DataSF: Resources

Open Data Release Toolkit

This work is licensed under a Creative Commons Attribution-ShareAlike 4.0 International License. Primary Author: Erica Finkle. Version 1.2

The purpose of this toolkit is to provide guidance and a consistent process for Office of the Chief Data Officer (OCDO) staff and department management and Data Coordinators in the City and County of San Francisco (CCSF) to release sensitive or protected datasets on the open data portal.

Table of Contents

- Background
- Process Overview
  - Deciding When and How to Release De-Identified Data
  - Documenting the Decision-Making Process
- Open Data Release Form
- Step 1: Identify Sensitive or Protected Datasets
  - Step 2: Identifiability Risk Assessment
    - Step 2A: Assess the Value of Publication
    - Step 2B: Assess the Risk of Publication
      1) Impact = individual expectation of privacy X repercussions (legal, financial, etc.)
        a) Individual expectation of privacy
        b) Repercussions
        c) Impact
      2) Risk rating = impact X likelihood of re-identification attempt
        a) Impact
        b) Likelihood of a re-identification attempt
        c) Risk rating
    - Step 2C: Weigh the Value of Publication Against the Risk of Publication
- Step 3: Privacy Solutions
  - Step 3A. Should the Dataset be Completely Closed?
  - Step 3B. Choose Identifiability Level
  - Step 3C: Choose De-Identification Methods
- Step 4: Accessibility Risk Assessment
  - Step 4A: Assess Likelihood of Successful Re-Identification
  - Step 4B: Is the De-Identified Dataset Still Useful?
  - Step 4C: Accessibility Risk Rating
  - Step 4D: Should the De-Identified Dataset be Published?
- Appendix A: Acknowledgements & Resources
- Appendix B: Purposes of De-Identifying Datasets for Release
- Appendix C: Identifiability Spectrum Model
- Appendix D: De-Identification Methods
- Appendix E: Special Considerations for Location Data
- Appendix F: Motivated Intruder Test
Background

Purpose of This Toolkit

The purpose of this Toolkit is to provide guidance and a consistent process for Office of the Chief Data Officer (OCDO) staff, department management and Data Coordinators in the City and County of San Francisco (CCSF) to release sensitive or protected datasets on the SF Open Data portal. Datasets published on the portal must be designated by departments as falling within one of three data classifications - public, sensitive or protected. Sensitive or protected data is generally data that is protected by privacy law or regulation, identifies individuals, could be misused to target individuals and/or poses other concerns.

Challenge: Following the Letter of the Law Is Not Enough

Privacy laws and regulations in various sectors and at the federal, state and local level help us to identify sensitive or protected categories of data. While privacy laws state how to protect data, most of them were written well before modern computing and statistical tools. These tools challenge the idea of simple de-identification of data. For example, you could publish record level data on probationers and still be within the law. But by combining that data with other sources or just doing some crosstabs, you might be able to identify the individuals in the dataset and associated sensitive or personal information about them. As a result, open data programs need a much more robust way of managing the release of sensitive or protected datasets.

Responsible Risk Management

We seek to publish data responsibly. This requires a balancing of competing factors, such as:

- the value of publishing the data,
- an individual's expectation of privacy,
- repercussions to an individual or the organization from re-identification, and
- the likelihood of re-identification.

Recognizing the challenges of de-identification, our focus is on practical tools to minimize risk. The Toolkit will guide you through a step-by-step process to:

1. Identify sensitive or protected raw data,
2. Perform a risk assessment regarding the identifiability of the data,
3. Choose and implement privacy solutions (e.g. de-identification methods), and
4. Perform a risk assessment regarding the accessibility of the de-identified data.

The Open Data Release Form contained in this Toolkit is also meant to serve as a way to facilitate and document the decision-making process.
Remember - Risk Can Be Managed, but Will Not Be Zero

We can never have absolute certainty (i.e. zero risk) that successful re-identification will not occur. Techniques for re-identification continually evolve, often faster than security measures intended to protect data. While re-identification risk will never be 0%, it can be reduced and managed. Acknowledging the supposed failures of “perfect anonymization”, the Open Data Release Toolkit moves forward by setting out a decision-making process to publish valuable information while managing risk of re-identification and sensitive attribute disclosure in a thoughtful way.

Scope of This Toolkit

This Toolkit applies to privacy risk in the specific context of publishing data for a government’s open data program. It does not address the entire data lifecycle. For instance, it does not cover data collection and data retention processes at the department or agency level. Rather, it ‘zooms in’ on risk management tools at the data release stage of the data lifecycle. This Toolkit also does not apply to public datasets, which are public records that can be disseminated without any concerns.

Version Notes

We’ll update this Toolkit as we learn more and receive feedback.

<table>
<thead>
<tr>
<th>Date</th>
<th>Version</th>
<th>Description of changes</th>
</tr>
</thead>
<tbody>
<tr>
<td>August 23, 2016</td>
<td>1.0</td>
<td>Creation of Toolkit</td>
</tr>
<tr>
<td>September 23, 2016</td>
<td>1.1</td>
<td>Re-order, re-title and link certain text among sections to improve flow of instructions and narrative; new text to provide additional context/examples; add acknowledgements &amp; resources; formatting</td>
</tr>
<tr>
<td>November 3, 2016</td>
<td>1.2</td>
<td>Addition of De-Identification Protocol in Step 3C “Choose De-Identification Methods”</td>
</tr>
</tbody>
</table>
Process Overview
Deciding When and How to Release De-Identified Data

The Toolkit will guide you through a step-by-step process to identify which datasets are sensitive or protected, perform a risk assessment regarding the identifiability of that data, choose and implement privacy solutions (e.g. de-identification methods), and perform a risk assessment regarding the accessibility of the de-identified data. The chart below illustrates this process:
Documenting the Decision-Making Process

The following Open Data Release Form should be used to document your decision-making. Refer to the remainder of this Toolkit for step-by-step guidance to fill out this form.

