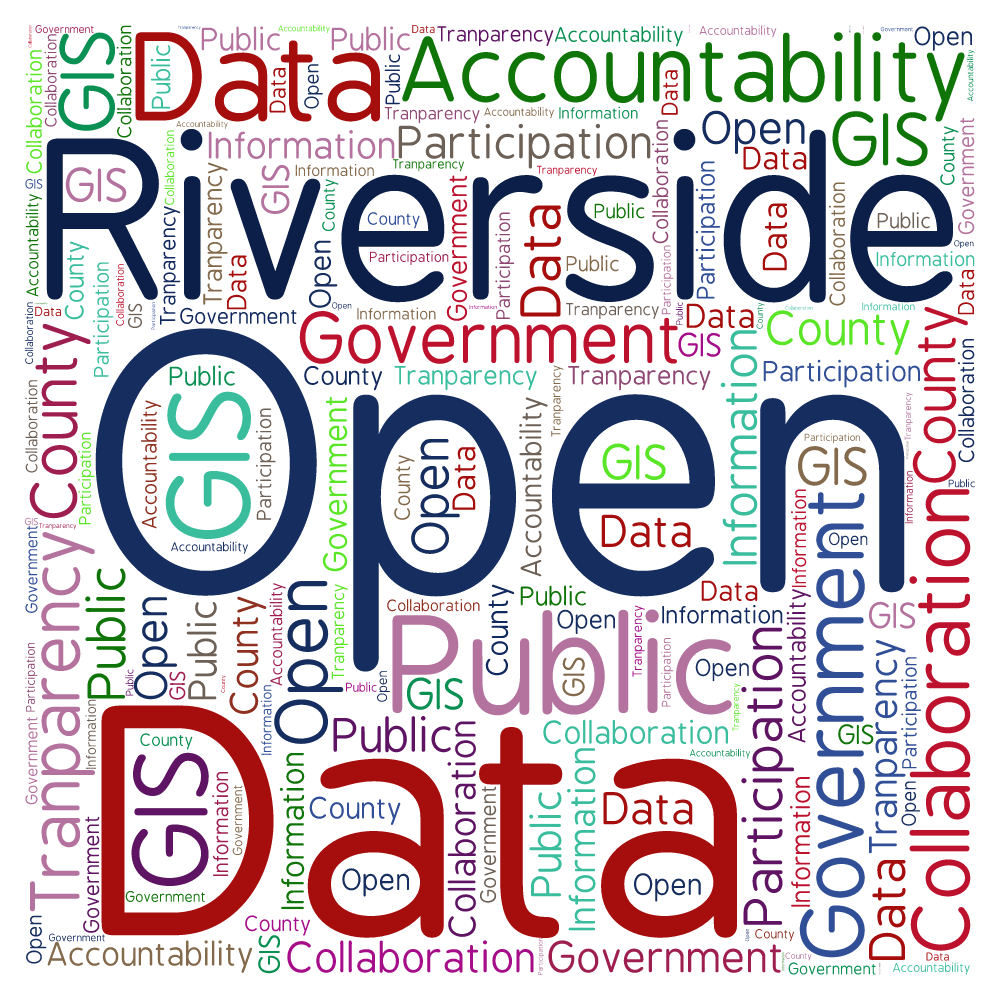
Open Data Handbook



Tom Mullen

Chief Data Officer

Riverside County, California

September 2016

1. **Contents**

[1 Background 4](#_1fob9te)

[1.1](#_3znysh7) Purpose of this Guidebook 4

[1.2](#_2et92p0) Why Release Open Data 4

[1.2.1](#_tyjcwt) Stimulate new ideas and services. 4

[1.2.2](#_3dy6vkm) Increase internal data sharing. 4

[1.2.3](#_1t3h5sf) Simplify responses to public information requests. 4

[1.2.4](#_4d34og8) Reduce unwanted web traffic. 4

[1.2.5](#_2s8eyo1) Changing how we use data. 4

[1.3](#_17dp8vu) RC Data Our Open Data Portal 4

[2 Roles and Responsibilities 5](#_3rdcrjn)

[3 How to Create your Data Catalog (Inventory) 5](#_26in1rg)

[3.1](#_35nkun2) Step 1: Identify data sources 6

[3.2](#_1ksv4uv) Step 2: Brainstorm and identify potential datasets in each data source 6

[3.3](#_44sinio) Step 3: Complete dataset inventory template 7

[4 Dataset Formatting 7](#_2jxsxqh)

[4.1](#_z337ya) Breadth and Depth of Data File 7

[4.2](#_3j2qqm3) Historical Data and Changes in Methodology 7

[4.3](#_1y810tw) Data File Format 8

[4.4](#_4i7ojhp) Use Vertical Rather that Horizontal Orientation 8

[4.5](#_2xcytpi) Header Row 10

[4.6](#_1ci93xb) Column Names 11

[4.7](#_3whwml4) Empty Cells in a Group of Rows 11

[4.8](#_2bn6wsx) Blank, ‘N/A” or Other Unknown Cells 12

[4.9](#_qsh70q) Subtotal or Total Rows or Other Grouped Data 12

[4.10](#_3as4poj) County Fields 13

[4.11](#_1pxezwc) Region Fields 14

[4.12](#_49x2ik5) Coded Fields 14

[4.13](#_2p2csry) Text Fields Must be Trimmed of Whitespace 14

[4.14](#_147n2zr) Numeric Fields (e.g. Money, Measures, Identification Numbers) 15

[4.14.1](#_3o7alnk) Money 15

[4.14.2](#_23ckvvd) Measures (e.g. Ratios, Quantities, Percentages) 16

[4.14.3](#_ihv636) Identification Numbers and Numeric Codes 16

[4.14.4](#_32hioqz) Date Fields 17

[4.14.5](#_1hmsyys) Geographic Co-Ordinates 19

[4.14.6](#_41mghml) Shapefiles Containing Lines and Polygons 20

[4.14.7](#_2grqrue) Shapefiles Containing Points –Converting to Tabular 20

[4.14.8](#_vx1227) Special Guidance for Converting Excel Data to CSV or TSV 20

[4.14.9](#_3fwokq0) Special Guidance for Data Export Programmers 22

[5 Data Publication 23](#_1v1yuxt)

