Open Data Aggregation and Suppression Guidelines

As agencies work to make more of their data available to the public, they must balance the goal of transparency with the need to protect individual privacy. While much of the data maintained by state agencies contains personally identifiable information (PII) or protected health information (PHI) and must be protected in its raw form, the data may still be valuable as open data after it has been de-identified. This document provides guidance on data aggregation and suppression practices that agencies can use to make more of their data publicly available.

> The Connecticut State Data Plan provides the following guidance on aggregating and publishing open data:
>
> - *Provide open data at the finest level of geographic and demographic granularity possible, with consideration of client/consumer data confidentiality, privacy, and deductive disclosure.*
> - *Aggregate private and sensitive data in consistent, meaningful and respectful ways, to enable policy makers to make better decisions, but protect the rights and dignity of persons for whom these data may be collected.*
>
> Connecticut State Data Plan, Principle 12, p. 8.

Open data and privacy preservation

Making government data open and accessible can have numerous benefits for civic engagement, innovation, and government effectiveness. However, it can also potentially impact the privacy of individuals whose information is collected by state agencies. Agencies must be aware of open data privacy implications and take the necessary steps to minimize risk. One of the primary open data privacy concerns is re-identification, or the discovery of an individual's personal information from a dataset that has been de-identified. Even after a dataset has been aggregated and the direct identifiers (e.g. name, social security number, etc.) have been removed, there may still be the possibility of inadvertently disclosing personal information about individuals represented in the dataset. Seemingly anonymous data can become revealing when combined with other datasets. Following the de-identification guidelines in this guide can mitigate the risks inherent in sharing private and sensitive data publicly.

- ➢ For more detailed information and guidance on how to weigh the risks and benefits of publishing government data as open data, see Harvard University's 2017 report, "Open Data Privacy: A risk-benefit, process oriented approach to sharing and protecting municipal data," available at http://nrs.harvard.edu/urn-3:HUL.InstRepos:30340010.

De-identification considerations

Most data collected by state agencies that contains personal information cannot be published in its raw form, but must be de-identified before being released as open data. There are many de-identification methods (e.g. removing fields, removing records, aggregating data, adding anonymous identifiers, etc.), and the method of de-identification will vary based on the data in question.