### Open Data Release Form

<table>
<thead>
<tr>
<th>Basic Information</th>
</tr>
</thead>
<tbody>
<tr>
<td>Department</td>
</tr>
<tr>
<td>Department contact</td>
</tr>
<tr>
<td>Contact details</td>
</tr>
<tr>
<td>Date</td>
</tr>
</tbody>
</table>

**Step 1: Identify sensitive or protected datasets**

1A. Dataset

1B. Relevant fields

**Step 2: Identifiability Risk Assessment**

2A. Value of publication

<table>
<thead>
<tr>
<th>Value of publication</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Moderate</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>

2B-1. Risk of publication - impact

(a) **Individual Expectation of Privacy**

<table>
<thead>
<tr>
<th>Expectation of Privacy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
</tr>
<tr>
<td>Moderate</td>
</tr>
<tr>
<td>High</td>
</tr>
</tbody>
</table>

(b) **Repercussions**

<table>
<thead>
<tr>
<th>Repercussions</th>
</tr>
</thead>
<tbody>
<tr>
<td>No discernable</td>
</tr>
<tr>
<td>Minor</td>
</tr>
<tr>
<td>Moderate</td>
</tr>
<tr>
<td>Major</td>
</tr>
</tbody>
</table>

(c) **Impact = individual expectation of privacy \times repercussions (legal, financial, etc.)**

<table>
<thead>
<tr>
<th>Impact Level</th>
<th>Repercussions</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>Very low</td>
</tr>
<tr>
<td>Moderate</td>
<td>Low</td>
</tr>
<tr>
<td>High</td>
<td>Very low</td>
</tr>
<tr>
<td>Low</td>
<td>Moderate</td>
</tr>
<tr>
<td>Moderate</td>
<td>Significant</td>
</tr>
<tr>
<td>High</td>
<td>High</td>
</tr>
</tbody>
</table>

DataSF.org: Open Data Release Toolkit - [Return to Top](#)
2B-2. Risk of publication - risk rating

(a) Impact: See 2b-1(c) above

(b) Likelihood of re-identification attempt
- Rare
- Unlikely
- Possible
- Probable

(c) Risk rating = impact X likelihood of re-identification attempt

<table>
<thead>
<tr>
<th>Risk Rating</th>
<th>Very low</th>
<th>Low</th>
<th>Moderate</th>
<th>Significant</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td>Rare</td>
<td>Very low</td>
<td>Very low</td>
<td>Moderate</td>
<td>Significant</td>
</tr>
<tr>
<td></td>
<td>Unlikely</td>
<td>Very low</td>
<td>Low</td>
<td>Significant</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Possible</td>
<td>Very low</td>
<td>Moderate</td>
<td>Significant</td>
<td>High</td>
</tr>
<tr>
<td></td>
<td>Probable</td>
<td>Very low</td>
<td>Significant</td>
<td>High</td>
<td>Extreme</td>
</tr>
</tbody>
</table>

2C. Weigh the value of publication against the risk of publication

<table>
<thead>
<tr>
<th>Value v. Risk</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Risk Rating</td>
<td>Very low risk</td>
<td>Low risk</td>
<td>Moderate risk</td>
</tr>
<tr>
<td></td>
<td>Significant risk</td>
<td>High risk</td>
<td>Extreme risk</td>
</tr>
</tbody>
</table>

- Moderate – high value. Very low – low risk
- Low – high value. Very low – moderate risk
- Low – high value. Low – significant risk
- Low – high value. Moderate – high risk
- Low – high value. Significant – extreme risk
- Low – moderate value. High – extreme risk

Step 3: Privacy Solutions

3A. Should the dataset be completely closed?

Given the result of Step 2C, should the dataset be completely closed?
- No
- Yes
  If “yes”, do not proceed.

3B. Identifiability spectrum level

If the answer to Step 3A above is “no”, then choose an identifiability spectrum level based on the results in Step 2C:
- Level 1: Readily identifiable data
- Level 2: Masked data
- Level 3: Obscured data
- Level 4: Aggregate data

3C. De-identification methods
### Step 4: Accessibility Risk Assessment

| 4A. Assess likelihood of successful re-identification | ❐ Rare  
| ❐ Unlikely  
| ❐ Possible  
| ❐ Probable |

| 4B. Is the de-identified dataset still useful? | ❐ None  
| ❐ Low  
| ❐ Medium  
| ❐ High |

| 4C. Accessibility risk rating | ❐ Very low  
| ❐ Low  
| ❐ Moderate  
| ❐ Significant  
| ❐ High  
| ❐ Extreme |

| 4D. Should the de-identified dataset be published? | ❐ Open  
| ❐ Limited Access  
| ❐ Closed |

### Planning

We plan to revisit the decisions in this form every...

- 6 months
- 1 year
- Other __________________________

### Notes
Step 1: Identify Sensitive or Protected Datasets

Does the Dataset Contain Sensitive or Protected Information?

Datasets published on the SF Open Data portal must be designated by Departments as falling within one of three data classifications - Public, Sensitive or Protected:

1. **Public**: This data could be publicly disseminated without any concerns
2. **Sensitive**: In its raw form, this data poses security concerns, could be misused to target individuals or poses other concerns
3. **Protected**: This data is protected by law or regulation and can only be shared or access internally and per organizational procedures; OR this information includes individually identified information

The Department should consult with its Deputy City Attorney to understand any applicable laws or regulations. To the extent there are any questions or concerns about the appropriate data classification, OCDO will work with the relevant Department and its legal counsel to resolve.

In some cases, an entire dataset may be clearly public or clearly private, but privacy risks often occur within just a small set of fields within a dataset or from a few unique and unpredictable entries. Even in the absence of specific legal prohibitions on disclosing personally identifiable information (PII), outlier conditions or rare events could lead to identification of individuals or sensitive attributes about individuals. For example, identifying that an arrestee is a minor of a certain age and ethnicity in a certain zip code, without providing any other information, might nonetheless serve to identify that particular individual. This is because age, ethnicity and zip code are each a “quasi-identifier” (see definition below) that, when combined, identify the individual.

The following are broad categories of data to watch out for:

- **Unique identifiers**: These are attributes that clearly identify individuals in the datasets, such as name, social security number, employee ID, etc.
- **Quasi identifiers**: These are attributes that are not alone a unique identifier, but when taken together with other data can potentially identify individuals in the dataset. Examples include birth date, ZIP code, gender, race or ethnicity.
- **Sensitive attributes**: Sensitive attributes, such as health conditions or financial information, should not be linkable to personal identities of individuals.

A more context-specific checklist of types of data to look for is summarized below:

- Obvious PII (e.g. name, SSN, birthday, phone number)
- Location data ([Appendix E](#) describes special considerations about location data)
- Specific entries within a given field that may be particularly sensitive (e.g. part of a 911 call)

---

● Unstructured or free-form fields (e.g. 311 requests from citizens)
● References to individuals (e.g. ensure no systematic connection between original PII and anonymous IDs used to replace the PII)
● Fields that are present in public, non-anonymized data (e.g. voter records, property assessments) may give rise to concerns about quasi-identifiers.

Step 2: Identifiability Risk Assessment

Risk assessments are not required to publish a protected or sensitive dataset on SF Open Data. However, it serves several important purposes, such as (i) making it easier for the public to understand how and why their information is being used, and (ii) re-assuring individuals and the public at large that CCSF is seeking to set and highlight best practices where a universally accepted standard does not exist. The risk assessment process also improves how CCSF uses information that may impact individual privacy.

