Appendix C: Model Benefit-Risk Analysis

Step 1: Evaluate the Information the Dataset Contains

Dataset: ________________________________

Consider the following categories of information:

- *Direct Identifiers:* These are data points that identify a person without additional information or by linking to other readily available information. "Personally Identifiable Information," or PII, often falls within this category. For example, they can be names, social security numbers, or an employee ID number. (*See, e.g.,* municipal guidance like Seattle's PII/Privacy in the Open Dataset Inventory). Publishing direct identifiers creates a *very high* risk to privacy because they directly identify an individual and can be used to link other information to that individual.

- *Indirect Identifiers:* These are data points that do not directly identify a person, but that in combination can single out an individual. This could include information such as birth dates, ZIP codes, gender, race, or ethnicity. (*See, e.g.,* municipal guidance like Seattle's PII/Privacy in the Open Dataset Inventory). In general, to preserve privacy, experts recommend including no more than 6-8 indirect identifiers in a single dataset.[1] If a dataset includes 9 or more indirect identifiers there is a *high* or *very high* risk to privacy because they can indirectly identify an individual.

- *Non-Identifiable Information:* This is information that cannot reasonably identify an individual, even in combination. For example, this might include city vehicle inventory or atmospheric readings. This data creates *very low* or *low* risk to privacy.

- *Sensitive Attributes:* These data points that may be sensitive in nature. Direct and indirect identifiers can be sensitive or not, depending on context. For example, this might include financial information, health conditions, or a criminal justice records. Sensitive attributes typically create *moderate*, *high*, or *very high* risk to privacy.

- *Spatial Data and Other Information that Is Difficult to De-identify:* Certain categories or data are particularly difficult to remove identifying or identifiable information from, including: geographic locations, unstructured text or free-form fields, biometric information, and photographs or videos.[2] If data to be included in a public dataset are in one of these formats, they may create a *high* or *very high* risk to privacy.

[1] *See* Khaled El Emam, *A De-Identification Protocol for Open Data*, IAPP (MAY 16, 2016), https://iapp.org/news/a/a-de-identification-protocol-for-open-data/.
[2] *See* GARFINKEL, *supra* note 9, at 32-33.

Consider how linkable the information in this dataset is to other datasets:

- Do any of the dataset's direct or indirect identifiers currently appear in other readily accessible open datasets (e.g., other municipal county, or state open datasets)? If this information is present in multiple open datasets, it increases the chances of identifying an individual and increases the risk to privacy.

- How often is the dataset updated? In general, the more frequently a dataset is updated—every fifteen minutes versus every quarter, for example—the easier it is to re-identify an individual and the greater the risk to privacy.

- How often is the information in this dataset requested by public records?

Consider how the information in this dataset was obtained:

- In what context was this data collected? Is this data collected under a regulatory regime? Are there any conditions, such as a privacy policy or contractual term, attached to the data? If the personal information in this dataset collected directly from the individual or from a third party?

- Would there be a reasonable expectation of privacy in the context of the data collection? For example, if the public has no notice of the data collection or data are collected from private spaces, there may be an expectation of privacy.

- Was the collection of the information in this dataset controversial? Was any of the information in this dataset collected by surveillance technologies (e.g., body-worn cameras, surveillance cameras, unmanned aerial vehicles, automatic license plate readers, etc.)?

- Has this dataset been checked for accuracy? Is there a mechanism for individuals to have information about themselves in this dataset corrected or deleted?

- Is there a concern that releasing this data may lead to public backlash or negative perceptions?

Step 2: Evaluate the Benefits Associated with Releasing the Dataset

List some of the foreseeable benefits of publishing the data fields included in this dataset and identify whether this use typically involves aggregate data or individual records. For example, measuring atmospheric data at particular locations over time may reveal useful weather patterns, and tracking building permit applications may reveal emerging demographic or commercial trends in particular neighborhoods.

Consider the likely users of this dataset. Who are the ideal users? Check all that apply.