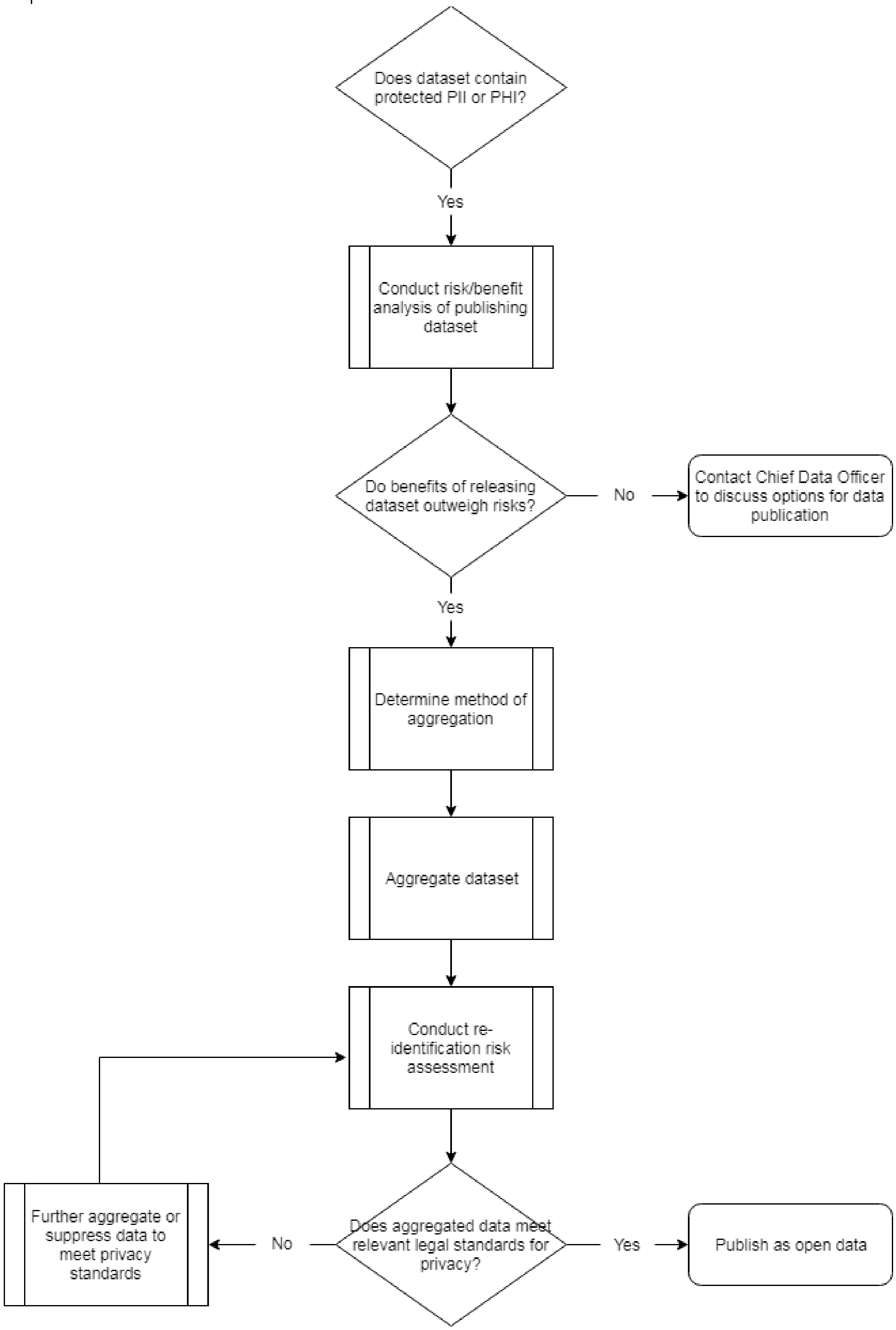
The need to protect individual privacy must be balanced with the goal of maximizing the utility of the data. Generally, datasets at the individual-level contain the greatest level of detail and are the most valuable for analysis. De-identification methods like removing fields, removing records, and creating anonymous identifiers are strategies for protecting privacy while still publishing data at the individual-level. While much of the data collected and maintained by state agencies cannot be made public at an individual-level due to privacy laws and protections, there are some exceptions, including the State Licenses and Credentials dataset provided by the Department of Consumer Protection and the Accused Pre-Trial Inmates in Correctional Facilities and Sentenced Inmates in Correctional Facilities datasets published by the Department of Correction.

Data that cannot be published at the individual level can still be published as open data if it is aggregated appropriately. The level of aggregation will vary depending on the dataset. In addition to being aggregated on a geographic level (e.g. town), data can also be aggregated across demographic (e.g. gender, race, etc.) and other relevant fields specific to a given dataset. After being aggregated, the data may still risk re-identification if the aggregated groups are too small. In these cases, some data may need to be suppressed, or a higher level of aggregation may be necessary.

When in doubt about the appropriate level of granularity at which to publish open data, contact the Chief Data Officer via email at Tyler.Kleykamp@ct.gov.

Data aggregation and suppression process

The flowchart below illustrates a suggested process for aggregating datasets containing PII or PHI to prepare for publication as open data.


Does dataset contain protected PII or PHI?
Yes
Conduct risk/benefit analysis of publishing dataset
Do benefits of releasing dataset outweigh risks?
No
Contact Chief Data Officer to discuss options for data publication
Yes
Determine method of aggregation
Aggregate dataset
Conduct re-identification risk assessment
Does aggregated data meet relevant legal standards for privacy?
No
Further aggregate or suppress data to meet privacy standards
Yes
Publish as open data

Model guidelines for data aggregation and suppression: Connecticut State Department of Education

The CSDE has a helpful document outlining its data suppression guidelines available on its website. This document discusses when to suppress data in aggregated datasets and may be a helpful starting place for agencies looking to publish private and sensitive data as open data.