The identifiability risk assessment should incorporate the following steps:

2A. Assess the value of publication
2B. Assess the risk of publication:
   1) Impact = individual expectation of privacy X repercussions (e.g. legal, financial, etc.)
   2) Risk rating = impact X likelihood of re-identification attempt
2C. Weigh the value of publication against the risk of publication

Step 2A: Assess the Value of Publication

There are a number of reasons, both practical and philosophical, why publishing data can be of value to your department and the people it serves. See Appendix B for examples of both the broad and specific purposes served by de-identifying datasets for release.

Departments already must designate the value of datasets to be published on the SF Open Data portal as low, medium, or high value. The definition of each is in the “Full Dataset” row in the table below. The assigned value for each dataset on the SF Open Data portal can be found in the Dataset Inventory.

Also consider the value each sensitive or protected field contributes toward potential use cases of the data. The final row in the table below defines low, medium, and high value in this context.

Taking into consideration the value of the full dataset and any sensitive or protected data fields that are key to creating value, assign a level (low, medium, high) to the value of publication.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full Dataset</td>
<td>This data has unclear value for either the public or other Departments; your department has never received requests for this data, nor heard of</td>
<td>This data may be useful for other departments or for people external to the city; your department has occasionally received requests for this data, or</td>
<td>Existing and ongoing requests for this data; this data addresses pressing information needs or pain points (within or without the city); or your</td>
</tr>
</tbody>
</table>

DataSF.org: Open Data Release Toolkit - Return to Top
Step 2B: Assess the Risk of Publication

1) Impact = individual expectation of privacy X repercussions (legal, financial, etc.)

a) Individual expectation of privacy

Each individual's expectation of privacy for his or her own records will vary. While it is impracticable to know this for each individual, you should consider what are the expectations of a reasonably informed person.

Factors that influence an individual's expectation of privacy include:

- Is the data about a person's private life or more public matters, such as working life?
- What is the context in which the data was collected? What is the context in which it is being shared? How do they differ, if at all?
- What notice (if any) about protection or disclosure of data has been provided to the individual?
- More broadly, what is likely to be a reasonable individual's expectation about the privacy of this information?

<table>
<thead>
<tr>
<th>Factor</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual's expectation of privacy</td>
<td>e.g. data is about public matters; data is being disclosed for a context the same as that for which it was collected; notice of disclosure was given at the time of collection</td>
<td>e.g. data could be considered public or private; data is being disclosed for a context similar or tangential to that for which it was collected; no notice was given at time of collection</td>
<td>e.g. data is about a person's private life; data is being disclosed for a context unrelated to that for which it was collected; notice of protection of data (i.e. no disclosure) was given at time of collection</td>
</tr>
</tbody>
</table>

b) Repercussions

The table below lists several types of repercussions if data is successfully re-identified. Both the individual and the organization can be impacted by a privacy harm. For many of the factors below, judgements must be made on a case-by-case basis taking into considering the particular context of the dataset.

Consider each type of repercussion, then make a determination as to a general level of repercussions (i.e. across all types).
Individual
Individual repercussions should be assessed by considering an average person in a given context. The special circumstances of an individual need not be taken into account unless they are common among a class of individuals in the particular context.

- Types of repercussions:
  - Physical or emotional (e.g. physical harm, property crime, distress, anxiety)
  - Reputational (e.g. embarrassment, shaming)
  - Financial (e.g. opportunity loss, fraud, identity theft)

Organization
Organizational repercussions should be assessed only where the relevant data, if released, could breach applicable privacy laws and regulations. These repercussions are **not** meant to apply where data is a public record.

- Types of repercussions:
  - Operational or compliance (e.g. sanctions under privacy laws or regulations, reveal non-public information about agency policies or practices)
  - Reputational (e.g. negative press or public backlash relating to breach of privacy laws or regulations)
  - Financial (e.g. fines, penalties, claims for compensation)

Level of repercussions:
1. No discernible (visible, recognized)
2. Minor
3. Moderate
4. Major

<table>
<thead>
<tr>
<th>Type</th>
<th>No discernible</th>
<th>Minor</th>
<th>Moderate</th>
<th>Major</th>
</tr>
</thead>
<tbody>
<tr>
<td>Individual - physical or emotional</td>
<td>No physical harm or emotional distress/ anxiety</td>
<td>Minor physical harm or emotional distress/ anxiety</td>
<td>Moderate physical harm or emotional distress/ anxiety</td>
<td>Severe physical harm or emotional distress/ anxiety</td>
</tr>
<tr>
<td>Individual - reputational</td>
<td>No embarrassment</td>
<td>Minor embarrassment</td>
<td>Moderate embarrassment</td>
<td>Severe shaming/ embarrassment</td>
</tr>
<tr>
<td>Individual - financial</td>
<td>No financial impact</td>
<td>Minor financial impact</td>
<td>Moderate financial impact</td>
<td>Major financial impact</td>
</tr>
<tr>
<td>Org – operational or compliance</td>
<td>No repercussions</td>
<td>Slightly increased oversight or operational restrictions; minor sanctions</td>
<td>Some increased oversight; moderate operational restrictions or sanctions</td>
<td>Significant increased oversight, operational restrictions, sanctions</td>
</tr>
<tr>
<td>Org - reputational</td>
<td>No repercussions</td>
<td>A few complaints from citizens</td>
<td>Some bad press; public protest and/or citizen complaints</td>
<td>Significant bad press; severe public backlash</td>
</tr>
<tr>
<td>Org - financial</td>
<td>No financial impact</td>
<td>Minor financial impact</td>
<td>Moderate financial impact</td>
<td>Major financial impact</td>
</tr>
</tbody>
</table>
c) Impact
What would be the impact of a successful re-identification attempt? To assign an impact level, use the individual's expectation of privacy and the general level of repercussions as calculated above.

<table>
<thead>
<tr>
<th>Impact Level</th>
<th>Repercussions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No discernable</td>
</tr>
<tr>
<td>Low</td>
<td>Very low</td>
</tr>
<tr>
<td>Moderate</td>
<td>Very low</td>
</tr>
<tr>
<td>High</td>
<td>Very low</td>
</tr>
</tbody>
</table>

2) Risk rating = impact X likelihood of re-identification attempt

a) Impact
What would be the impact of a successful re-identification attempt? Use the impact level assigned in 2B-1(c) above.

b) Likelihood of a re-identification attempt
There are two factors relating to re-identification: (1) the likelihood someone will attempt re-identification, and (2) the likelihood of re-identification being successful. The first factor should be considered here. The second factor will be considered in Step 4A, after you implement de-identification methods.