- ☐ Individuals
- ☐ Community Groups
- ☐ Journalists
- ☐ Researchers
- ☐ Companies or Private Entities
- ☐ Other Government Agencies or Groups
- ☐ Other: ____________________________

Assess the scope of the foreseeable benefits of publishing the dataset:

Qualitative Value	**Quantitative Value**	**Description**
Very High	10	The dataset will likely have *multiple compelling and important* utilities for individuals, the community, other organizations, or society.
High	8	The dataset will likely have a *compelling and important* utility for individuals, the community, other organizations, or society.
Moderate	5	The dataset will likely have a *clear* utility for individuals, the community, other organizations, or society. While the utility is clear, it is not as urgent as a "high" value.
Low	2	The dataset will likely have a *limited* utility for individuals, the community, other organizations, or society.
Very Low	0	The dataset will likely have *negligible* utility for organizations, the community, other organizations, or society.

Next, assess the likelihood that the desired benefits of releasing this dataset would occur:

Qualitative Value	Quantitative Value	Description
Very High	10	The benefit is *almost certain* to occur.
High	8	The benefit is *highly likely* to occur.
Moderate	5	The benefit is *somewhat likely* to occur.
Low	2	The benefit is *unlikely* to occur.
Very Low	0	The benefit is *highly unlikely* to occur.

Combining your rating of the foreseeable benefits of the dataset with the likelihood that these benefits will occur, assess the overall benefit of this dataset:

Likelihood of Occurrence	**Impact of Foreseeable Benefits**				
	Very Low Impact	**Low Impact**	**Moderate Impact**	**High Impact**	**Very High Impact**
Very High Likelihood	Low Benefit	Moderate Benefit	High Benefit	Very High Benefit	Very High Benefit
High Likelihood	Low Benefit	Moderate Benefit	Moderate Benefit	High Benefit	Very High Benefit
Moderate Likelihood	Low Benefit	Low Benefit	Moderate Benefit	Moderate Benefit	High Benefit
Low Likelihood	Very Low Benefit	Low Benefit	Low Benefit	Moderate Benefit	Moderate Benefit
Very Low Likelihood	Very Low Benefit	Very Low Benefit	Low Benefit	Low Benefit	Low Benefit

Step 3: Evaluate the Risks Associated with Releasing the Dataset

Consider the foreseeable privacy risks of this dataset:[3]

- *Re-identification (and false re-identification) impacts on individuals*
  - Would a re-identification attack on this dataset expose the person to identity theft, discrimination, or abuse?
  - Would a re-identification attack on this dataset reveal location information that could lend itself to burglary, property crime, or assault?
  - Would a re-identification attack on this dataset expose the person to financial harms or loss of economic opportunity?
  - Would a re-identification attack on this dataset reveal non-public information that could lead to embarrassment or psychological harm?

- *Re-identification (and false re-identification) impacts on the organization*
  - Would a re-identification attack on this dataset lead to embarrassment or reputational damage to the City of Seattle?
  - Would a re-identification attack on this dataset harm city operations relying on maintaining data confidentiality?
  - Would a re-identification attack on this dataset expose the city to financial impact from lawsuits, or civil or criminal sanctions?
  - Would a re-identification attack on this dataset undermine public trust in the government, leading to individuals refusing to consent to data collection or providing false data in the future?

- *Data quality and equity impacts*
  - Will inaccurate or incomplete information in this dataset create or reinforce biases towards or against particular groups?
  - Does this dataset contain any incomplete or inaccurate data that, if relied upon, would foreseeably result in adverse or discriminatory impacts on individuals?
  - Will any group or community's data be disproportionately included in or excluded from this dataset?
  - If this dataset is de-identified through statistical disclosure measures, did that process introduce significant inaccuracies or biases into the dataset?

---

[3] Special thanks to Simson Garfinkel and Khaled El Emam whose works provide a foundation for articulating this analytic framework. *See* DE-IDENTIFICATION OF PERSONAL INFORMATION 32-33 (NIST 2015), DE-IDENTIFYING GOVERNMENT DATASETS SP 800-188; Khaled El Emam, *A De-Identification Protocol for Open Data*, IAPP (MAY 16, 2016), https://iapp.org/news/a/a-de-identification-protocol-for-open-data/; KHALED EL EMAM, GUIDE TO THE DE-IDENTIFICATION OF PERSONAL HEALTH INFORMATION (2013).