[5.1](#_4f1mdlm) Data Submission Packet 23

[5.2](#_2u6wntf) Internal Review and Approval 23

[5.3](#_19c6y18) Dataset Title 23

[5.4](#_3tbugp1) Dataset Description 23

[5.5](#_28h4qwu) Reviewers 23

[5.6](#_nmf14n) Guidance on Private or Sensitive Data 24

[5.7](#_37m2jsg) Completing the Metadata 24

[5.7.1](#_1mrcu09) Category Required 24

[5.7.2](#_46r0co2) Data Steward Contract Required (private) 25

[5.7.3](#_2lwamvv) Geographic unit (Conditionally required) 25

[5.7.4](#_111kx3o) Tags (Optional) 26

[5.7.5](#_3l18frh) Program Link (Optional) 26

[5.7.6](#_206ipza) Data Notes (Optional) 26

[5.8](#_4k668n3) Data Dictionary 26

[5.8.1](#_2zbgiuw) Document All Fields 26

[5.8.2](#_1egqt2p) List and Define Categories 27

[5.8.3](#_3ygebqi) Include Allowable Ranges for Numeric Data Where Applicable 27

[5.8.4](#_2dlolyb) Define Methodologies where Appropriate 27

[6 Appendices 28](#_sqyw64)

[6.1](#_3cqmetx) Appendix A. Data Catalog Templates 28

[6.2](#_4bvk7pj) Appendix B. Definitions 29

[6.3](#_2r0uhxc) Appendix C. Resources & Credits 30

# Background

## Purpose of this Guidebook

The purpose of this guidebook is to provide guidance on the publication of open data in the County of Riverside and to help Data Coordinators in their new role. We will update this guide as the roles, responsibilities and processes evolve

## Why Release Open Data

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

### Stimulate new ideas and services.

By releasing open data, county departments may help to stimulate new and innovative ideas from the 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 throughout the county.

### 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. The open data portal can provide a platform to share data internally. Combining information from different departments could provide valuable insights into how our county works and how departments may better serve those that live and work in Riverside.

### Simplify responses to public information requests.

Open data releases can be an effective way of responding to requests for information. 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 county. This puts unnecessary stress on the county’s technology infrastructure and unneeded burden on the 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.

## RC Data Our Open Data Portal

RC Data is the County of Riverside’s open data portal and is hosted by a vendor, [Socrata](http://www.socrata.com/). RC Data allows users to find, visualize and use our data to discover new insights into county services or develop novel applications based on our published data.

# Roles and Responsibilities

Below is an overview of the roles and general responsibilities in support of open data. We expect to modify or change these as we learn more and as the roles become more refined.

|  |  |
| --- | --- |
| **Role** | **General Responsibilities** |
|  |  |
| Data Coordinators | Data Coordinators are designated for each department as the main point of contact and accountability for open data in their department. General responsibilities include:   * Inventorying department data sets * Establishing a plan and timeline for publishing them * Ensuring department head or designee approval is obtained prior to publication of data * Providing reports on progress in implementing the open data plan * Serving as a key point of contact |
| Chief Data Officer | The Chief Data Officer is accountable for the county’s overall implementation of the open data policy. General responsibilities include:   * Creating processes, rules and standards to implement the open data policy, including but not limited to:   + Providing guidance and assistance to departments and their Data Coordinators in releasing open data   + Creating data and metadata standards and guidelines * Maintaining the open data website * Assisting departments with analysis of county datasets |
| Data Stewards | Data Stewards are individuals representing data owners of individual databases, datasets, or information systems. In general, a data steward has business knowledge of the data and can answer questions about the data itself. General responsibilities likely include:   * Managing the dataset or source and authorizing changes to it * Managing access to and use of the data, including documentation * Managing accuracy, quality and completeness of the data |
| Data Custodian | Data Custodians are IT professionals that assist with the technical implementation of individual databases, datasets, or information systems. Not all systems or data sources will have a data custodian. General responsibilities likely include:   * Implementing technical changes requested by the data steward/owner * Administration and maintenance for the database or system |

# How to Create your Data Catalog (Inventory)

The first major task for departments and their Data Coordinators is to create a data catalog (or inventory) of your department’s data. Below we provided guidance on how to do this. But we recognize that this will be in part a learning process. As a result, we will make changes to this guidance based on your feedback. After the initial catalog creation, this process may continue as new datasets are created, discovered or requested by other parties.

Follow the 3 major steps below to conduct your data inventory:

1. Identify data sources
2. Brainstorm and identify datasets in each data source
3. Complete dataset inventory template (for each dataset)

Appendix A. includes templates to support this process.

## 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:

* 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?

For each of the data sources:

* Provide a name and brief description of the data sources
* Capture any technical details and point of contacts

Appendix A. includes a template for this process.

* Data Catalog Template

## 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 subject matter experts, data stewards, management team or analysts in this process.

To help brainstorm, use the questions below:

* What data populates your monthly or quarterly reports?
* What departmental data is currently publicly available 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?
* What data is frequently requested by the public?
* What data do other departments ask for?
* What kinds of open data are similar agencies across the country publishing?

**Caution:** Don’t exclude any datasets based on privacy or confidentiality concerns! Our goal is to have a holistic picture of our data. Based on this big picture, we can then decide what we should publish. Step 3 provides a means to capture privacy and confidentiality concerns.

## Step 3: Complete dataset inventory template

For each dataset you identify in Step 2, complete the inventory template. Include:

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

Appendix A. includes a template for this process.

* Data Catalog Template

# Dataset Formatting

Maximizing data discovery, exploration, analysis, reuse, and understanding is greatly influenced by the way data is published and presented. Ever cognizant of the end-users and their reuse of data for trending, research, innovation, and development, format and quality of data is critical (including, but not limited to accuracy, completeness, timeliness, consistency). Standardizing the data publishing model maximizes the utility and interoperability of the data for a multiple range of uses and end- users. Adherence to the standards outlined within this document will result in data that is agile and adaptable, enabling filtering within data files, cross-sector correlations, and integration with third- party visualization and analytic software.