The likelihood someone will attempt re-identification can vary widely based on differing assumptions and hypotheses made about external actors. Therefore, we recommend the following practices:
- Seek input from interested stakeholders in your department (and externally where feasible) on the potential motives listed below and any others
- Come to a consensus decision about the level of likelihood below
- Revisit your decision on the level of likelihood on a regular basis to continuously assess the risk in light of new information

Assess whether someone may attempt re-identification by considering potential actors and motives or reasons for re-identification, including:
- Potential actors:
  - General public — anyone who has access to the data.
  - Expert — a computer scientist skilled in re-identification.
  - Insider — a member of an organization that collected or holds the data and has more background information than the general public.
  - Information broker — an organization that systematically collects both identified and de-identified information to re-identify.
  - Family, Friends, Acquaintances — a person who has specific information
- Potential motives or reasons:
  - Finding out personal data about someone else for nefarious reasons
  - Finding out personal data about someone else for financial gain
○ Causing mischief by embarrassing others
○ Revealing newsworthy information about public figures
○ Political or activist purposes (e.g. a campaign against a particular organization or person)
○ Curiosity (e.g. to know who was involved in an incident on a crime map)
○ Use by the private sector, in combination with a company’s own data, for marketing, data analysis, etc.
○ Certain uses of the data by your agency may create a motivation for attackers to attempt re-identification
○ Members of vulnerable populations (e.g. children, disabled, those in the criminal justice system) may be more susceptible to re-identification
○ Datasets that are seemingly innocuous may be used simply as a link to help re-identify another dataset that is sensitive.

<table>
<thead>
<tr>
<th>Factor</th>
<th>Rare</th>
<th>Unlikely</th>
<th>Possible</th>
<th>Probable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood of re-ID attempt</td>
<td>Would not expect this to happen</td>
<td>Could happen but few obvious motives</td>
<td>There are potential motives</td>
<td>Clear motives exist; has happened elsewhere (e.g. other jurisdictions)</td>
</tr>
</tbody>
</table>

**c) Risk rating**

To assign a risk rating, weigh the impact and the likelihood of a re-identification attempt.

<table>
<thead>
<tr>
<th>Risk Rating</th>
<th>Impact Level</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Very low</td>
</tr>
<tr>
<td>Likelihood</td>
<td>Very low</td>
</tr>
<tr>
<td>Rare</td>
<td>Very low</td>
</tr>
<tr>
<td>Unlikely</td>
<td>Very low</td>
</tr>
<tr>
<td>Possible</td>
<td>Very low</td>
</tr>
<tr>
<td>Probable</td>
<td>Very low</td>
</tr>
</tbody>
</table>

**Step 2C: Weigh the Value of Publication Against the Risk of Publication**

To determine whether the dataset should be completely closed and, if not, decide on the desired level of identifiability of the dataset, you should weigh the value vs. risk.

<table>
<thead>
<tr>
<th>Value v. Risk</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Low</td>
</tr>
<tr>
<td>Risk Rating</td>
<td>Very low risk</td>
</tr>
<tr>
<td></td>
<td>Low risk</td>
</tr>
<tr>
<td></td>
<td>Moderate risk</td>
</tr>
<tr>
<td></td>
<td>Significant risk</td>
</tr>
</tbody>
</table>
Step 3: Privacy Solutions

Once you performed the identifiability risk assessment, you should determine whether the dataset should be completely closed. If not, decide on the desired level of identifiability of the dataset and the de-identification methods to implement. The identifiability risk assessment serves as a tool for decision-making, rather than as a scoring mechanism that maps the factors above 1:1 with the privacy solutions. Many datasets will fall somewhere in the middle of the spectrum, and thus require careful consideration of the appropriate identifiability levels and the best de-identification methods.

Step 3A. Should the Dataset be Completely Closed?

A rule of thumb is that the dataset should be completely closed if the result of weighing benefit versus risk in Step 2C is

Step 3B. Choose Identifiability Level

If the dataset should not be completely closed, then you should choose the appropriate identifiability level from the table below given the weighting of risk and value in Step 2C above. These levels are based on an identifiability spectrum model. The discrete levels characterize specific stages that a dataset would go through as it is increasingly de-identified. See Appendix C for more information.

<table>
<thead>
<tr>
<th>Level</th>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Level 1</td>
<td>Readily identifiable data</td>
<td>Individual level data that is clearly identifiable. For example, includes names, dates of birth, etc.</td>
</tr>
<tr>
<td>Level 2</td>
<td>Masked data</td>
<td>Data where “identifying variables” (e.g. name, ID number) are redacted/ masked/ transformed; for example, through randomization and creating reversible or irreversible pseudonyms.</td>
</tr>
<tr>
<td>Level 3</td>
<td>Obscured data</td>
<td>Data where “identifying” AND “quasi-identifying” (e.g. gender, race, ethnicity, age) variables are redacted/ masked/ transformed; for example, banding/generalizing the categories that make up race or ethnicity.</td>
</tr>
<tr>
<td>Level 4</td>
<td>Aggregate data</td>
<td>Data that is aggregated; for example, through aggregation methods, non-stratified counts, frequencies, query systems, etc.*</td>
</tr>
</tbody>
</table>

*Note that individuals or sensitive attributes can be identified even from aggregate data via data linkages or if the cell size for a given crossing of some combination of variables can lead someone to identify a particular individual. For example, the sample size of a variable – say race – is small enough to deduce who that individual might be with or without additional data. The minimum cell size (meaning no results will be released for any cell of a table with a number smaller than “X”) will vary. Consult with OCDO to help you define it.
Step 3C: Choose De-Identification Methods

Primary Methods

Different kinds of data require different kinds of de-identification techniques. The appropriate de-identification method will depend upon the protected or sensitive data fields specific to your dataset.

For most datasets, we recommend following Khaled El Emam's De-Identification Protocol for Open Data. According to this protocol, you should:

1. **Classify Variables.** Determine which variables in the dataset are direct identifiers, quasi-identifiers or sensitive attributes, and which variables are not to be considered. For definitions of each of these classifications, see above Step 1: “Identify Sensitive or Protected Datasets”.

2. **Remove or pseudonymize direct identifiers.** If a direct identifier is not going to be used to link records then it should be removed. Pseudonymization techniques include one-way hashing and encryption.

3. **K-Anonymize quasi-identifiers or sensitive attributes.** K-anonymity is where each cell is aggregated such that at least k individuals exhibit each feature within the data. Khaled El Emam recommends k>=11 to de-identify public data. The indirect identifiers or sensitive attributes can be generalized, truncated, or redacted to achieve k-anonymity.
   a. **Caution:** For datasets with more than six to eight indirect identifiers or sensitive attributes, including for longitudinal data, either publish aggregated data (rather than individual level) or allow only limited access to the dataset (rather than open data portal release).

**Appendix D** contains descriptions of several de-identification methods used to protect against re-identification risk and examples of how each method is used. We also recommend consulting the draft NIST Special Publication 800-188 De-Identifying Government Datasets - Section 4 (Technical Steps for Data De-Identification), Section 5 (Requirements for De-Identification Tools) and Appendix C (Specific De-Identification Tools).