The document provides the following suppression rules:

Suppression of Cell Counts:

1. *If any cell is ≤ 5 the value is suppressed (this includes a total).*
2. *If cell is ≤ 5 and only one value is suppressed in a row or column, the next highest value in that row or column is also suppressed. If there are multiple occurrences of this value, randomly suppress one occurrence. This is referred to as complementary suppression.*
3. *Totals are retained whenever possible.*
4. *Fields with a value of 0 are not suppressed.*
5. *All categories by which data are parsed (e.g., race, EL) are presented in report tables even if there are no data for categories.*

Suppression of Computed Statistics:

When cell counts are small, suppression of statistics (e.g., average, percent of total) protects confidentiality and ensures that statistics based on a very small sample size are not interpreted as equally representative as those based on a sufficiently larger sample size. Suppress a statistic if any one of the following conditions is true:

a. the count associated with the statistic has been previously suppressed
b. numerator is ≤ 5
c. denominator is < 20

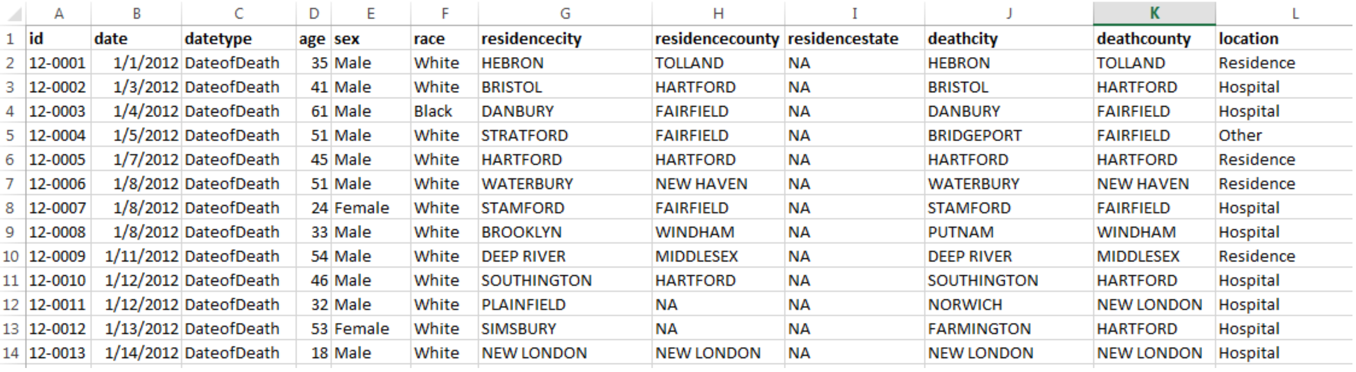
- ➢ More detail about the SDE's suppression guidelines is available in their "Data Suppression Guidelines" document, available at: http://edsight.ct.gov/relatedreports/BDCRE%20Data%20Suppression%20Rules.pdf.

Examples: Individual-level and aggregated datasets

Below are three examples of de-identified data based on the Accidental Drug Overdose Related Deaths, 2012-2018 dataset published by the Office of the Chief Medical Examiner.

Individual-level datasets

The dataset pictured below is an example of de-identified data at the individual level. Each row represents an individual death related to a drug overdose. In this case, the method of replacing individual's names with a unique identifier was sufficient for de-identifying this dataset. The name of each deceased individual has been replaced with a unique identifier in column A. In many cases, this level of de-identification would not be appropriate (e.g. in the case of education data), but, as this example shows, in some cases publishing de-identified individual-level data is possible.



