset: _____

Proceed to **Part B** unless the data contain only factual information that does not pertain to persons, businesses, or specific entities (for example, data is about weather, earthquakes, geographical features, etc.)

Part B

Privacy/Confidentiality Checklist and Guidance

This Checklist is for reviewing data files for risks of disclosing information that needs to be protected because of Federal privacy or data confidentiality policy. This Checklist should be completed prior to submitting the information for public release through Data.gov to ensure that the necessary steps are taken to safeguard privacy and confidentiality. For additional information refer to the complete Checklist on Disclosure Potential of Proposed Data Releases and the *Report on Statistical Disclosure Limitation* (Statistical Working Paper No. 22).

This Checklist is for the general reader and is not a substitute for having privacy/confidentiality experts review the data for disclosures prior to the data's release. Use of this checklist provides guidance for following best practices for safeguarding data privacy and confidentiality when publicly releasing information. Agencies are expected to follow their existing disclosure review practices based on statutory, regulatory and policy requirements.

Formal Disclosure Review

1) Did the dataset undergo any formal, comprehensive privacy/disclosure review?

   _____ **Yes**

   _____ **No**—*Guidance*: A dataset should be reviewed for risk of disclosing identifiable information about persons or businesses. Some agencies use formal privacy committees (e.g.: disclosure review boards) to review microdata files and tabular data to identify risks and suggest solutions to reduce any such risks. In addition, the capability to match a dataset to external data increases the risk of identifying a person or business.

If the answer is Yes, specify or attach a brief description of the disclosure review program and STOP HERE.

_____

_____

_____

_____

Data Privacy/Confidentiality Requirements

2) Were the data collected with a promise of privacy/confidentiality? That is, were the respondents told that their identity or information would be protected in public releases? This information may be found in the OMB package, Institutional Review Board application, cover letter, questionnaire, script, or other materials.

_____ **Yes**—*Please specify legal authority:*

_____

_____

_____

_____

_____ **No**

3) Are there any exemptions under FOIA that would allow your agency to withhold information on any persons or businesses listed in the file?

_____ **Yes**---*Guidance*: List the exemption your agency relies upon for withholding information contained in this dataset from a requestor.

_____

_____

_____ **No**

If the answer to question 2 is No AND the answer to question 3 is No, then STOP HERE. Otherwise, proceed.

Microdata File Section

*If the dataset is a **microdata file**,*[5] *then complete this section. If it is not a microdata file, go to the Tabular Data Section on page 10.*

Information in a microdata file may reveal a respondent's identity *directly*, by including their name or some other identifying particular exclusively associated with them (address, social security number, telephone number) or *indirectly*, by containing details about them which, in combination, describe them exclusively (zip code + date of birth + marital status + occupation). Indirect re-identification can be difficult to protect against. Appropriate safeguards require detailed knowledge of the microdata file and existing public information that may be used to disclose or confirm an individual's identity.

[5] A **microdata file** consists of a series of records at the respondent level. Each record contains values of variables for a person, household, establishment, company, facility or other reporting unit.

Direct Identifier – information that exclusively identifies a person or business. Examples include name, social security number, birth date, account or other identification numbers, street address, e-mail address, telephone or fax number, and certificate or license numbers. See Appendix A for a more comprehensive list of direct identifiers.

4) Does the file have direct identifiers?

   _____ **Yes**—*Guidance*: All direct identifiers must be deleted prior to public release of the data. Personal contact information is also considered to be a direct identifier and must be deleted from the file.

   _____ **No**

Indirect Identifier – information that, when used in combination with other data, could lead to the identification of a person or business. Geographic and demographic information can be indirect identifiers. For example, a person may be identified by combining details such as race/ethnicity + gender + zip code + marital status + occupation. A business may be identified by combining details such as State + sales type + industry code + product code. See Appendix B for examples of indirect identifiers.

5) Does each record in the file have a unique identification (ID) variable?

   _____ **Yes**—*Guidance*: Assess how the record ID variable is constructed. The ID variable should be randomly generated and randomly assigned to each record. If the ID variable is tied to any unique attribute of the person or business or relates to any geographic information, then it should be replaced with a randomly generated and assigned ID variable.

   _____ **No**

6) Does the file contain geographic information? Geographic information includes zip code, city, county, State, Census division, metropolitan status codes, latitude and longitude coordinates, block group, or Census tract.

   _____ **Yes**—*Guidance*: The smaller the geographic unit, the greater the risk of identifying a person or business. Review the level of geographic information and consider reducing the amount and specificity of geographic information publicly released. Remove zip code, city, and county variables or collapse into broader geographic groupings.

   _____ **No**

7) Does the file contain demographic information that describes the person or business? For ***persons***, demographic information includes date of birth, age, gender, race/ethnicity, occupation, income, marital status, mortgage/rent payments, property value, vehicle types and number, citizenship status, military status, number of children, or language spoken.

   For ***businesses***, demographic information includes type of industry, business, facility, product type, number of locations, market share information, etc.

   _____ **Yes**—*Guidance*: Demographic information may cause a person or business to be more identifiable. Certain other demographic variables such as age may need to be deleted, top- or bottom-coded[6], or recoded into broader categories to avoid sparse distributions—especially at the high and low ends of a distribution. Apply appropriate disclosure limitation methods to prevent the demographic variables from being used to identify respondents.

   _____ **No**

8) Does the file contain contextual information about the area where the person or business is located? Contextual information is descriptive information that can be used to identify a respondent or group of respondents.

   _____ **Yes**—*Guidance*: Unique combinations or very low or high concentrations of these variables could identify the location of the person or business, which increases disclosure risk. Some contextual information can be used to identify unique groups within a population. Certain variables may need to be deleted or recoded into broader categories to prevent linking to other files or reveal any identifiable attributes about a person or business.