Location Data and Geo-Masking

Prior to releasing location data, consult **Appendix E** for special risks to consider, guiding principles and geo-masking methods. Locations or geographic coordinates raise a special set of concerns. In some circumstances, location data – such as zip codes, supervisor district, census tract, GPS data or other map references – will constitute PII (e.g. where information about a place or property is, in effect, also information about the individual associated with it).

Geo-masking is used to provide privacy protection for individual address information while maintaining spatial resolution for mapping purposes. When location data is geo-masked, OCDO should report the specific approach used together with the dataset. Unless the presence of geo-masking is explicitly and clearly noted, those seeking to use the dataset may misinterpret artificial “hot-spots” resulting from geo-masking techniques.
Frequency of publication, dates and timestamps

Privacy risk often can be further mitigated by reducing the frequency or timeliness of publication, so that it covers more events, is harder to identify a recent case, or does not reveal additional data such as time or date of the event. For example:

- Preserve the day of the week or the month (e.g. for purposes of identifying weekly or seasonal trends), but remove other time information such as specific date or year.
- Generalizing dates to no greater specificity than a year (e.g. March 1, 2016 becomes 2016).
- Systematically adjust dates by a random amount.
- Fields that contain time stamps may require further review and subsequent aggregation.

Other data modalities to consider removing

- **Unstructured text.** This can contain direct identifiers or additional information that can serve as a quasi-identifier.
- **Photos and video.** There are various types of biometric methods for matching photos of individuals against datasets containing photos and identifiers.
- **Sequence information.** Genetic sequences and other kinds of sequence information can be identified by matching to existing databanks containing sequences matched to identities.

Step 4: Accessibility Risk Assessment

Just as the identifiability of data can fall at various points on a spectrum, so can the accessibility of data. For instance, the Open Data Institute (ODI) has developed a [Data Spectrum](#) illustrating various levels of data accessibility, which can then be grouped at the broadest level as “open,” “shared,” and “closed.”

Step 4A: Assess Likelihood of Successful Re-Identification

How to assess

The likelihood of successful re-identification depends on factors such as the de-identification methods, the availability of other linkable data and the skill of and resources available to a motivated intruder. The likelihood can vary widely simply by changing an underlying assumption or hypothesis (e.g. what a motivated intruder already knows). Also, the likelihood will increase over time as techniques improve and more data becomes available. Therefore, we recommend:

- Seek input from interested stakeholders in your department (and externally where feasible) on the motivated intruder test
- Seek to calculate a specific re-identification probability or penetration test if feasible
- Revisit your decision on the level of likelihood on a regular basis to continuously assess the risk in light of new information

Consider the ‘motivated intruder’

Conduct a ‘motivated intruder test’ from [Appendix F](#), which considers whether an intruder...
would be able to achieve re-identification if motivated to attempt this. Appendix F summarizes this test based on Anonymisation: managing data protection risk code of practice by the UK Information Commissioner’s Office (ICO). The ICO document lays out the test in more detail.

**What is the format and volume of available information? How does this relate to the format or amount of data being considered for publication?**

**Take stock of the following:**
- The smaller the location or geographic area, the more susceptible the population may be to having their identities disclosed.
- The more traits revealed about an individual, the greater the likelihood of re-identification.
- Does the data have the characteristics needed to facilitate data linkage (e.g. is the same code number used to refer to the same individual in different datasets)?
- What other ‘linkable’ information is available publicly or easily?
- What technical measures might be used to achieve re-identification?
- What is the volume of data? Some large datasets have a high degree of unicity (i.e. development of unique identifying profiles derived from traditionally de-identified data).
- How often will new data be published?

**Consider quantifying the re-identification probability or penetration testing if feasible.**
- **Re-identification probability.** Different kinds of re-identification probabilities include Known Inclusion Re-Identification Probability, Unknown Inclusion Re-Identification Probability, Recording Matching Probability and Inclusion Probability. See the NIST draft De-Identifying Government Datasets for descriptions of each probability.
- **Penetration testing.** This involves checking the de-identification of datasets by purposefully trying to find and exploit vulnerabilities. The idea is to perform penetration testing to identify and remedy any privacy concerns prior to public release of a dataset.

These processes can be very resource-intensive, and therefore cannot be performed for all datasets. However, the City may be able to conduct penetration testing on certain datasets via trusted partnerships with the academic and civic tech communities. If a penetration test has been carried out, you will want to consider any re-identification vulnerabilities it reveals.

**Assign level of likelihood**

<table>
<thead>
<tr>
<th>Factor</th>
<th>Rare</th>
<th>Unlikely</th>
<th>Possible</th>
<th>Probable</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood of re-ID attempt</td>
<td>No ‘other’ info identified; format and volume of info to be published prevent data linkage (e.g. aggregate tables)</td>
<td>Could happen but ‘other’ info not readily available; format and volume of info to be published do not raise any red flags.</td>
<td>‘Other’ info may be available to at least some individuals; format and volume of data to be published could potentially allow for re-identification</td>
<td>‘Other’ info readily available and linkable; large dataset with high degree of unicity; has happened elsewhere (e.g. other jurisdictions)</td>
</tr>
</tbody>
</table>
Step 4B: Is the De-Identified Dataset Still Useful?

Taking into consideration the value of the full dataset and any sensitive or protected data fields that have been removed, use your best judgment to assess the utility of the de-identified dataset. Consider whether the de-identification methods have introduced bias or inaccuracies into the dataset.

<table>
<thead>
<tr>
<th>Data Type</th>
<th>None</th>
<th>Low</th>
<th>Moderate</th>
<th>High</th>
</tr>
</thead>
<tbody>
<tr>
<td>Full de-identified dataset(s)</td>
<td>No beneficial use cases and no compelling use cases for the dataset</td>
<td>Few beneficial use cases for the dataset and no compelling use cases for the dataset</td>
<td>Some beneficial use cases or potentially a compelling use case for the dataset</td>
<td>Several beneficial use cases for the dataset and/or few very compelling use cases</td>
</tr>
</tbody>
</table>

Step 4C: Accessibility Risk Rating

To obtain a risk rating, combine the likelihood of successful re-identification with the utility of the de-identified data per the table below.

<table>
<thead>
<tr>
<th>Risk Rating</th>
<th>Utility Level</th>
<th>High</th>
<th>Medium</th>
<th>Low</th>
<th>None</th>
</tr>
</thead>
<tbody>
<tr>
<td>Likelihood</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Rare</td>
<td>Very low</td>
<td>Very low</td>
<td>Low</td>
<td>Moderate</td>
<td></td>
</tr>
<tr>
<td>Unlikely</td>
<td>Very low</td>
<td>Low</td>
<td>Moderate</td>
<td>Significant</td>
<td></td>
</tr>
<tr>
<td>Possible</td>
<td>Low</td>
<td>Moderate</td>
<td>Significant</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Probable</td>
<td>Moderate</td>
<td>Significant</td>
<td>High</td>
<td>Extreme</td>
<td></td>
</tr>
</tbody>
</table>

Step 4D: Should the De-Identified Dataset be Published?