## Breadth and Depth of Data File

As highlighted in the Open Data Handbook, there is high value in the depth and breadth of data (e.g., granularity, breadth of history - years of coverage, etc.). Breadth and depth of data optimizes usefulness, allowing for applied predictive analytics and modeling. Release of data with high information content enhances reuse (internally and externally, and within and across sectors), and the ability to analyze, trend, inform, and generate new and valuable insights.

## Historical Data and Changes in Methodology

Historical data is highly valuable, even when methodologies have changed, providing a look back and informing the path forward.

There are instances when statute, regulation, or polices result in changed thresholds and/or methodologies in how data is collected and reported. Such changes should not eliminate historical data from consideration or publication. If a change in methodology dictates a change in the structure of the information collected or calculation of the data collected this can be accommodated by publishing multiple datasets. What is critical to making this data valuable and understandable to the public is comprehensive documentation for each of the datasets. For example, if a particular dataset’s methodology changed in 2002, the historical data for the years 1970-2001 can be a standalone dataset, capped at the year of the old methodology (with explanatory documentation). A new dataset would display the data beginning with the effective date for the new methodology (e.g., beginning in 2002).

## Data File Format

Only comma-separated value (CSV) and format data files are accepted; ASCII or UTF-8 character encoding is required. While there is no official standard for CSV, the Internet Engineering Task Force (IETF) has published a memo regarding CSV format at [https://tools.ietf.org/html/rfc4180.](https://tools.ietf.org/html/rfc4180)

If your data is currently maintained in **Excel**, ***it must be saved in CSV format prior to publication***. See section 4.24 - Special Guidance for Converting Excel Data to CSV.

## Use Vertical Rather that Horizontal Orientation

Horizontal data orientation should be restructured to vertical whenever feasible. Vertical datasets are more easily understood and sortable, as well as more useful for creating visualizations on the Open NY platform.

**Not Acceptable** – horizontal data orientation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Company** | **Year** | **Skin Disorders** | **Respiratory Conditions** | **Poisoning** | **Other Illnesses** |
| ABC | 2009 | 0 | 1 | 0 | 2 |

**Acceptable** – vertical data orientation

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Year** | **Work Related Injury** | **Number of Cases** |
| ABC | 2009 | Skin Disorders | 0 |
| ABC | 2009 | Respiratory Conditions | 1 |
| ABC | 2009 | Poisoning | 0 |
| ABC | 2009 | Other Illnesses | 2 |

This is especially true for datasets containing data by year, especially numerous years. Years should have their own rows, rather than columns, in the data. Presentation, consumption, ease of utility, and refresh of the data become much more difficult as data files get wider with numerous columns. Moreover, horizontal orientation (as opposed to vertical) restricts ability to create visualizations in a variety of tools and makes it difficult to perform analytics and observe time-based trends in a single view. Vertical data orientation not only makes the data machine-readable, but human readable

**Not Acceptable** – horizontal data orientation - each subsequent year of data requires the addition of another column

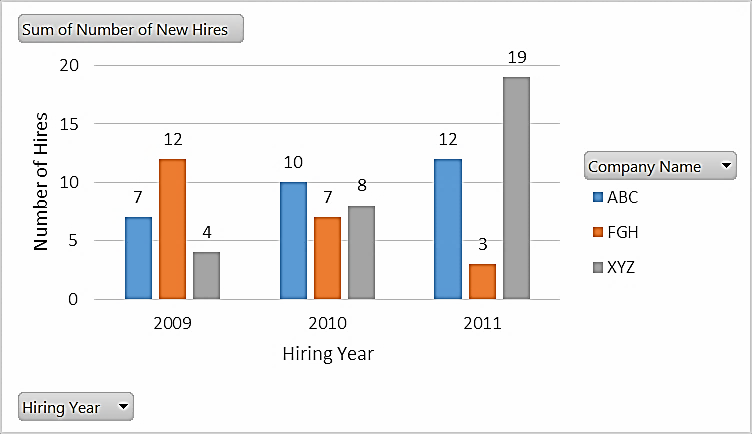
|  |  |  |  |
| --- | --- | --- | --- |
| **Company Name** | **2009 New Hires** | **2010 New Hires** | **2011 New Hires** |
| ABC | 7 | 10 | 12 |

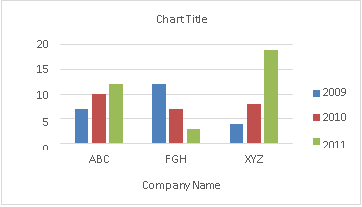
**Acceptable** – vertical data orientation - additional years can be appended to the data file as needed

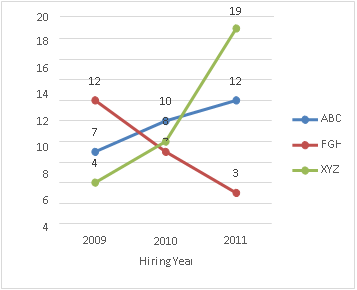
|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Company Name** | | **Hiring Year** | | **Number of New Hires** | |
| ABC | | 2009 | | 7 | |
| ABC | 2010 | | 10 | |
| ABC | 2011 | | 12 | |
| XYZ | 2009 | | 4 | |
| XYZ | 2010 | | 8 | |
| XYZ | 2011 | | 19 | |
| FGH | 2009 | | 12 | |
| FGH | 2010 | | 7 | |
| FGH | 2011 | | 3 | |

Vertical orientation of the submitted data file provides maximum flexibility for reuse and application, as the data can then be sorted by year or to create visualizations with the data rolled up by year.

The benefit of this more vertical orientation is that any grouping of variables is possible facilitating complex analyses of the data. An end-user can utilize any number of third party business intelligence and/or data analysis visualization tools to create pivot tables or graphs to identify patterns and trends over time.







Vertical orientation of the submitted data file also provides maximum flexibility and utility to flip to a horizontal visualization of the data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Sum of Number of New Hires** |  |  |  |  |
|  | 2009 | 2010 | 2011 | Grand Total |
| ABC | 7 | 10 | 12 | 29 |
| FGH | 12 | 7 | 3 | 22 |
| XYZ | 4 | 8 | 19 | 31 |
| Grand Total | 23 | 25 | 34 | 82 |

## Header Row

Data should contain one and only one header row. Multi-row headers are not acceptable.