- *Public trust impacts*
    - Does this dataset have information that would lead to public backlash if made public?
    - Will local individuals or communities be shocked or surprised by the information about themselves in this dataset?
    - Is it likely that the information in this dataset will lead to a chilling effect on individual, commercial, or community activities?
    - Is there any information contained within the dataset that would, if made public, reveal nonpublic information about an agency's operations?

Consider who could use this information improperly or in an unintended manner (including to re-identify individuals in the dataset). Check all that apply.

- [ ] General public (individuals who might combine this data with other public information)
- [ ] Re-identification expert (a computer scientist skilled in de-identification)
- [ ] Insiders (a municipal employee or contractor with background information about the dataset)
- [ ] Information brokers (an organization that systematically collects and combines identified and de-identified information, often for sale or reuse internally)
- [ ] "Nosy neighbors" (someone with personal knowledge of an individual in the dataset who can identify that individual based on the prior knowledge)
- [ ] Other: ______________________________________

Assess the scope of the foreseeable privacy risks of publishing the dataset:

Qualitative Value	**Quantitative Value**	**Description**
Very High	10	The dataset will likely have *multiple severe or catastrophic* adverse effects on individuals, the community, other organizations, or society.
High	8	The dataset will likely have a *severe or catastrophic* adverse effect on individuals, the community, other organizations, or society.
Moderate	5	The dataset will likely have a *serious* adverse effect on individuals, the community, other organizations, or society.
Low	2	The dataset will likely have a *limited* adverse impact on individuals, the community, other organizations, or society,
Very Low	0	The dataset will likely have a *negligible* adverse impact on individuals, the community, other organizations, or society.

Next, assess the likelihood that the foreseeable privacy risks of releasing this dataset would occur:

Qualitative Value	**Quantitative Value**	**Description**
Very High	10	The risk is *almost certain* to occur.
High	8	The risk is *highly likely* to occur.
Moderate	5	The risk is *somewhat likely* to occur.
Low	2	The risk is *unlikely* to occur.
Very Low	0	The risk is *highly unlikely* to occur.

Combining your rating of the foreseeable risks of the dataset with the likelihood that these risks will occur, assess the overall risk of this dataset:

Likelihood of Occurrence	Impact of Foreseeable Risks				
	Very Low Impact	**Low Impact**	**Moderate Impact**	**High Impact**	**Very High Impact**
Very High Likelihood	Low Risk	Moderate Risk	High Risk	Very High Risk	Very High Risk
High Likelihood	Low Risk	Moderate Risk	Moderate Risk	High Risk	Very High Risk
Moderate Likelihood	Low Risk	Low Risk	Moderate Risk	Moderate Risk	High Risk
Low Likelihood	Very Low Risk	Low Risk	Low Risk	Moderate Risk	Moderate Risk
Very Low Likelihood	Very Low Risk	Very Low Risk	Low Risk	Low Risk	Low Risk

Step 4: Weigh the Benefits against the Risks of Releasing the Dataset

Step 4A: Combine the overall scores from the benefit and risk analyses to determine the appropriate solution for how to treat the dataset.

Benefit	Risks				
	Very Low Risk	**Low Risk**	**Moderate Risk**	**High Risk**	**Very High Risk**
Very High Benefit	Open	Open	Limit Access	Additional Screening	Additional Screening
High Benefit	Open	Limit Access	Limit Access	Additional Screening	Additional Screening
Moderate Benefit	Limit Access	Limit Access	Additional Screening	Additional Screening	Do Not Publish
Low Benefit	Limit Access	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish
Very Low Benefit	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish	Do Not Publish

- *Open*: Releasing this dataset to the public presents low or very low privacy risks and the potential benefits of the dataset substantially outweigh the potential privacy risks.
- *Limit Access*: Releasing this data presents moderate to very low privacy risks and the potential benefits of the dataset outweigh the potential privacy risks. In order to reduce the privacy risk, limit access to the dataset (such as by attaching contractual/Terms of Service terms to the dataset prohibiting re-identification attempts).
- *Additional Screening:* Releasing this dataset presents high privacy risks and the benefits could outweigh the potential privacy risks, or releasing this dataset presents privacy risk and the potential benefits do not outweigh the potential privacy risks. In order to reduce the privacy risk, formal application and oversight mechanisms should be considered (such as a disclosure review board, data use agreements, or a secure data enclave).
- *Do Not Publish*: Releasing this dataset presents very high to moderate privacy risks and the potential privacy risks of the dataset substantially outweigh the potential benefits. This dataset should remain closed, unless the risk can be reduced or there are countervailing public policy reasons for publishing it.