Given the limited scope and context of this Toolkit, we assign dataset(s) to one of three levels of accessibility – open, limited access, or closed:

- **Open**: publish on SF Open Data portal (and elsewhere if desired)
- **Limited access**: limited access to certain parties (e.g. researchers) upon request and with their agreement that they will not seek to re-identify the data
- **Closed**: accessed only by internal staffers or where otherwise allowed by law

Generally, de-identified datasets with:

- **extreme** or **high risk** will be closed or limited access
- **significant** or **moderate** risk will be limited access or open
- **low** or **very low** risk will be open
“Limited access” is particularly appropriate for handling of de-identified data derived from sensitive source materials where there is a significant likelihood of successful re-identification. For example, limited access to data might be granted to access de-identified datasets where obvious identifiers have been removed but certain quasi-identifiers (e.g. gender, race/ethnicity) remain.

The line between “limited access” and “closed” can be blurry. **OCDO plans to make available further guidance on “limited access” at a future date.** For purposes of using this Toolkit prior to the release of such guidance, the decision between these two categories should be based on the outcome of the accessibility risk assessment.

Where a dataset will be “open”, please refer to the Publishing section of the DataSF website for publishing submission, submission guidelines, publishing guidelines, automation services and publishing plans.
Appendix A: Acknowledgements & Resources

Acknowledgements

Below are a handful of thanks - I may have missed some, if so, my apologies!

**Thought partners, colleagues and advisors.** Susan Crawford, Catherine Crump, Gabe Cunningham, Matt Daley, Ben Green, Michael Mattmiller, Yves-Alexandre de Montjoye, and Jane Wiseman.

**World's Most Amazing Team Ever.** Joy Bonaguro, Janine Heiser, and Jason Lally. Read more about them. A special thanks to Joy Bonaguro for her brilliant and game-changing thoughts, ideas, suggestions, feedback, and edits.

Resources

This Toolkit was developed with reference to a few amazing resources:

<table>
<thead>
<tr>
<th>Title</th>
<th>Attribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>A de-identification protocol for open data</td>
<td>Khaled El Emam, IAPP blog post</td>
</tr>
<tr>
<td>Anonymisation: managing data protection risk code of practice</td>
<td>UK Information Commissioner’s Office</td>
</tr>
<tr>
<td>Balancing Utility and Privacy of High-Dimensional Datasets: Mobile Phone Metadata</td>
<td>Alejandro Noriega Campero (2015) Massachusetts Institute of Technology</td>
</tr>
<tr>
<td>Data De-identification: An Overview of Basic Terms</td>
<td>Privacy Technical Assistance Center, U.S. Department of Education</td>
</tr>
<tr>
<td>Data Spectrum</td>
<td>Open Data Institute</td>
</tr>
<tr>
<td>De-Identifying Government Datasets</td>
<td>DRAFT NIST Special Publication 800-188, Simson L. Garfinkel</td>
</tr>
<tr>
<td>Title</td>
<td>Author(s)</td>
</tr>
<tr>
<td>----------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td><strong>Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification</strong></td>
<td>Gregory S. Nelson of ThotWave Technologies, Chapel Hill, NC</td>
</tr>
<tr>
<td>[Insert link to Harvard paper once finalized]</td>
<td>[Insert attribution]</td>
</tr>
</tbody>
</table>
Appendix B: Purposes of De-Identifying Datasets for Release

De-identifying datasets for release can serve broader purposes, such as:

- **Stimulate new ideas and services.** By releasing open data, city departments may help to stimulate new and innovative ideas from our local technology community. There is great potential for open data to act as the fuel for new solutions and even new businesses that can address common problems or challenges facing those that live in, work in or travel to the City. View applications that have been built using the City's data.

- **Increase internal data sharing.** Open data can also help with some of our internal challenges accessing data between departments. Right now, analysts often rely on personal relationships to access data from other departments. DataSF can provide a platform to share data internally. Combining information from different departments could provide valuable insights into how our city works and how departments may better serve those that live and work in San Francisco.

- **Simplify Sunshine Requests.** Open data releases can be an effective way of responding to requests for data through the City's Sunshine Ordinance. One open data release may address multiple requests for information than can be repetitive and costly to respond to if addressed on an individual basis.

- **Reduce unwanted web traffic.** Publishing open data can also help reduce unwanted web traffic on department websites, which is often the result of “data scraping” by individuals seeking to obtain data in bulk from the City. This puts unnecessary stress on the city's technology infrastructure and unneeded burden on city IT staff.

- **Changing how we use data.** Ultimately, open data can serve as a platform to change how we use, share and consume our data externally and internally; transform data into services, and foster continuous improvement in decision-making and the business of government.

Open data may also serve specific purposes that depend upon the agency/department or subject matter of the dataset. While it is impossible to list all of these, a few examples are:

- **Research.** Certain agencies may seek to make de-identified data public to promote research transparency and allow others to reproduce and build upon their results.

- **Criminal justice.** A criminal justice dataset may make it possible to measure and examine criminal justice outcomes over time, promote the responsible presentation and use of crime data, and meet concrete research needs.

- **Health.** Publishing health data could lead to more efficient scientific research, such as new kinds of studies previously not possible that involve the combination of multiple data sources. It can also help avoid duplicative research and discovery of complementary datasets held by other City agencies.

- **Transportation/utilities.** Open data can reduce data transaction costs, increase service efficiency and/or identify new business opportunities for services.
Appendix C: Identifiability Spectrum Model

In his paper “Practical Implications of Sharing Data: A Primer on Data Privacy, Anonymization, and De-Identification”, G. S. Nelson of ThotWave Technologies shares a model developed by Khaled El Elmam of the Children's Hospital of Eastern Ontario that considers “identifiability” of data as a spectrum, rather than as a binary option. The discrete levels characterize specific stages that a dataset would go through as it is increasingly de-identified. Note that, on this spectrum, even at the highest level, de-identification involves some level of risk because the techniques for re-identification continually evolve, often faster than security measures intended to protect data.