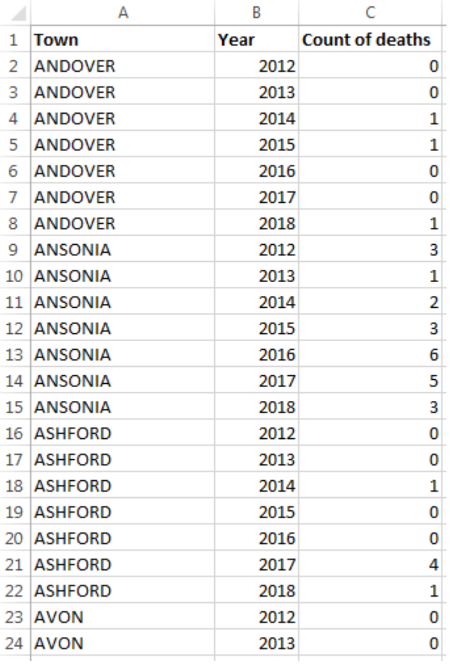
	A	B	C	D	E	F	G	H	I	J	K	L
1	**id**	**date**	**datetype**	**age**	**sex**	**race**	**residencecity**	**residencecounty**	**residencestate**	**deathcity**	**deathcounty**	**location**
2	12-0001	1/1/2012	DateofDeath	35	Male	White	HEBRON	TOLLAND	NA	HEBRON	TOLLAND	Residence
3	12-0002	1/3/2012	DateofDeath	41	Male	White	BRISTOL	HARTFORD	NA	BRISTOL	HARTFORD	Hospital
4	12-0003	1/4/2012	DateofDeath	61	Male	Black	DANBURY	FAIRFIELD	NA	DANBURY	FAIRFIELD	Hospital
5	12-0004	1/5/2012	DateofDeath	51	Male	White	STRATFORD	FAIRFIELD	NA	BRIDGEPORT	FAIRFIELD	Other
6	12-0005	1/7/2012	DateofDeath	45	Male	White	HARTFORD	HARTFORD	NA	HARTFORD	HARTFORD	Residence
7	12-0006	1/8/2012	DateofDeath	51	Male	White	WATERBURY	NEW HAVEN	NA	WATERBURY	NEW HAVEN	Residence
8	12-0007	1/8/2012	DateofDeath	24	Female	White	STAMFORD	FAIRFIELD	NA	STAMFORD	FAIRFIELD	Hospital
9	12-0008	1/8/2012	DateofDeath	33	Male	White	BROOKLYN	WINDHAM	NA	PUTNAM	WINDHAM	Hospital
10	12-0009	1/11/2012	DateofDeath	54	Male	White	DEEP RIVER	MIDDLESEX	NA	DEEP RIVER	MIDDLESEX	Residence
11	12-0010	1/12/2012	DateofDeath	46	Male	White	SOUTHINGTON	HARTFORD	NA	SOUTHINGTON	HARTFORD	Hospital
12	12-0011	1/12/2012	DateofDeath	32	Male	White	PLAINFIELD	NA	NA	NORWICH	NEW LONDON	Hospital
13	12-0012	1/13/2012	DateofDeath	53	Female	White	SIMSBURY	NA	NA	FARMINGTON	HARTFORD	Hospital
14	12-0013	1/14/2012	DateofDeath	18	Male	White	NEW LONDON	NEW LONDON	NA	NEW LONDON	NEW LONDON	Hospital

Individual-level dataset

Aggregated datasets

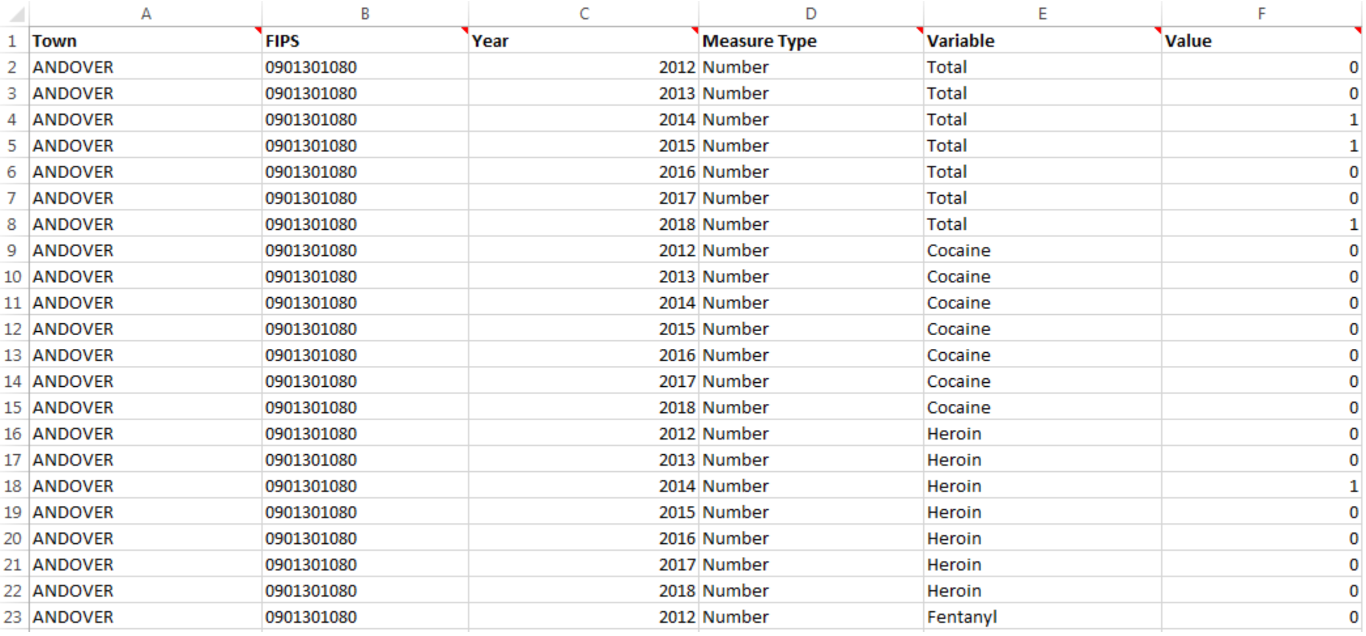
Below are two examples of how the Accidental Drug Related Deaths dataset could be aggregated to further protect the personal information of individuals represented in the dataset. In the first case, the number of deaths is shown as a count by town and year. In the second, the number of deaths is shown by town, year, and the drugs related to the death. The method of aggregation should be selected with the goal of providing meaningful information that can help policy makers make better decisions, while also protecting the privacy of individual's whose data is represented in the data.