**Not Acceptable** – multi-row header; column names uppercase

|  |  |  |
| --- | --- | --- |
| **PROJECT** | **CONTACT** | **CONTACT** |
| **NAME** | **NAME** | **TITLE** |
| ABC | John Smith | Project Director |

**Acceptable**- single row header; column names in title case

|  |  |  |
| --- | --- | --- |
| Project | Contact | Contact Title |
| ABC | John Smith | Project Director |

## Column Names

##### Column names must be clear and in plain English, instead of the source system database naming conventions.

##### Do not use underscores in column names.

##### Avoid use of abbreviations, use title case for field names, and be sure that the names match that in the Data Dictionary.

##### Codes should not be used. However, if any codes absolutely must be used, they must be fully explained in the dataset documentation (i.e., data dictionary).

##### Column names should be kept to less than 50 characters in length whenever practicable where shortening will not result in misinterpretation.

**Not Acceptable** – cryptic column names

|  |  |  |
| --- | --- | --- |
| **prjnme** | **ctnme** | **ctttl** |
| ABC | John Smith | Project Director |

**Acceptable** – clear, reasonably long column names

|  |  |  |
| --- | --- | --- |
| **Project Name** | **Contact Name** | **Contact Title** |
| ABC | John Smith | Project Director |

**Not acceptable** – very long and/or unclear column names

|  |  |
| --- | --- |
| **CountyofResidence1YearAgoPopulation1Yearan dOver\_MarginofError** | **CountyofResidence1YearAgoNonmovers\_Estimate** |
| 2052 | 215081 |

**Acceptable** – use of acronyms to shorten long names (fully document in data dictionary)

|  |  |
| --- | --- |
| **COR 1 Year Ago Population >1 Yr MOE** | **COR 1 Year Ago Nonmovers Estimate** |
| 2052 | 215081 |
|  |  |

## Empty Cells in a Group of Rows

A group of rows related to one entity should repeat the entity for all rows in the group.

**Not Acceptable** – empty fields in-group of rows

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Year** | **Work Related Injury** | **Number of Cases** |
| ABC | 2009 | Skin Disorders | 0 |
|  | 2009 | Respiratory Conditions | 1 |
|  | 2009 | Poisoning | 0 |

**Acceptable** – field repeated for all rows in group

|  |  |  |  |
| --- | --- | --- | --- |
| **Company** | **Year** | **Work Related Injury** | **Number of Cases** |
| ABC | 2009 | Skin Disorders | 0 |
| ABC | 2009 | Respiratory Conditions | 1 |
| ABC | 2009 | Poisoning | 0 |

## Blank, ‘N/A” or Other Unknown Cells

Blank fields, when left unexplained, often lead to confusion – particularly when the column is numeric.

* If the blank field represents zero, then the field should be zero.

##### If the blank field represents “not collected” or “unknown”, then this should be explained in the metadata or data dictionary.

Equally important is consistency.

##### If data is numeric, it should not also be presented as text within the same column.

##### Fields containing “N/A”, “-“ or “unknown” *should not be mixed* within a numeric column as this will make sorting and analysis difficult.

See section 4.14 “Numeric Fields” for an example of sorting problems that can occur when numeric data is presented as text format data.

## Subtotal or Total Rows or Other Grouped Data

Avoid including subtotal and total rows unless absolutely necessary. End-users may create a visualization that, unknown to them, includes these subtotal or total rows, skewing the totals in the visualization.

Example of a dataset**, Not Acceptable and Improperly Formatted**, containing subtotal/total rows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **County** | **Year** | **Number of Cases** |
| Capital | Albany | 2009 | 27 |
| Capital | Saratoga | 2009 | 13 |
| Capital | Subtotal | 2009 | 40 |
| Western | Erie | 2009 | 15 |
| Western | Subtotal | 2009 | 15 |
|  | Grand Total | 2009 | 55 |

This results in the Number of Cases being erroneously doubled

|  |  |  |
| --- | --- | --- |
| **Region** | **Year** | **Number of Cases** |
| Capital | 2009 | 80 |
| Western | 2009 | 30 |

If subtotal and total rows must be included to satisfy a specific use case, the data must contain an additional “type” column to provide the user an easy way to filter out subtotal and total rows when necessary.

**Not Acceptable** – no easy way to filter out subtotal and total rows

|  |  |  |
| --- | --- | --- |
| **Area** | **Year** | **Number of Cases** |
| Erie | 2000 | 15 |
| Niagara | 2000 | 3 |
| Western New York | 2000 | 35 |
| Region 2 Subtotal | 2000 | 53 |
| Statewide Total | 2000 | 102 |

**Acceptable** – subtotal and total rows can be easily filtered out

|  |  |  |  |
| --- | --- | --- | --- |
| **Region** | **County** | **Year** | **Number of Cases** |
| Western New York | Erie | 2000 | 15 |
| Western New York | Niagara | 2000 | 3 |
| Region 2 | Subtotal | 2000 | 35 |
| Western New York | Total | 2000 | 53 |
| Statewide | Total | 2000 | 102 |

## County Fields

Standardizing the county field across datasets allows for correlating data across datasets, aggregation across agencies and sectors, and interoperability and linkage. It allows mappings to equivalent elements. This process is facilitated when the values of county fields are consistent among datasets. County fields should only contain the county name. Do not include “County”, “County of”, etc. in county fields. Do not use abbreviations, e.g. “ALBA” for “Albany”.

Where known, an additional column can be included identifying County FIPS (Federal Information Processing Standards) Codes, a.k.a. ANSI INCITS-31-2009 Codes1. This helps avoid inconsistency issues and can provide more reliable correlation. If this code can be made available in addition to county, it should be included as an additional field in a separate column.