If the above table results in an "Open" categorization, then record the final benefit-risk score and continue preparing to publish the dataset. If the above table does *not* result in an "Open" categorization, then proceed to Step 4B by applying appropriate de-identification controls to mitigate the privacy risks for this dataset. The de-identification methods described below will be appropriate for some datasets, but not for others. Advances are always being made in de-identification techniques, and some tools may require disclosure control experts to properly implement. In the long-term, municipalities should strive to incorporate the expertise of disclosure control professionals and to implement mathematically provable privacy protections like differential privacy.

Consider the level of privacy risks you are willing to accept, the overall benefit of the dataset, and the operational resources available to mitigate re-identification risk. Note that the more invasive the de-identification technique, the greater the loss of utility will be in the data, but also the greater the privacy protection will be.

Technical Controls[4]

Method	Description	Privacy Impact	Utility Impact	Operational Costs
Suppression	Removing a data field or an individual record to prevent the identification of individuals in small groups or those with unique characteristics.	Removing the field removes the risk created by those fields, and lowers the likelihood of linking one dataset to another based on that information. Removing individual records can also effectively protect the privacy of those individuals. Suppression cannot guarantee absolute privacy, because there is always a chance that the remaining data can be re-identified using an auxiliary dataset.	This approach removes all utility added by the suppressed field or record, and could skew the results or give false impressions about the underlying data.	This is a relatively low-cost method of de-identification. Removing entire fields of data can be both a quick and relatively low-tech process. When removing records one-by-one, particularly large datasets, there is a risk that some records may be overlooked.[5]
Generalization/Blurring	Reducing the precision of disclosed data to minimize the certainty of individual identification, such as by replacing precise data values with ranges or sets.	The more specific a data value is, the easier it will generally be to single out an individual. However, even relatively broad categories cannot guarantee absolute privacy, because there is always a chance that the remaining	Generalizing data fields can render data useless for more granular analysis, and may skew results slightly or give false impressions about the underlying data.	Generalizing data fields can be a quick and straightforward process for reducing the identifiability of particular fields after the initial thresholds are set. In order to determine the appropriate level of generalization for particular data types, additional

[4] Special thanks to the Berkman Klein Center for Internet & Society at Harvard University whose work provides a foundation for this analytic framework. BEN GREEN ET AL, OPEN DATA PRIVACY (2017), https://dash.harvard.edu/handle/1/30340010; Micah Altman et al., *Towards a Modern Approach to Privacy-Aware Government Data Releases*, 30 BERKELEY TECH. L.J. 1968 (2015), https://cyber.harvard.edu/publications/2016/Privacy_Aware_Government_Data_Releases.
[5] *See* Fitzpatrick, *supra* note 9.

Method	Description	Privacy Impact	Utility Impact	Operational Costs
		data can be re-identified using an auxiliary dataset.		research or expert consultation may be required.
Pseudonymization	Replacing direct identifiers with a pseudonym (such as a randomly generated value, an encrypted identifier, or a statistical linkage key).	Pseudonymization removes the association between an individual and their data, and replaces it with a less easily identifiable key, lowering but not eliminating the risk of re-identification. Pseudonymization can be reversed in many circumstances, and are often considered personally identifiable information by privacy and data protection authorities.	Pseudonymization can allow for information about an individual to be linked across multiple records, increasing its utility for a wide variety of purposes.	Pseudonymization can appear relatively straightforward and cost-effective, however creating *irreversible* pseudonyms suitable for open data release can require significant effort.[6] Most successful re-identification attacks on openly released data have come from data that was inadequately pseudonymized.[7]
Aggregation	Summarizing the data across the population and then releasing a report based on those data (such as contingency tables or summary statistics),	Aggregating data can be an effective method for protecting privacy as there is no raw data directly tied to an individual, however experts recommend minimum cell sizes of 5-10 records.[8]	Aggregation is more useful for examining the performance of a group or cohort. Because the raw data is not presented, it cannot be relied on to generate additional insights.	This method of de-identification requires slightly more expertise than simply removing fields or records. After an initial learning curve, the method can be