   _____ **No**

9) If the data come from a statistical sample, does the file include analytic weight variables?

   _____ **Yes**—*Guidance*: Weights are typically the product of several components that together could be used as an identification tool. In general, the components used to create the final analytic weight should not be publicly released.

   _____ **No**

[6] A top-code is an upper limit on all published values of a variable. Any values greater than this upper limit are replaced by the upper limit. Similarly, a bottom-code is a lower limit on all published values for a variable. Different limits may be used for different quantitative variables, or for different subpopulations.

10) If the data come from a statistical sample, does the file contain variance estimation variables[7] such as sampling strata, primary and secondary sampling units?

_____ **Yes**—*Guidance*: Variance estimation variables may have been constructed based on geographic information. Assess these variables for the risk of identifying a respondent. In general, ensure that the published variance estimation variables do not link directly to geography or other variables used to stratify or select the sample. Release only those variance estimation variables that are necessary for data analysis.

_____ **No**

11) Which of the following disclosure limitation methods were applied to variables in the file to be released?

_____ None

_____ Suppression of the values for a variable

_____ Top or bottom coding[8]

_____ Data swapping[9]

_____ Collapsing categories

_____ Data blurring[10]

_____ Other (*specify or attach*): _____

_____

_____

[7] Variables that group the data based on the stratification in the sample design.

[8] Top or bottom coding should be limited to a small percent of records based on a frequency distribution of the records.

[9] Data swapping should be limited to a small percent of records and on a set of predetermined variables that are used as swapping variables. The values of the swapping variables are then switched between records.

[10] Blurring involves aggregating values across small sets of respondents for selected variables and replacing a reported value (or values) by the aggregate value from that set of respondents.

Tabular Data Section

*If some data are in **tabular format**, complete this section.*

12) Was a formula or measurement used to identify any table cells that could disclose identifiable information of a respondent?

_____ **Yes**

_____ **No**—*Guidance*: Disclosure limitation methods are applied to cells for which a **linear sensitivity measure** indicates that some respondent's data may be closely estimated. Some sensitivity measures are: the threshold rule, the dominance rule, and p-percent rule. See the Report on *Report on Statistical Disclosure Limitation* (Statistical Working Paper No. 22), for further details. Apply a sensitivity measure to test the table cells to assess the risk of disclosing identifiable information from the published statistical aggregate values. Application of a disclosure limitation method to the table(s) reduces disclosure risk (for examples, see question 14).

13) Was a disclosure limitation method applied to sensitive cells in the table or set of tables? A sensitive cell is a value in a table cell that permits a user to closely estimate the contribution of an individual person or business. Sensitive cells can be identified by applying a sensitivity measure to the table cell value.

_____ **Yes**

_____ **No**—*Guidance*: Applying a disclosure limitation method to the table(s) will reduce disclosure risk.

14) Which of the following disclosure limitation methods were applied to the table cells in the file to be released?

_____ None

_____ Cell Suppression

_____ Rounding

_____ Collapsing rows and/or columns

_____ Other (*specify or attach*): _____

_____

Additional comments supporting publication of dataset:

_____

Agency Point of Contact for this dataset:

_____

APPENDIX A

Examples of Direct Identifiers for Persons such as Individuals or Businesses

These are examples of direct identifiers. This list is not comprehensive because definitions of direct or indirect identifiers vary across different legal authorities.

- Name (individual, parents, children, firms, subsidiaries)
- Mailing Address
- Telephone or FAX number
- E-mail address
- Web Universal Resource Locator (URL)
- Social Security Number (in any form)
- Employment/Federal Identification Number (EIN or FIN)
- Identification and serial numbers such as credit and debit cards, account numbers, driver's license number, license plates, device identifiers, vehicle identifiers
- Certificate/License or Business License number
- Dates of birth or death
- Biometric records such as fingerprints, DNA, iris or retina scan, facial recognition or photographic images
- Photographic images of business locations and facilities

APPENDIX B

Examples of Indirect Identifiers for Persons such as Individuals or Businesses

The list shows examples of common indirect identifiers. Any variable that may be used in combination with other variables in a file or across files to identify a respondent may be an indirect identifier.

- Dates, including date of—
  - admission, treatment, discharge or other health-related events
  - opening, closing, or sale of a business
  - other notable events such as enrollment in schools, accidents, crimes, etc.
- Locations such as—
  - place of birth
  - location of company or facility
- Demographic characteristics such as—
  - gender, age, race, ethnicity, or cultural heritage
  - marital status, number or age of children, household composition
  - income, education, or occupation
- Descriptive characteristics such as—
  - type of business, NAICS[11] code or product code; sales type
  - type of parent, enterprise, or subsidiaries or their NAICS or industry product codes
  - number of employees, establishments, or subsidiaries
  - annual payroll, sales, revenue, or market share
- Physical characteristics of individuals such as—
  - medical history information
  - medical conditions
  - height, weight, or build
  - skin, hair or eye color
- Physical characteristics of facilities such as—
  - size or capacity
  - number of floors, housing units, or buildings

[11] NAICS-North American Industrial Classification System, the current system of coding industry operations. NAICS replaced the Standard Industry Classification (SIC) in 1987.

- Geographic information for residence, event, business operation, etc., such as—
    - city, county, or state
    - zip code
- Contextual area information for residence, event, business operation, etc., such as—
    - number of heating or cooling degree days
    - measures of air quality
    - proximity to a major facility such as an airport, jail, university, hospital, power plant
    - natural disasters, catastrophes, accidents, or other distinguishing events
    - statistics on crime, health, unemployment, poverty, public assistance, foreign population, etc.
    - government expenditures