**Not Acceptable** – non-standard county fields

|  |  |  |
| --- | --- | --- |
| **County** | **Year** | **Number of Cases** |
| Albany County | 2009 | 27 |
| County of Saratoga | 2009 | 13 |

**Acceptable** – standard county fields, inclusion of FIPS code if possible

|  |  |  |  |
| --- | --- | --- | --- |
| **County** | **County FIPS** | **Year** | **Number of Cases** |
| Albany | 36001 | 2009 | 27 |
| Saratoga | 36091 | 2009 | 13 |

## Region Fields

Region is sometimes a field in agency data, but inconsistency in how agencies delineate their regions can confuse end-users of the data. If an agency dataset contains an agency-specific region, the column heading should indicate the agency whose regions are being referred to, perhaps by including the agency’s acronym. This should be further explained in the data dictionary and/or overview document.

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | **ESD Region** | **Contact Name** | **Contact Title** |
| ABC | Mohawk Valley | John Smith | Project Director |

## Coded Fields

Similarly to column names as described in section 4.6 above, coded fields can be extremely useful for charting or for application developers, but can be easily misunderstood by those not familiar with your data when not adequately explained. If possible, coded fields should be accompanied by a field providing the text equivalent – this enhances the ability to identify patterns and trends and perform rapid sorts and filters on the codes without sacrificing readability for the user. If this is not possible, coded fields should be explained in the data dictionary document.

**Not Acceptable** – unclear coded fields

|  |  |  |
| --- | --- | --- |
| **Project Name** | **Status** | **Phase** |
| ABC Project | A | ST |
| XYZ Project | H | DC |

**Acceptable** – text equivalents accompany coded fields

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project Name** | **Status** | **Status Description** | **Phase** | **Phase Description** |
| ABC Project | A | Active | ST | Project Started |
| XYZ Project | H | On Hold | DC | Design Complete |

## Text Fields Must be Trimmed of Whitespace

Text fields must be trimmed of leading or trailing whitespace (space-padding), otherwise searching, sorting and filtering the data on the Open Data platform will not work as expected. The following VB script can be used in Excel to remove trailing whitespace from all fields. Please note that any CSV or TSV must be properly imported into Excel first, and then saved to CSV or TSV after the edit.

Sub NoSpaces() Dim c

As Range

For Each

c In Selection.Cells c = Trim(c)

Next End

Sub

## Numeric Fields (e.g. Money, Measures, Identification Numbers)

Do not mix text in a field that is intended to contain numeric data. Mixing text and numeric data in the same column will result in the entire column being stored as text. This will severely hamper a user’s ability to analyze the numeric data, and result in unexpected sorting problems since the numbers will sort as text, not as numbers.

Numeric data stored in a numeric column will sort as expected, e.g. 1 before 2, 2 before 3, 3 before 17, etc. Numeric data stored in a text column, however, will sort *“alphabetically”*, e.g. “1” before “17”, “17” before “2”, “2” before “3”, etc. The example below highlights the differences in how sorted numeric data will display when stored as numeric versus text:

|  |  |
| --- | --- |
| Numeric Data Stored as Numeric, Sorted | Numeric Data Stored as Text, Sorted |
| 1 | 1 |
| 2 | 17 |
| 3 | 2 |
| 9 | 3 |
| 17 | 457 |
| 65 | 65 |
| 457 | 9 |

### Money

##### Numeric data that represents money should be provided with either no decimal places or two decimal places. Do not vary the number of decimal places used to format the values throughout the data – consistency is key.

##### Do not include currency symbols, or commas for place-separators.

##### Negative values should be preceded with a minus sign, not placed within parentheses.

**Not Acceptable** – monetary values containing currency symbols, place-separators, varying decimal places, and parentheses

|  |  |
| --- | --- |
| **Project Name** | **Cost** |
| ABC Project | $1,345,231.768 |
| XYZ Project | (654,213.5) |

**Acceptable**- monetary values containing no decimal places or two decimal places; negative values expressed with leading minus sign.

|  |  |
| --- | --- |
| Project Name | Cost |
| ABC Project | 1345231.77 |
| XYZ Project | -654213.50 |
| DEF Project | 458374 |

### Measures (e.g. Ratios, Quantities, Percentages)

Varying decimal places are acceptable. Do not include commas for place-separators. Negative values should be preceded with a minus-sign, not placed within parentheses.

**Not Acceptable** – numeric values containing place-separators and parentheses

|  |  |
| --- | --- |
| **Project Name** | **Performance Ratio** |
| ABC Project | 6,213.891 |
| XYZ Project | (4,165.271) |

**Acceptable** – numeric values containing no place-separators and negative values expressed with leading minus-sign

|  |  |
| --- | --- |
| **Project Name** | **Performance Ratio** |
| ABC Project | 6213.891 |
| XYZ Project | -4165.271 |

### Identification Numbers and Numeric Codes

Identification numbers and numeric codes such as SWIS codes, FIPS codes, DOS ID numbers, etc., where leading zeroes are significant, must be provided as text values to prevent the leading zeroes from being truncated. This can be a problem particularly when Excel is used to prepare the data, as Excel defaults to the “General” format for columns and assumes that any value that looks like a number must be a quantity, resulting in truncation of leading zeroes.

**Not Acceptable** – identification number provided as number type data (leading zeroes truncated)

|  |  |
| --- | --- |
| **SWIS Code** | **Municipality** |
| 10100 | Albany |
| 10300 | Cohoes |

**Acceptable** – identification number provided as text type data (leading zeroes retained)

|  |  |
| --- | --- |
| **SWIS Code** | **Municipality** |
| 010100 | Albany |
| 010300 | Cohoes |

### Date Fields

#### Full Dates

*Full dates must be provided in MM/DD/YYYY format****.***

*Example: 09/02/2013*

*The importance of standardizing this format is that this is the only way to display trends over time. It is critical for conducting analyses, time series, and inform decision-making.*

#### Month, Year

Full dates are much more preferable to month, year for analyzing trends over time and should be provided any time the source system can support it. If only monthly data is available, the next best option is to provide a full date set to the last day of the month, e.g. 09/30/2015 – the display can be masked to show only month and year while still retaining the ability to trend.