[6] *See* GARFINKEL, *supra* note 9, at 17.
[7] *See* Ira Rubinstein & Woodrow Hartzog, *Anonymization and Risk,* 91 WASH. L REV. 703 (2016), http://digital.law.washington.edu/dspace-law/bitstream/handle/1773.1/1589/91WLR0703.pdf?sequence=1&isAllowed=y; Jules Polonetsky, Omer Tene & Kelsey Finch, *Shades of Gray: Seeing the Full Spectrum of Practical Data De-Identification,* 56 SANTA CLARA L. REV. 594 (2016).
[8] *See* Khaled El Emam, Comment Letter on Proposed Rule to Protect the Privacy of Customers of Broadband and Other Telecommunications Services; Khaled El Emam, *Protecting Privacy Using k-Anonymity,* 15 J. AM. MED. INFORMATICS ASS'N (2008).

Method	Description	Privacy Impact	Utility Impact	Operational Costs
	rather than releasing individual-level data.			implemented without significant costs. Expert consultants or guidance from federal statistical agencies may provide guidance in setting minimum cell sizes or addressing particular data types.[9]
Visualizations	Rather than providing users access to raw microdata, data may be presented in more privacy-protective formats, such as data visualizations or heat maps.	When data is released in non-tabular formats, individual data records are typically more obscure and harder to link to other auxiliary datasets, protecting individual privacy.	Data released in these sorts of formats may still be highly useful for a range of purposes, although not all. These formats may also limit the ways in which datasets can be combined or built on to generate new insights. Visualizations and other alternative data formats may also be more engaging to the lay public than raw tabular data.	These are fairly low-cost approaches to limiting privacy risks, with numerous public resources readily available to Open Data program staff. Data that update frequently may be harder to maintain.
Perturbation	An expert adds "noise" to the dataset (such as swapping values from one record to another, or replacing one value with an artificial value), making it difficult to	The false data in the field makes re-identification much less likely to occur. The noise makes it difficult to determine if re-identification is associated with a specific individual.	Utility decreases as the amount of noise in the data increases. The proportionate amount of legitimate data is reduced as false data is added.	This is costly in that it requires an expert. The type of noise, as well as the amount to be added will have a drastic difference, and to ensure a retention in utility, it must be

[9] *Id*.

Method	Description	Privacy Impact	Utility Impact	Operational Costs
	distinguish between legitimate values and the "noise."			completed by an expert. However, research shows that "even relatively small perturbations to the data may make re-identification difficult or impossible."[10]
k-Anonymity	A technique to measure and limit how many individuals in a dataset have the same combination of identifiers. K-anonymity suppresses or generalizes identifiers and perturbs outputs until a particular k-value is reached.	Privacy protection is greater as the value of "k" increases. Experts recommend that the k-value for open datasets should be at least k=11 (that is, for every combination of identifiers in a dataset, there should be at least 11 equivalent records).[11]	As with the above controls, the negative impact on utility increases as k-value increases. In order to achieve k=11, significant portions of some datasets may need to be suppressed or generalized.	This is a costly, complex, and time-consuming method. An expert in de-identification and k-anonymity is necessary to ensure that the k-value is correct and will provide the desired level of protection and utility. Subsequent research has led to additional requirements for the diversity of sensitive attribute within k-anonymous datasets (l-diversity) and statistical relationship to the original data (t-closeness).[12]
Differential Privacy	A formal mathematical definition of privacy, which may be satisfied by a range of techniques	Differential private solutions increase privacy for all individuals in a dataset and provide	As with other above tools, differential private solutions decrease	Differential privacy requires an expert to calculate the leakage threshold, the amount of noise to add,

[10] *See* GARFINKEL, *supra* note 9, at 29.
[11] El Emam, *supra* note 42.
[12] *See* GARFINKEL, *supra* note 9, at 12.