Note: Use vertical rather than horizontal orientation when including data by year, as in these two examples. Years should have their own rows, rather than columns, in the data. This will make it easier to update each year and will facilitate data visualization.



	A	B	C
1	**Town**	**Year**	**Count of deaths**
2	ANDOVER	2012	0
3	ANDOVER	2013	0
4	ANDOVER	2014	1
5	ANDOVER	2015	1
6	ANDOVER	2016	0
7	ANDOVER	2017	0
8	ANDOVER	2018	1
9	ANSONIA	2012	3
10	ANSONIA	2013	1
11	ANSONIA	2014	2
12	ANSONIA	2015	3
13	ANSONIA	2016	6
14	ANSONIA	2017	5
15	ANSONIA	2018	3
16	ASHFORD	2012	0
17	ASHFORD	2013	0
18	ASHFORD	2014	1
19	ASHFORD	2015	0
20	ASHFORD	2016	0
21	ASHFORD	2017	4
22	ASHFORD	2018	1
23	AVON	2012	0
24	AVON	2013	0

Aggregated dataset #1: Accidental Drug Related Deaths by Town, 2012-2018



	A	B	C	D	E	F
1	**Town**	**FIPS**	**Year**	**Measure Type**	**Variable**	**Value**
2	ANDOVER	0901301080	2012	Number	Total	0
3	ANDOVER	0901301080	2013	Number	Total	0
4	ANDOVER	0901301080	2014	Number	Total	1
5	ANDOVER	0901301080	2015	Number	Total	1
6	ANDOVER	0901301080	2016	Number	Total	0
7	ANDOVER	0901301080	2017	Number	Total	0
8	ANDOVER	0901301080	2018	Number	Total	1
9	ANDOVER	0901301080	2012	Number	Cocaine	0
10	ANDOVER	0901301080	2013	Number	Cocaine	0
11	ANDOVER	0901301080	2014	Number	Cocaine	0
12	ANDOVER	0901301080	2015	Number	Cocaine	0
13	ANDOVER	0901301080	2016	Number	Cocaine	0
14	ANDOVER	0901301080	2017	Number	Cocaine	0
15	ANDOVER	0901301080	2018	Number	Cocaine	0
16	ANDOVER	0901301080	2012	Number	Heroin	0
17	ANDOVER	0901301080	2013	Number	Heroin	0
18	ANDOVER	0901301080	2014	Number	Heroin	1
19	ANDOVER	0901301080	2015	Number	Heroin	0
20	ANDOVER	0901301080	2016	Number	Heroin	0
21	ANDOVER	0901301080	2017	Number	Heroin	0
22	ANDOVER	0901301080	2018	Number	Heroin	0
23	ANDOVER	0901301080	2012	Number	Fentanyl	0

Aggregated dataset #2: Accidental Drug Related Deaths by Town and Drug, 2012-2018