If full dates cannot be provided, express the year and month in two separate numeric columns: year (YYYY) and month (MM). Month names (“October”, “Oct”) can be included in addition to the numeric month.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Project Name** | **Contact Name** | **Contact Title** | **Start Year** | **Start Month** |
| ABC | John Smith | Project Director | 2013 | 10 |

Or

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Name** | **Contact Name** | **Contact Title** | **Start Year** | **Start Month** | **Start Month Name** |
| ABC | John Smith | Project Director | 2013 | 10 | October |

If only month and year is available and this data lends itself to trending, agencies should consider adding a concatenated year and month column to help make this possible.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Project Name** | **Contact Name** | **Contact Title** | **Start Year** | **Start Month** | **Start Year-Month** |
| ABC | John Smith | Project Director | 2013 | 10 | 201310 |

#### Time Only Fields

There are two acceptable formats for time fields:

##### Military time, a.k.a. 24-hour time, in HH:MM:SS or HH:MM format. Midnight in 24-hour time is 00:00 and noon is 12:00.

Acceptable – Military time is available

|  |  |
| --- | --- |
| **Event Type** | **Event Time** |
| Departure | 08:45:36 |
| Arrival | 16:20:25 |

##### 12-hour time, in HH:MM:SS AM PM or HH:MM AM/PM format.

Acceptable – Military time is not available

|  |  |  |
| --- | --- | --- |
| **Location Name** | **Open** | **Close** |
| Acme Office Supply | 08:00 AM | 06:00 PM |

Please note that this guideline is for time columns separate from date columns. For combined date and time columns, see section 4.14.4.4 below.

#### Combined Date and Time

When the full date and time is available, it should be provided in a single field.

* Military time, a.k.a. 24-hour time, in HH:MM:SS or HH:MM format. Midnight in 24-hour time is 00:00 and noon is 12:00.

Acceptable – Military time is available

|  |  |
| --- | --- |
| **Event Type** | **Event Date and Time** |
| Departure | 03/17/2015 08:45:36 |
| Arrival | 03/17/2015 16:20:25 |

#### 12-hour time, in HH:MM:SS AM/PM or HH:MM AM/PM format.

Acceptable – Military time is not available

|  |  |
| --- | --- |
| **Event Type** | **Event Date and Time** |
| Departure | 03/17/2015 08:45:36 AM |
| Arrival | 03/17/2015 04:20:25 PM |

#### Zip Codes

Five-digit or nine-digit Zip Codes are acceptable. Consistency within a dataset is critical. Nine-digit Zip Codes can be provided as either:

* hyphenated, e.g. 12345-9876

##### non-hyphenated, e.g. 123459876

**Do not mix both formats within the same dataset.** Either one or the other should be used, but not both.

**Not Acceptable** – inconsistent formatting of nine-digit Zip Codes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facility Name** | **Street Address** | **City** | **State** | **Zip** |
| XYZ Medical Care Center | 803 Genesee Street | Rochester | NY | 14611-0123 |
| ABC Incorporated | 654 Main Street | Rochester | NY | 146110123 |

**Acceptable** – consistent formatting of nine-digit Zip Codes

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facility Name** | **Street Address** | **City** | **State** | **Zip** |
| XYZ Medical Care Center | 803 Genesee Street | Rochester | NY | 14611-0123 |
| ABC Incorporated | 654 Main Street | Rochester | NY | 14611-0123 |

#### Phone Numbers

Phone numbers must include area code. Area codes are mandatory. Most important is **consistency**. Whether the phone number is expressed as 518-000-1111 or 5180001111 or (518)-000-1111, do not mix formats within the same column. Consistency with a single format within a single dataset is critical.

#### Address Data

Clean address data is very valuable as it can add another dimension to your data – geographic locations that can be mapped. The geographic location of an address can be the most sought after by app developers (e.g., locations of farmers’ markets, wineries & breweries, etc.).

To ensure accurate mapping, addresses must be broken into four columns: street address, city, state and zip code. If latitude and longitude have not been provided, then a complete and clean address must be submitted for the platform to auto-generate latitude and longitude for those datasets which lend themselves to mapping.

If geocoding and mapping of address data is desired, then street address data must not contain P.O. Boxes or other information that cannot be accurately geocoded such as “corner of”, “in front of”, “across from”, etc. If this is the case then latitude and longitude should be provided.

**Not Acceptable** – not formatted for mapping

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facility Name** | **Street Number** | **Street Address** | **City** | **Zip** |
| XYZ Medical Care Center | 803 | Genesee Street | Rochester, NY | 14611 |

**Acceptable** – formatted for mapping

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facility Name** | **Street Address** | **City** | **State** | **Zip** |
| XYZ Medical Care Center | 803 Genesee Street | Rochester | NY | 14611 |

### Geographic Co-Ordinates

Easily adaptable for mapping are geographic coordinates (latitude, longitude) **specified in decimal degrees**. For ease of use, coordinates specified in other systems (e.g. UTM, degree minute-second, etc.) should be converted to decimal degrees if map visualizations are required. If geographic coordinates are available, they should be provided as separate columns.

If UTM coordinates are given, the documentation must indicate that New York State is in zone 18, and the columns should be named “UTM Northing” and “UTM Easting”.

|  |  |  |
| --- | --- | --- |
| **Facility Name** | **Latitude** | **Longitude** |
| XYZ Medical Care Center | 42.944116999999 | -76.446226899999 |

### 

### Shapefiles Containing Lines and Polygons

The Open Data Platform supports mapping complex information such as lines (e.g. bus routes) or polygons (e.g. Census Blocks). This data must be provided in Shapefile format. To map multiple layers, each layer should be stored as a separate Shapefile.

### Shapefiles Containing Points –Converting to Tabular

Shapefiles containing point data is often rich sources of tabular data that can be mapped and analyzed on the platform. The .DBF file included in Shapefiles can be opened using Excel and the data exported as CSV, just like other sources of tabular data. In most cases, Shapefiles containing only point data will render better as maps and be more usable when converted to tabular CSV format.

### Special Guidance for Converting Excel Data to CSV or TSV

Excel files (xls, xlsx) will not be accepted since they can contain features that cause the import to fail such as merged cells, macros, data spanning tabs, and formulas. This section provides guidance for preparing data currently maintained in Excel into a CSV or TSV format. Excel may be used to generate CSV and TSV file formats; however, care must be taken to ensure that cell formats within Excel do not corrupt the data.