Method	Description	Privacy Impact	Utility Impact	Operational Costs
	if the result of an analysis of a dataset is the same before and after the removal of a single data record.	mathematical guarantees against a wider range of re-identification attacks than traditional de-identification techniques. Some differential privacy solutions rely on limiting the number of queries completed to prevent maintain a proven minimum privacy threshold (often known as the "privacy budget"). The more queries performed on a function, the more the total "leakage" increases. The leakage can never decrease, and there is an acceptable level of leakage that can occur before a privacy risk becomes likely and the dataset must be abandoned. Non-interactive differential privacy solutions such as synthetic data also provide	the accuracy of analysis performed on the dataset. The amount of noise is calibrated to the amount of privacy protection offered, and in larger datasets may be negligible.[15] In other deployments, the level of utility in a differentially private dataset may be dependent upon the number of queries to be made in the dataset. Once the leakage threshold is hit, the dataset can no longer be used. However, if the desired task can be accomplished under the leakage threshold, the dataset retains great utility with little risk to privacy. In other cases, such as synthetic data (see below), differentially private tools may be non-interactive and so not limited by query	and other statistical nuances. It may also require an interactive query system to be established, or trained users who can create data summaries for release and use. Therefore, it carries a higher operational cost than other methods of de-identification. Differential privacy is an active research area, and while to date it has only been applied to a few operational system,[18] differential privacy tools for use by non-experts in privacy, computer science, and statistics are also currently in development.[19]

[15] *Comment by Alexandra Wood, Micah Altman, Suso Baleato, and Salil Vadhan* to Future of Privacy Forum (Oct. 3, 2017), *available at* https://fpf.org/wp-content/uploads/2018/01/Wood-Altman-Baleato-Vadhan_Comments-on-FPF-Seattle-Open-Data-Draft-Report.pdf.
[18] *See* GARFINKEL*, supra* note 9, at 7-9.
[19] *See* Wood et al., supra note 56. (citing e.g., Marco Gaboardi et al., PSI (Ψ): A Private Data Sharing Interface, Working Paper (2016), *available at* https://arxiv.org/abs/1609.04340).

Method	Description	Privacy Impact	Utility Impact	Operational Costs
		strong privacy protection when sharing statistics,13 as "the privacy loss budget can be spent in creating the synthetic dataset, rather than in responding to interactive queries."[14]	amounts, such as by enabling data or data summaries to be released and used.[16] Datasets that may otherwise be too sensitive to share in individual-level formats could still be safely analyzed in differentially private formats, as well.[17]	
Synthetic Data	A process in which seed data from an original dataset is used to create artificial data that has some of the statistical characteristics as the seed data.[20] Datasets may be partially synthetic (in which some of the data is inconsistent with the original dataset) or fully synthetic (in which there is no one-to-one mapping between any	Synthetic datasets can make it very difficult and costly to map artificial records to actual people, and supports mathematical privacy guarantees with differential privacy that can remain in force "even if there are future data releases."[22]	Synthetic data "can be confusing to the lay public," as they may contain artificial individuals who "appear quite similar to actual individuals in the population."[23] The utility of synthetic data also depends on the model used to create it. Synthetic databases, unlike some differential privacy deployments, do not need to be released via	Synthetic databases may be confusing to both researchers and lay people, requiring additional efforts to educate data users about the dataset's contents and limitations.

[13] *See* Wood et al., supra note 56 (citing Census, Google, Apple, Uber).
[14] Garfinkel, *supra* note 9, at 52.
[16] *See* Wood et al., supra note 56.
[17] *See* Wood et al., supra note 56.
[20] Garfinkel, *supra* note 9, at 48-49.
[22] *Id*. at 51.
[23] *Id*.

