Open Durham Guidebook:

Data Stewards Edition

City and County of Durham

(Version 0.5)

This document is adapted from a similar document made available by the City and County of San Francisco. All credit goes to them.

Table of Contents

1. Background

Purpose of this Guidebook and Version Notes

The purpose of this guidebook is to provide guidance to Data Stewards in the City and County of Durham. Data Stewards should use this guidebook to help them in their new role. We'll update this guide as the role and responsibilities of the Data Stewards evolve.

Date	Version	Description of changes
March 1st, 2017	0.1	Adoption of Guidebook from the City and County of San Francisco

Why release and publish data?

Before beginning, you should note the difference between releasing and publishing data. They are as follows.

> **Releasing data**: Submitting a data release form to the Open Data Program so that they have certain control over a dataset that your department produces. You will still have ownership over this dataset but the Open Data Program will have the ability to transform this data.
>
> **Publishing data**: Making a dataset publicly available on the data portal.

One of the first questions departments often ask when learning about data is "Why should my department release and/or publish data?" There are a number of reasons, both practical and philosophical, why releasing data can benefit your department and the people it serves.

Improve internal data sharing

The Durham data portal can help City and County employees solve internal challenges in accessing data between departments. Right now, data sharing is an ad-hoc affair occurring as data needs arise organically. The Durham data portal is a platform that allows for that process to be streamlined. Bridging the gap between data silos will lead to reduced response times, allow accurate data to be shared in real time, and improve strategic coordination between departments.

Simplify responses to external data requests

Making data available on the Durham data portal is an effective way to respond to data requests. Making data publicly available can save both time and money over the long term as publishing just one dataset can address multiple requests at the same time while also eliminating the chance of future requests for the same data set.

Accelerate the problem solving process

Solving problems within government often requires cooperation across multiple different users, teams, departments, and organizations. Using the data portal as a central terminal for published data will allow internal users to have an up-to-date, authoritative library of information that they can access at any time. The ability to cross-examine and analyze different data streams unlocks the potential for making data driven decisions that weren't possible before.

Change the culture around data

The purpose of Open Durham is to change how data is consumed both internally and externally. Durham can become a data driven culture where value is derived from sharing and aligning information.

What is open data?

Open Durham strives to increase the quality and access of open data in Durham. According to the Open Knowledge Foundation:

"Open data is data that can be freely used, re-used and redistributed by anyone - subject only, at most, to the requirement to attribute and sharealike."

The full Open Definition gives precise details as to what this means. To summarize the most important:

> **Availability and Access**: the data must be available as a whole and at no more than are reasonable reproduction cost, preferably by downloading over the internet. The data must also be available in a convenient and modifiable form.
>
> **Re-use and Redistribution**: the data must be provided under terms that permit re-use and redistribution including the intermixing with other datasets.
>
> **Universal Participation**: everyone must be able to use, re-use and redistribute - there should be no discrimination against fields of endeavour or against persons or groups. For example, 'non-commercial' restrictions that would prevent 'commercial' use, or restrictions of use for certain purposes (e.g. only in education), are not allowed

To learn more about Open Data, visit any of the following links:

- Open Data - Wikipedia
- Open Data: What is It and Why Should You Care? (GovTech, 2014)
- What Is Open Data? (Open Data Handbook)
- What Is Open Data? (UK ODI)
- 8 Principles of Open Data (opengovdata.org)
- 5 Star Data (5stardata.info)

Open Durham: Our Data Portal

OpenDurham (https://opendurham.nc.gov/page/home/) is the City and County of Durham's data portal. The Durham portal is hosted by a vendor, OpenDataSoft. The Durham data portal allows users to:

- Access raw tabular data
- Create charts and graphs

- Layer spatial data and build maps (also possible on the GIS portal)
- Create dashboards that can track progress and manage goals
- Build applications for public and internal benefit

2. Roles and Responsibilities

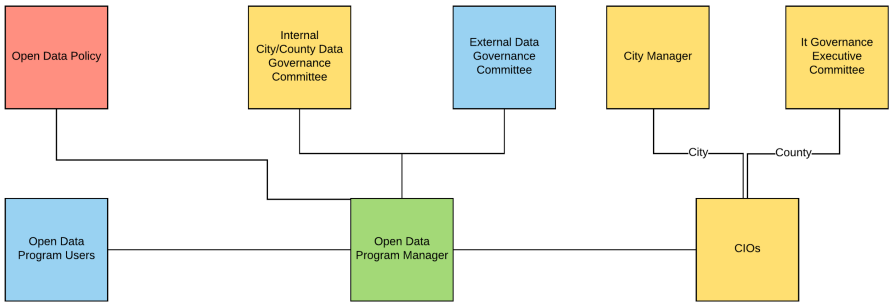
Below we included an overview of the roles and general responsibilities in support of City and County Open Data program. The open data policy ITP-5 (available here on CODI) provides greater detail on both the Data Stewards and the Open Data Program Manager. We expect to modify or change these as we learn more and as the roles become more refined.