The following is excerpted from Nelson's paper:

- **Level 1**: Data that is clearly identifiable. For example, data includes names, Social Security Numbers (SSNs), biometrics and dates of birth or other identifying information.
- **Level 2**: Data that is masked or obscured. For example, modify the “identifying” variables through randomization and creating reversible or irreversible pseudonyms.
- **Level 3**: Masked identifiers and non-identifiers. As with Level 2, the identifiers (such as name and date of birth) are masked, but with Level 3, we also mask variables that are considered to be quasi-identifiers (e.g. gender).
- **Level 4**: Managed data. Actively manage (and measure) the degree to which re-identification can occur. If the risk is low (according to an established benchmark) then the data is considered managed with or without personal information being considered identifiable.
- **Level 5**: Aggregate data that cannot physically identify individuals. Through the use of aggregation methods, non-stratified counts, frequencies or rates are shared. Note that not all aggregated data meet this requirement if the cell size for a given crossing of some combination of variables can lead someone to identify a particular individual. For example,

---


the sample size of one particular variable – say race – is small enough to deduce who that individual might be (with or without additional data.)

Nelson also illustrates the comparable risk present in the five levels of the model described above.

- **Level 1**: All individual level data (open)
- **Level 2**: Identifier columns are redacted/hidden/transformed (de-identified); data is coded or masked (surrogate keys); data is masked irreversibly (anonymized)
- **Level 3**: Quasi-identifier columns are redacted/hidden/transformed
- **Level 4**: Data is masked and the identifiability is measurable (low risk for re-identification)
- **Level 5**: Data is aggregated and individual data cannot be re-identified
## Appendix D: De-Identification Methods

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Redaction</td>
<td>The process of expunging sensitive data from records prior to disclosure</td>
<td>All identifiers and quasi-identifiers are dropped from the dataset</td>
</tr>
<tr>
<td>Record suppression</td>
<td>Removing certain fields or categories in datasets to prevent the identification of individuals in small groups or those with unique characteristics. When the combination of quasi-identifiers presents too high a risk of re-identification to be released, often used in public health reporting, geospatial analytics or secondary use datasets. Usually requires additional suppression of non-sensitive data to ensure adequate privacy protection (e.g., complementary suppression of one or more non-sensitive cells in a table so that the values of the suppressed cells may not be calculated by subtracting the reported values from the row and column totals)</td>
<td>Drop the cells where the number of persons for any combination of zip code, age category and race is below a given threshold (e.g., 5 people)</td>
</tr>
<tr>
<td>Cell Suppression</td>
<td>Suppressing or masking the value of a single field</td>
<td>A field in a health care record containing a very rare disease</td>
</tr>
<tr>
<td>Randomization</td>
<td>Retains the direct identifiers (name, phone number), but replaces their values with simulated (random) values</td>
<td>Algorithm which randomly replaces the date of birth for persons</td>
</tr>
<tr>
<td>Shuffling</td>
<td>Data for one or more variables are switched with another record. All of the values in the dataset are real, but they are assigned to the wrong people</td>
<td>Distinct values of a variable are randomly assigned to records</td>
</tr>
<tr>
<td>Creating Pseudonyms or Surrogate</td>
<td>The creation of aliases can be done in one of two ways where a variable such as SSN is replaced with a surrogate. Refers to the unique descriptor that can be used to match individual-level records across de-identified data files from the same source (e.g., for the purposes of comparing over time). Depending on the need, this can be done so that a key can be used to restore the original value or irreversibly (anonymized). Has the advantage that it can be recreated accurately at a later point in time on a different dataset</td>
<td>Applying a one-way hash to the variable using a secret (protected) key. A hash is a function that converts a value to another value (the hash value) but you cannot reverse the hash value back to the original value</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>Description</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Sampling</td>
<td>● Taking a random sample of a dataset&lt;br&gt;● Can also be taken using stratification to ensure that the proportion of class variables are the same as the original (e.g., age groups, gender, race)</td>
<td>Randomly select a sample (e.g., 10%) based on the original dataset size</td>
</tr>
<tr>
<td>Banding/Generalization</td>
<td>● Rare quasi-identifiers can be aggregated to provide better de-identification or anonymization.</td>
<td>A low race/ethnicity can be combined with others under an “other” category.</td>
</tr>
<tr>
<td>Adding Noise</td>
<td>● Often used to introduce noise or randomness in continuous variables&lt;br&gt;● May have limited protection as there are methods used in signal processing techniques to remove the noise&lt;br&gt;● Can still re-identify if not enough noise is added relative to the density of data points, but too much noise can make the data less useful&lt;br&gt;● Can be hard to communicate noise to the public – may be seen as misleading or intentional obfuscation of the truth</td>
<td>Adjusting location according to randomly generated distances, offsetting birthdays, jittering</td>
</tr>
<tr>
<td>Character Scrambling</td>
<td>● Rearrangement of the order of the characters in a field&lt;br&gt;● This has limited value as it may be quite easy to reverse and is not a reliable way to protect information</td>
<td>For example, “SMITH” may be scrambled to “TMHIS”</td>
</tr>
<tr>
<td>Character Masking</td>
<td>● Character masking is when the nth character or characters of a string are replaced with another character&lt;br&gt;● Simple methods that only replace the first or last character has limited use as the values can be reconstructed with little effort</td>
<td>Replace SMITH with SMIT* or <em>M</em>T*</td>
</tr>
<tr>
<td>Truncation</td>
<td>● A variant of character masking in that the nth character is removed rather than replaced with a special character</td>
<td>Replace SMITH with MITH or SMIT or SITH</td>
</tr>
<tr>
<td>Encoding</td>
<td>● The value is replaced with another meaningless value&lt;br&gt;● Most effectively used when creating a surrogate value for unique values</td>
<td>Replace SMITH with X&amp;T%#</td>
</tr>
<tr>
<td>Blurring</td>
<td>● Used to reduce the precision of the data&lt;br&gt;● Convert a continuous variable into categorical data elements, aggregating data across small groups of respondents, and reporting rounded values and ranges instead of exact counts</td>
<td>Replace an individual’s actual reported value with the average group value (on more than one variable with different groupings for each variable)</td>
</tr>
<tr>
<td>Masking</td>
<td>● Used to “mask” the original values in a dataset&lt;br&gt;● The purpose of this technique is to retain the structure and functional usability of the data, while concealing information that could lead to the identification, either directly or indirectly, of an individual value</td>
<td>Replace sensitive information with realistic but fake data, or modify original data values based on predetermined masking rules (e.g., by applying a</td>
</tr>
<tr>
<td>Perturbation</td>
<td>Involve making small changes to the data to prevent identification of individuals from unique or rare population groups. Data perturbation is a data masking technique in that it is used to “mask” the original values in a dataset to avoid disclosure.</td>
<td>Swap data among individual cells to introduce uncertainty, so that the consumer of the data does not know whether the real data values correspond to certain records, and introduce “noise,” or errors (e.g., by randomly misclassifying values of a categorical variable).</td>
</tr>
<tr>
<td>---</td>
<td>---</td>
<td>---</td>
</tr>
<tr>
<td>Rounding</td>
<td>Cells are rounded to a multiple. The data user will not have information on what action was taken with that particular cell. This action will prevent the identification of a unique individual and will also prevent the use of linear methods to determine the attributes of one individual through combinations of cells and totals. E.g. cell with value of 10,000 for all people doing some activity up to the present date. The following month, the figure in that cell rises to 10,001. If an intruder compares the tables it would be easy to deduce a cell of 1. Rounding prevents this.</td>
<td>Cells rounded to the nearest multiple of three (as a result, each cell will have one added to it, one subtracted from it, or be left alone).</td>
</tr>
<tr>
<td>Tabular reporting/Aggregation</td>
<td>Produce tabular (aggregated) data. One form of aggregating data is k-anonymity, where each cell is aggregated such that at least k individuals exhibit each feature within the data. The efficacy of this approach is limited in datasets that contain many quasi identifiers / sensitive attributes or include outliers / rare cases.</td>
<td>Present a summary table instead of individual records.</td>
</tr>
<tr>
<td>Query systems</td>
<td>Users access data through online portals that return answers to specific questions. This allows the data processor to monitor and, as needed, limit amount of information revealed.</td>
<td></td>
</tr>
<tr>
<td>Differential privacy</td>
<td>Differential privacy is a formal mathematical definition of privacy, which provides a provable guarantee of privacy against a wide range of potential attacks. It is not a single tool, but rather a standard, which many tools have been devised to satisfy. Some differentially private tools utilize an interactive query-based mechanism, and others are non-interactive, i.e., enabling data or data summaries to be released and used.</td>
<td></td>
</tr>
</tbody>
</table>
Appendix E: Special Considerations for Location Data