Method	Description	Privacy Impact	Utility Impact	Operational Costs
	record in the original dataset and the synthetic dataset).[21]		interactive query systems, as "the privacy loss budget can be spent in creating the synthetic dataset, rather than in responding to interactive queries."[24]	

Administrative and Legal Controls

Method	Description	Privacy Impact	Utility Impact	Operational Costs
Contractual provisions	Data is made available to qualified users under legally binding contractual terms (such as commitments not to attempt to re-identify individuals or link datasets, to update the information periodically, or to use data in noncommercial and nondiscriminatory ways).	Contractual controls alone do not necessarily reduce the risk of re-identification, but when complementing the technical controls above can provide more flexible and contextual privacy protections. Contractual terms are more robust when backed up by audit requirements and penalties for noncompliance.	Contractual provisions do not impede utility for acceptable data uses, although the compliance costs may deter some potential data users. Contractual terms prohibiting commercial uses may deter certain categories of users (such as businesses or data brokers).[25]	Consistent contractual provisions must be developed and deployed, but this is a less extensive process than many of the technical measures above. Contractual provisions can also be tailored to the specific risk profiles of each dataset. There may be legal limits on how governments can restrict the use of data as well.[26]
Access fees	Charging users for access to data increases accountability and may	Because fees are likely to deter many casual browsers of a particular datasets, the	The deterrent effect of access fees on the general public will impede the	Introducing access fees comes with initial and ongoing administrative overhead, and

[21] *Id. at* 49-54.
[24] *Id*. at 52.
[25] *See* Jan Whittington et al., supra note 13, at 1962.
[26] *Id.* at 1963.

	discourage improper use of data.	likelihood of accidental re-identification of an individual by a curious friend, neighbor, or acquaintance generally decreases. Tiered fee structures (e.g., that charge more for commercial access or remote versus in-person data access) may also lower the risk of re-identification by other actors. Charging fees may also introduce registration and audit capabilities, allowing Open Data program staff to identify which data users accessed which datasets.	potential utility of the dataset and could limit access by some marginalized or vulnerable communities (e.g., those without credit cards, technological sophistication, or new market entrants).	requires thoughtful determination of when particular datasets or classes of users warrant the use of fees.
Data enclaves	Physical or virtual environments are created that enable "authorized users to access confidential data and analyze the data using provided statistical software."[27]	Risks of re-identification are almost entirely removed by restricting external access to even de-identified data and introducing accountability and oversight measures. Technical controls may not need to be as strict, when complemented by administrative and legal	Data utility can be maximized for qualified researchers, as privacy protections are no longer purely technical. Researchers may be limited in what research questions can be asked and in the format of their results.	There are significant operational costs to maintaining a secure data enclave, including establishing policies and procedures for granting qualified researcher queries, for processing queries on de-identified data, for establishing the enclave, and

[27] *See* Micah Altman et al., *supra* note 23, at 40; GARFINKEL, *supra note* 9 at ix.

		safeguards (such as requiring researchers to apply for access, describe the proposed research, agree to confidentiality laws and penalties, audit logs, and authentication measures).	But data utility is completely removed for any individual or organization that is not approved to access the dataset.	for monitoring the program over time.
Tiered access controls	Systems in which data are made available to different categories of users through different mechanisms.[28]	Tiered access controls permit municipalities to craft more granular and contextual privacy protections depending on the sensitivity and identifiability of the data, and may support more accountability mechanisms (e.g., providing more sensitive or identifiable data only to potential data users who sign enforceable data use agreements or have their research questions vetted in advance).	Limiting access to some datasets to particular types of users may increase the utility of data to those who qualify for greater access but decrease it for those who do not or cannot satisfy the access requirements. This may deter some members of the public from engaging with certain open datasets, but it may also provide municipal data leaders more oversight and insight into which data are most valuable to users.	Establishing and monitoring an access-control system may require meaningful operational overhead. Consistent access terms and conditions will need to be defined, and deployed, and enforced. Access models that intend to do individualized vetting of some subsets of data users will likely require additional staffing.
Ethical and/or disclosure review board	Particularly risky or ambiguous policy decisions about a dataset are escalated to an advisory group with broad expertise	Review boards with diverse backgrounds and subject matter expertise can more robustly debate the benefits and risks of releasing a	A review board may determine that a dataset's utility ultimately outweighs its impact on individual privacy; it may also	Establishing and maintaining an accountable and transparent body of experts can be a challenging operational endeavor,

[28] *See* Wood et. al., supra note 56.

	and community engagement for further review.[29]	dataset and can address any additional dimensions not captured by the privacy risk assessment.	determine that the benefits do *not* outweigh the risks.	although guidance and models from academic data research are available.[30]