Comment [1]: Change to link to Durham's open data policy

Role	General Responsibilities
Data Stewards	Data Stewards are designated for each department as the main point of contact and accountability for data in their department. General responsibilities include: ● Releasing and publishing department data sets ● Serving as a key point of accountability for timelines and questions about data sets ● Implementing privacy, metadata and other standards and practices
Open Data Program Manager	The Open Data Program Manager is designated by the Chief Information Officer for the City and County of Durham and is accountable for the city's overall implementation of the open data policy. General responsibilities include: ● Creating processes, rules and standards to implement the data policy, including but not limited to: ○ Prioritizing data sets for publication ○ Determining what datasets are appropriate for public disclosure ○ Creating data licensing and metadata standards and guidelines ○ Providing guidance and assistance to City and County departments in releasing data ○ Providing guidance and assistance to City and County departments in assessing and, where appropriate, improving the accuracy, completeness, interoperability and other quality dimensions of data ○ Facilitating creation of department implementation plans and reporting ● Maintaining the data portal ● Maintaining the City and County data inventory ● Presenting an annual citywide and countywide data report ● Assisting departments with analysis of city and county datasets ● Facilitate and manage the Data Governance Committee
Data Governance Committee	The Open Data Program Manager will convene a committee of City and County employees that are internal stakeholders. This committee will pave the way for data governance within the City and County. General Responsibilities include: ● Attend a 1-2 hour meeting once a month to discuss the culture of data within the City/County. ● Provide guidance and input on how the City/County should progress in its effort to encapsulate data into the framework of the organization. ● Contribute to the prioritization of data publication.

	• Contribute to a set of standards around data quality, metadata, and technical capacity. • Help the City/County grow in its analytical capacity through individual BI/data analysis projects. • Promote the need for data literacy within the City/County.

Open Data Program Accountability Chart



3. How to Initialize the Data Release Process

The first major task for all departments is to initialize the release data to the Open Data program. Below, we provided guidance on how to do this. We recognize that this will be a learning process. As a result, we will make changes to this guidance based on your feedback. After the initial data release is completed, this process may continue as new datasets are created, discovered or requested by other parties.

Follow the 3 major steps below to facilitate the publication of your data:

1. Identify data sources
2. Brainstorm and identify datasets in each data source
3. Complete dataset release form (one for each dataset)

Step 1: Identify data sources

Your data may be housed in a variety of places - from inside information systems or databases to shared drives and folders. This can also include 3rd party vendors and data hosted on vendor systems. Step 1 is about identifying the major data sources in your department.

Questions to help identify and discover data sources (available in the Data Brainstorming Canvas):

- What information systems does your department use?
- What databases does your department use?
- What applications capture information or are used in your business processes?

- Are some data resources kept in spreadsheets (on shared or individual drives)?
- What information are we already publishing and where did that information come from?

Steps 2 and 3 Guidance

Below we describe the next 2 steps of the data release process. Visit the Detailed Data Release Guide for Steps 2 and 3 for more detail.

Step 2: Brainstorm and identify potential datasets in each data source

Some of your information sources may be fairly straightforward (e.g. a single sheet in a spreadsheet). In these cases, you have already identified the dataset.

In addition, you may already have a list of datasets you are publishing or plan to publish.

But others, like relational databases, may be very complex. Identifying subsets of the database that could serve as datasets, probably requires some brainstorming. You may want to include data analysts, managerial staff, or any other relevant personnel that can provide valuable insight.

To help brainstorm, use the questions below:

- What data populates your monthly or quarterly reports?
- What departmental data is currently publicly available on the data portal or elsewhere online?
- What data does your department use for internal performance and trend analysis?
- What information is published as a performance metric?
- What data is reported to federal, state or local agencies?
- Talk with your Information Officer - what data has been requested before?
- What data do other departments ask for?
- What kinds of data are similar agencies across the country publishing?

Note: Don't exclude any datasets based on privacy or confidentiality concerns! Our goal is to have a holistic picture of our data. Privacy and confidentiality concerns will be taken care of on a case by case basis during the data release process.