Risks of Location Data

The risks that can emerge from the disclosure of location data are still emerging, and there is no simple rule for handling this information. Consider the following:

- The more precise a piece of geographic data, the more possible it becomes to analyze it or combine it with other information, resulting in disclosure of identifying information.
- The context of the related information being published.
- The different geographic units used – overlapping geographic units (e.g. zip code, census tract, supervisor district) create the risk of jigsaw of identification whereby the overlap among the geographic units makes it possible to identify, say, a single household.
- The less dense a population or the fewer instances in a geographic region, the more noise is required to protect privacy.
- If geo-masked data is published multiple times, each with a different set of random noise added, it will provide more details about the underlying data.
- Spatial coordinates and addresses should be thought of as the same information presented in different ways, and so should be treated with the same sensitivity. Spatial coordinates are often overlooked in practice because they are not human-readable. However, spatial coordinates can be equally revealing of location as address information.

Guiding Principles for Release of Location Data

The UK Information Commissioner’s Office has developed the following principles when considering the disclosure of geographic datasets to the public:

- Increase a mapping area to cover more properties or occupants;
- Avoid the publication of location data on a household level;
- Reduce the frequency or timeliness of publication, so that it covers more events, is harder to identify a recent case, or does not reveal additional data such as time or date of the event. Publishing data very frequently or in real-time poses a greater privacy risk;
- Remove the final ‘octet’ on IP addresses to degrade the location data they contain; and
- Use formats, such as heat maps, that provide an overview without allowing the inference of detailed information about a particular place or person.

Geo-Masking

When location data is geo-masked, OCDO should report the specific approach used together with the dataset. Unless the presence of geo-masking is explicitly and clearly noted, those seeking to use the

---


DataSF.org: Open Data Release Toolkit - Return to Top
A dataset may misinterpret artificial “hot-spots” resulting from geo-masking techniques.

<table>
<thead>
<tr>
<th>Method</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aggregation</td>
<td>• Reducing the precision of identifiers to the associated street block, census tract, zip code, etc.</td>
</tr>
</tbody>
</table>
| Spatial k-anonymity   | • Spatial k-anonymity uses algorithms to intelligently cluster data such that each cluster represents a region with at least k data points. Because the density of data varies across geographic regions, each cluster will be of a different size.  
  • Challenging to implement. Care must be taken to determine which fields to include in the aggregation. For instance, ten police entries that all report a different type of incident may not be appropriate to group together, since any information about a specific event in that neighborhood will be enough to uniquely identify the corresponding entry. |
| Donut geomasking      | • A form of random perturbation in which the distance each data point is shifted is restricted by minimum and maximum distances. This is done to ensure that the shared data is not too close to the original entry but also that the dataset still broadly represents the true conditions. |
| Triangular displacement|                                                                                                                                                                                                     |
| Random direction and fixed radius |                                                                                                                                                                                                         |
| Random perturbation within a circle |                                                                                                                                                                                                         |
| Guassian displacement |                                                                                                                                                                                                         |
| Bimodal Gaussian displacement |                                                                                                                                                                                                         |

See [https://agislab.files.wordpress.com/2014/04/patient-privacy.pdf](https://agislab.files.wordpress.com/2014/04/patient-privacy.pdf) for an illustration of each of these methods.
Appendix F: Motivated Intruder Test

This is primarily a summary of the test based on Anonymisation: managing data protection risk code of practice by the UK Information Commissioner’s Office (ICO). The ICO document lays out the test in more detail.

Who Is a ‘Motivated Intruder’?

The ‘motivated intruder’ test considers whether a motivated intruder would be able to achieve re-identification if motivated to attempt this. It assumes the person is reasonably data competent, has access to resources such as the internet, libraries and all public documents, and would employ investigative techniques such as making enquiries of people who may have additional knowledge of the identity of the data subject or advertising for anyone with information to come forward.

What ‘Other’ Information Is ‘Out There’ for a Motivated Intruder to Discover?

The other information needed to perform re-identification could be information available to certain organizations, to certain members of the public, or that is available to everyone because it has been published on the internet, for example.

It is worth making a general assessment of the risk of one individual or a group of individuals with prior knowledge using that knowledge to re-identify some individuals in a dataset. However, the privacy risk posed by those with particular personal knowledge, in all but exceptional instances, should not solely or significantly dictate risk mitigation measures. Consider how likely it is that a person or group will both have prior knowledge and seek out this type of data for nefarious purposes.

What Practices Might Be Undertaken by a Motivated Intruder? How Readily Available Is the ‘Other’ Information?

Although the general public may not be skilled in re-identification, many resources on the internet make it easy to obtain datasets, specialized tools, and even relevant experts for purposes of re-identification. Additional practices can include:

- Carrying out a web search to discover whether a combination of date of birth and zip code data can be used to reveal a particular individual's identity
- Searching newspaper archives to see whether it is possible to associate a victim's name with crime map data
- Using social networking to see if it is possible to link anonymized data to a user's profile
- Using a voter register and local library resources to try to link anonymized data to someone's identity
- A company combining public data with its own data for purposes of marketing, data
analysis, etc.