Step 4B: After determining and applying appropriate privacy controls and mitigations for the dataset, re-assess the overall risks and benefits of the dataset (Steps 1-3). Note any mitigation steps taken, and record the final benefit-risk score:

Benefit	**Risks**				
	Very Low Risk	**Low Risk**	**Moderate Risk**	**High Risk**	**Very High Risk**
Very High Benefit	Open	Open	Limit Access	Additional Screening	Additional Screening
High Benefit	Open	Limit Access	Limit Access	Additional Screening	Additional Screening
Moderate Benefit	Limit Access	Limit Access	Additional Screening	Additional Screening	Do Not Publish
Low Benefit	Limit Access	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish
Very Low Benefit	Additional Screening	Additional Screening	Do Not Publish	Do Not Publish	Do Not Publish

If the score is still not "Open," consider using another mitigation method. If this is not possible, then determine whether to publish the dataset. If there may be countervailing public policy factors that should be considered, move on to Step 5.

- *Open*: Releasing this dataset to the public presents low or very low privacy risks and the potential benefits of the dataset substantially outweigh the potential privacy risks.

[29] *See generally* CONFERENCE PROCEEDINGS: BEYOND IRBS: ETHICAL GUIDELINES FOR BIG DATA RESEARCH, FUTURE OF PRIVACY FORUM (Dec. 10, 2015), https://fpf.org/wp-content/uploads/2017/01/Beyond-IRBs-Conference-Proceedings_12-20-16.pdf.
[30] *See* 45 C.F.R. 46.102; OMER TENE & JULES POLONETSKY, BEYOND IRBS: ETHICAL GUIDELINES FOR BIG DATA RESEARCH 1 (Dec. 2015), https://bigdata.fpf.org/wp-content/uploads/2015/12/Tene-Polonetsky-Beyond-IRBs-Ethical-Guidelines-for-Data-Research1.pdf.

- *Limit Access*: Releasing this data presents moderate to very low privacy risks and the potential benefits of the dataset outweigh the potential privacy risks. In order to reduce the privacy risk, limit access to the dataset (such as by attaching contractual/Terms of Service terms to the dataset prohibiting re-identification attempts).
- *Additional Screening:* Releasing this dataset presents high privacy risks and the benefits could outweigh the potential privacy risks, or releasing this dataset presents privacy risk and the potential benefits do not outweigh the potential privacy risks. In order to reduce the privacy risk, formal application and oversight mechanisms should be considered (such as a disclosure review board, data use agreements, or a secure data enclave).
- *Do Not Publish*: Releasing this dataset presents high or very high privacy risks and the potential privacy risks of the dataset substantially outweigh the potential benefits. This dataset should remain closed, unless the risk can be reduced or there are countervailing public policy reasons for publishing it.

Step 5: Evaluate Countervailing Factors

Sometimes, a dataset with a very high privacy risk is still worth releasing into the open data portal in light of public policy considerations. For example, a dataset containing the names and salaries of elected officials would likely be considered high-risk due to the inclusion of a direct identifier. However, there is a compelling public interest in making this information available to citizens that outweighs the risk to individual privacy.

Additionally, there are always risks associated with maintaining and releasing any kind of data relating to individuals. Two key considerations when deciding whether to release the data irrespective of a potentially high or very high risk to individual privacy are:

1. If you are on the edge between two categories, analyze the dataset holistically but err on the side of caution. A dataset that is not released immediately can still be released at another date, as additional risk mitigation techniques become available. A dataset that has been released publicly, however, cannot ever be fully pulled back, even if it is later discovered to pose a greater risk to individual privacy. Be particularly cautious about moving data from an original recommendation of *Do Not Publish* to *Open*, and ensure that the potential benefits of releasing the data are truly so likely and compelling that they outweigh the existing privacy risks.

Any time you deviate from the original analysis, document your reasoning for doing so. This will not only help you decide whether the deviation is, in fact, the correct decision, but also provides accountability. Should the need arise, you will have a record of your reasoning, including analysis of the expected benefits and the recognized risks at the time. Where personally identifiable information is published notwithstanding the privacy risk, accountability mechanisms help maintain trust in the Open Data program that may otherwise be lost.