When importing into Excel, format the cells of the blank workbook in the ‘Number’ tab of the ‘Format Cells’ menu from the default value ‘General’ to ‘Text’ format; this ensures that the cell is displayed exactly as entered. Without this change in formatting, numeric content entered may be subject to interpretation, typically as a date, by Excel. For example, if ’11-1961’ is entered into a ‘General’ format cell in Excel, it will appear as ‘Nov-61’, since Excel recognizes this as a date format. If the cell is formatted as ‘Text’, then ’11-1961’ displays as entered.

#### Black Rows and Columns

In preparing data for publication using Excel, *it is critical to check for and remove any inadvertently created blank rows or columns*. Excel will allow for blank rows and columns at the end of datasets if you click outside the active data and save it. That is, if you click outside the border of the active data and save it as CSV, Excel will read these blank rows and columns as part of the data and save the blank rows and/or columns. This may not appear to be part of the spreadsheet but will, in fact, result in blank rows and columns being included when saving as CSV.

***Care must be taken to ensure that any blank rows and columns have been removed prior to creating your CSV files.***

An easy way to determine whether such blank rows or columns are present in Excel is to press [Ctrl]+[End] inside the spreadsheet and see if this takes you beyond your data in the spreadsheet. If it does, the data file contains blank rows and columns that need to be deleted. Delete the blank rows and columns, save the Excel file, and then press [Ctrl]+[End] again – this should now take you to the last row and column in your data.

#### Merged Cells

Merged cells are not acceptable, and cannot be reproduced in a CSV or TSV.

**Not Acceptable** – merged cells

|  |  |  |
| --- | --- | --- |
| **Project Name** | **Contact** | |
| **Name** | **Title** |
|  |  |  |
| ABC | John Smith | Project Director |

**Acceptable** – single cells

|  |  |  |
| --- | --- | --- |
| **Project Name** | **Contact Name** | **Contact Title** |
| ABC | John Smith | Project Director |

#### Empty Rows and Empty Columns

Empty rows and empty columns among the data is not acceptable. Empty rows adversely impact sorting and analysis, and empty columns adversely impact initial import and refresh

* Care must be taken when importing data from, for example, reports which may have blank rows. If blank rows are present the data should be cleansed to ensure that any blank rows and columns have been removed prior to creating your CSV files.

**Not Acceptable** – Empty Rows and Empty Columns

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Company** | | **Year** | |  | | **Work Related Injury** |
| ABC | | 2009 | |  | | Skin Disorders |
| ABC | 2009 | |  | | Respiratory Conditions | |
| ABC | 2009 | |  | | Poisoning | |
|  |  | |  | |  | |
| XYZ | 2009 | |  | | Skin Disorders | |
| Stat Job | 2009 | |  | | Respiratory Conditions | |
| XYZ | 2009 | |  | | Poisoning | |

#### Calculated Fields

These data fields should be expanded to include each data component especially when the creation of visualizations will rely upon this data. For example, if two figures were added together to create a summary value, you should include three columns: one for the first added value, a second column for the second added value, and a third column for the sum of the two.

**Not Acceptable** – calculated field missing components

|  |  |  |
| --- | --- | --- |
| **Facility Name** | **Year** | **Total Tax** |
| XYZ Medical Care Center | 2013 | 23784 |

**Acceptable** – components included with calculated field

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Facility Name** | **Year** | **Assessed Value** | **Rate** | **Total Tax** |
| XYZ Medical Care Center | 2013 | 297300 | 9.00 | 26757 |

#### Multiple Data Items in a Cell

A cell may contain only one item of information; multiple lines within a cell will cause the Import process to fail. Cells that contain collections of data are impossible to evaluate and could cause problems for end-users of your data.

**Not Acceptable** – collection of data in a field

|  |  |
| --- | --- |
| **Project Name** | **Status** |
| ABC | 1/3/13 – Started  1/12/13 – Design Complete 1/20/13 – Development Complete |

**Acceptable (Horizontal)** – one item of information per field

|  |  |  |  |
| --- | --- | --- | --- |
| **Project Name** | **Started** | **Design Complete** | **Development Complete** |

### Special Guidance for Data Export Programmers

#### Commas, Backslashes and Quotation Marks

Commas and quotation marks have significant meaning in CSV files (comma-separated value files). Commas indicate the separation between field values and quotation marks indicate where text values begin and end (particularly important when a text value itself contains an embedded comma).

But what happens when the text value itself contains an embedded quotation mark? Export utilities provided by database platforms typically handle this well, but if the export program is being developed by the agency it is important to know how to handle this situation.

To signal that a quotation mark is a part of the text value and not an indicator of the beginning or end of a text value, you must immediately precede the quotation mark with a quotation mark, and surround the text value with quote marks. Here are some examples:

|  |  |
| --- | --- |
| **Text Value** | **Export As** |
| This is some "quoted" data | "This is some ""quoted"" data" |
| "This" is some quoted data | """This"" is some quoted data" |
| This is some quoted "data" | "This is some quoted ""data""" |

The backslash is an escape character, which indicates that the next character has some special meaning (e.g. “\n” is not the letter “n”, but is the newline character).

#### Blank/Null Values

Every column must be accounted for in a CSV or TSV, regardless if the source value is blank or null for a particular row. That is, if a dataset consists of ten columns, every row in the dataset must contain ten columns. This is accomplished in a CSV or TSV file by including the separating commas or tab characters with nothing in between.

# Data Publication

Now that you have created your data catalog and formatted your datasets, you are now ready to start publishing your data.

The following is the data submission and publication process for publishing data to the county’s open data portal. Data Coordinators will work internally with departmental business units to identify, describe, prepare, obtain approval and publish data to the open data portal.

## Data Submission Packet

A data submission template has been created to help with organizing the necessary information that will be published to the open data portal. We have organized this guide to match the order within the template.

* Excel: Data Submission Template

## Internal Review and Approval

Before submitting your dataset for publication, you must obtain internal approval from your department head or his/her designee. This helps ensure compliance with security, privacy and confidentiality laws as well as ensuring that your department leadership is prepared for the release of your department’s information. To facilitate internal approval, please provide the following:

## Dataset Title

Make sure your title is descriptive and as concise as possible.