Step 3: Complete the dataset release form

For each dataset you identify in Step 2, complete the release form. Include:

- New datasets (identified via brainstorming)
- Existing datasets, including already published datasets

Appendix A. includes the templates

4. Publishing

Before data is made publicly available on the open data portal, there are a set of checks and balances in place to ensure that data is not published before it is ready to be, and that all datasets have been reviewed by all the necessary parties. That process is as follows:

Uploading Data

Datasets will be uploaded to the Portal by either the Open Data Program Manager, or the Data Steward depending on the level of difficulty and technical expertise required to upload the dataset. The Open Data Program Manager and the Data Steward will work together to identify the best method, and will reevaluate as necessary to best meet the needs of the Open Data Program. The Open Data Program will work with the Technology Solutions department to determine whether a direct connection can be made from the Open Data Portal to the database holding the dataset, or whether the dataset should be uploaded directly to the portal as a spreadsheet.

Note: Uploading a dataset to the portal does not automatically make the dataset public.

Metadata/Data Configuration

According to the John Hopkins University Center for Government Excellence metadata is:

Put simply, metadata is descriptive information about data. Metadata enables visitors find and use published data effectively. Good metadata reduces the need for visitors to seek personal assistance, helps prevent misinterpretation of data, and encourages higher data quality. Without it, a catalog of published data could not exist. Metadata is generally divided into two types:

- Metadata that provides an overview of the data. This kind of metadata helps people find the data through internet searches, while navigating your portal, or even while navigating other data portals which might include your catalog.
- Metadata that provides details about specific parts of your data. This kind of metadata enables people to use your data effectively, by helping them understand the various elements it includes and potential limitations.

The Open Data Program Manager and the Data Steward will collaborate to complete any preparations that need to be made to the dataset before making it public. This means completing an informal quality assessment of the dataset. The Data Steward's subject matter expertise and the Open Data Program Manager's data governance expertise will allow them to ensure that the following quality standards are met:

- The dataset is the most complete, accurate, and current version appropriate for publishing
- The dataset has been spot checked for common errors such as missing and misplaced values.
- Any missing data points are left as null, but the meaning of null is defined in the dataset's metadata.
- Columns are formatted appropriately.
- Metadata is complete, concise, and free of jargon.
- Metadata explain the process used to create the data and summarize any changes.
- Metadata clearly explain any limitations or omissions for each dataset.
- Metadata clearly identify an update frequency and plan.

Data Updates

A successful open data program requires up-to-date data. When submitting the Dataset Release Form, Data Stewards are asked to provide a schedule for keeping data refreshed. Data Stewards are responsible (with 'as needed' assistance from the Open Data Program Manager) for all manual updates to datasets. Whenever possible, Data Stewards will update an existing dataset instead of creating a new dataset.

Nevertheless, Data Stewards may have limited time to monitor the freshness of datasets posted on the portal. The Open Data Program Manager will employ two tactics to help ensure that the Greensboro data portal remains up to date:

- **Active monitoring of dataset freshness.** The Open Data Program Manager will track each dataset on the portal. The Open Data Program Manager will alert Data Stewards when a dataset is identified as out of date and needs to be refreshed.
- **Automation.** For datasets that need to be updated at least once per month or more, the Open Data Program Manager will work with Data Stewards and the TS database administrator to implement automatic updates, where appropriate.

Reviews and Approvals

Once a dataset has been uploaded to the data portal, a set of reviews take place to ensure that the dataset is ready to be made publicly available. This is where personally identifiable information that was captured in the Data Release Form, as well as any other areas of concern are reviewed. That review process is as follows.

1. Data Steward Review & Approval
2. Open Data Program Manager Review & Approval
3. CIO Review and Approval

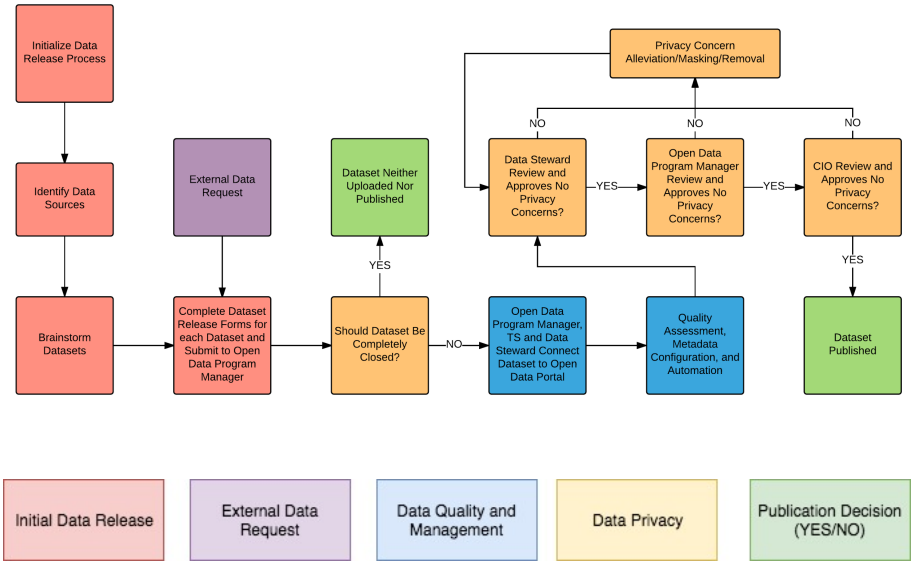
Any changes made to the dataset by one of the reviewers during this process triggers a fail-safe that starts the review process over from the start at the Data Steward. While personally identifiable information was identified in the Data Release Form, other information of concern might arise in the data publishing process. It should be handled as follows:

- **Minor concerns**: such as a column that can be easily removed, are communicated by the Open Data Program Manager, Data Steward or CIO for mitigation.
- **Significant concerns**: represent issues that are not easily resolved but do not immediately disqualify the dataset, such as the possibility that anonymized individuals in a dataset could be easily re-identified when the dataset is combined with another dataset on the portal. Once the reviewing party identifies such concerns, the dataset is forwarded to the Attorney's office for review.
- **Overwhelming concerns**: while overwhelming concerns should be caught before datasets are even uploaded to the Portal, there is a chance that such concerns are discovered in the review process. Overwhelming concerns immediately disqualify a dataset from being posted to open data. Examples include datasets that are protected by law or that would pose a security risk. Datasets identified as overwhelmingly concerning are not uploaded to the portal.

For datasets that show evidence of one or more of these classes, proper risk mitigation techniques will be used, including but not limited to:

- Redaction (e.g., deleting a column of names).
- Anonymization (e.g., reducing the precision of addresses).
- Aggregation (e.g., averaging a dataset by age group).
- Constraining access (e.g., limiting access to city staff).
- Forgoing publication and discarding the dataset.

Workflow



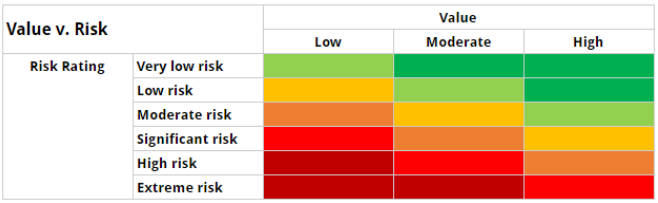
5. Appendices

Appendix A. Data Release Form

The data release form will help streamline the data prioritization and publication process.

- Google: If you have a google account and want to use google, use the Google template (copy the template into your own account - do not edit the document)
- Word: Download word template

Basic Information	
Department	[*Name of department*]
Data Steward	[*Name of department contact*]
Contact details	[*Email and phone number*]
Date	[*Date form finalized*]
Step 1: Dataset and Metadata	
1A. Dataset	[*Name of dataset*]
1B. Relevant fields	(1) *List ALL data fields.*

	(2) List any location/ geo data fields *(3) List any personally identifiable information* *(4) Description* *(5) Source* *(6) Refresh/Automation Schedule*
Step 2: Identifiability Risk Assessment	
2A. Value of publication	Value of publication ❑ Low ❑ Moderate ❑ High
2B. Risk of Publication	Risk of Publication ❑ Very low risk ❑ Low risk ❑ Moderate risk ❑ Significant risk ❑ High risk ❑ Extreme risk
2C. Weigh the value of publication against the risk of publication	 ❑ **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: Risk Response	
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.
Step 4: Personally Identifiable Information (PII)	
4A. Is there PII in this dataset?	❑ Yes ❑ No If "no", skip to Step 5
4B. Should this PII be masked, not included, or reason to not publish the dataset?	❑ Masked ❑ Not included ❑ Reason to not publish the dataset If "masked", proceed to Step 4C, otherwise, skip to Step 5
4C. How should this PII be	*[Describe preferred method for masking PII]*

masked?	
Step 5: Planning	
5A. We plan to revisit the decisions in this form every...	❑ 3 months ❑ 6 months ❑ 1 year ❑ Other _____
5B. Next date for review	[*Insert date*]
Notes	
[*Insert any important notes*]	

Appendix B. Resources & Credits

Below are the origional acknowledgements provided by the City and County of San Francisco. The City and County of Durham appreciates their hard work in putting together this document, and for allowing other municipalities to adopt it for their own purposes.

This guidebook was created using input from a number of resources, including:

Title	Attribution	License
City of Philadelphia data Guidebook	City of Philadelphia, Office of Innovation and Technology	Creative Commons Attribution-ShareAlike 4.0 International license
New York State data Handbook	New York State data Initiative	
data Handbook	Open Knowledge Foundation	
Sunlight Foundation data Guidelines	Sunlight Foundation	

Appendix C. Responding to Open Data Concerns

http://labs.centerforgov.org/open-data/addressing-concerns/