* Spell out acronyms or avoid jargon if at all possible
* Include date ranges if appropriate

## Dataset Description

Make sure your description addresses the following points where appropriate.

* Who collects the data or who is the source of the data (business unit or role classification)
* How is the data collected
* Why is this data collected and what is it used for
* What are the programs or policies that this data supports

## Reviewers

A list of the department contacts that should review the proposed publication. This includes the Data Coordinator and the department head or designee.

As needed, flag fields/columns that have privacy issues or are sensitive. During implementation, we will work together to figure out the best way to address these concerns. Read below for additional guidance.

## Guidance on Private or Sensitive Data

* If your dataset has private or sensitive data, there are several ways to mitigate the risk. The table below provides an overview of the basic methods.
* During the process of reviewing the packet for publication, we will work with you on implementing the appropriate method. So at this stage, focus on flagging areas of concern - not how to implement it.

|  |  |  |
| --- | --- | --- |
| **Method** | **What it is** | **Best for** |
| De-identification: Remove the privacy implicating columns. | The simplest way to avoid any privacy issues is to simply not publish the columns that include private data. For example, if a dataset is a list of users and includes their name, address or other information, you can simply remove those columns from the dataset. | Datasets that include individual information that is not necessary for consuming and understanding the data. |
| Aggregation: Summarize the data based on a data property. | Sometimes de-identifying the data is not sufficient. Your data might need to be aggregated either by geography or some other factor such as a category in the dataset. | Datasets where the individual records pose privacy risk even if the identifying columns are removed. A common example of this is health related data.  If the individual records (rows) are important to publish, use one of the other methods. |
| Obfuscation: Mask or transcribe the data. | Obfuscation can happen in a number of ways but a common case is with address data. Sometimes we want to retain a proxy of the address without aggregating the data. | Datasets where the information in the column is important to retain in some form. A common example is street address. Street addresses can be related to zip codes, the census block or other small geographies. This obscures the address without losing meaningful geographic information. |

## Completing the Metadata

As part of publishing data, you must complete the metadata (data about your dataset) requirements described below.

### Category Required

The category field groups similar datasets together regardless of who is publishing the data. The category is a drop down menu that contains the list of categories in the table below. If your dataset falls into multiple categories, select the best fit. You will be able to use tags to add additional ways to categorize the data.

|  |  |
| --- | --- |
| **Category** | **Description** |
| Economy and Community | Data on elections, businesses, community and economic development |
| Administrative and Fiscal Services | Data on finance, budgets, taxes, employees, internal county services |
| Transportation | Data about public transit, biking, driving, parking, and flying |
| Public Safety and Emergency Management | Data on crime, fire, accidents, enforcement and emergency management |
| Health and Social Services | Data about public health, mental health, social services, animal services, veteran services, |
| Geographic Locations and Boundaries | Geographic data on city locations and boundaries; Available as shape files, jsons, kml |
| Land Use and Environment | Data on permitting, construction, land use, sustainability and environmental conditions, building inspections, code enforcement, flood control, waste management |
| Culture and Recreation | Data on arts, museums, parks and open spaces, events |

### Data Steward Contract Required (private)

Data Stewards are individuals in charge of individual databases, datasets, or information systems. In general, a data steward has business knowledge of the data and can answer questions about the data itself. General responsibilities likely include:

* Managing the dataset or source and authorizing changes to it
* Managing access to and use of the data, including documentation
* Managing accuracy, quality and completeness of the data

You must supply the name and email address of the Data Steward. This information will be private; meaning only internal county staff working on the open data program will be able to see this.

### Geographic unit (Conditionally required)

This field is important to help users find geographic data at different scales and levels of specificity. You may have multiple geographic units included in your data. Please pick the one that provides the more granular designation. For example, you may indicate the supervisorial district as well as an address. The address is the more precise field and should be indicated as your geographic unit. Your options are (mostly in order of specific to general):

* Latitude/longitude
* Street address
* Intersection or street segment
* Parcel (Assessor Parcel Number - APN)
* Census block
* Census block group
* Census tract
* Supervisorial District
* Zip code
* City
* Raster
* Other

### Tags (Optional)

Tags (or keywords) are short ways to categorize your dataset. These will be available search terms for users so creating tags that are short and descriptive is ideal. You can also use common categorization schemes like program or division names within your department. These are most useful if you use them consistently across datasets in the department. Over time, we may come out with more specific guidance, but for now two to five keywords should be plenty. These must be comma separated.

### Program Link (Optional)

If you have a program website related to the collection of the data, you can indicate that here. Often, departments will have a programmatic website that describes the collection of data and methodologies in more detail.

### Data Notes (Optional)

If you have additional things you want to note that don’t belong in the data description or elsewhere, you can put that here. This will likely be a little used field, so don’t feel obligated to use this.

## Data Dictionary

The data dictionary is one of the most important parts of documentation for your dataset. This document provides all of the information a user will need to interpret a dataset without you. This is good practice as it helps reduce the number of inbound questions and makes sharing data much easier. Generally, you’ll develop this once, unless the underlying structure of the data changes

### Document All Fields

Regardless of how “simple” or “self explanatory” a field is, there should be documentation for every field. At a minimum the definition should have:

* the column name
* the column type
  + category
  + text
  + numeric
  + boolean (true/false)
* a short description

For Example

|  |  |  |
| --- | --- | --- |
| Column Name | Type | Description |
| PdId | text | This is a unique identifier for the incident and is the concatenation of the incident number and the incident type. |
| IncidntNum | text | Identification number for crime incidents; for a single incident, there can be multiple records. For each incident there will be at least a primary record and may be additional. |
| Category | category | This is the incident category, which can be one of x values. Please see the attached list of categories for additional information. |

### List and Define Categories

If you have a categorical field that has a controlled list of options, you should list those options and the relevant definition for those options within your data dictionary. For example, you may have a list of types for your data that may or may not be self explanatory. Regardless, you should list these out with their definitions and reference this in your data dictionary.

### Include Allowable Ranges for Numeric Data Where Applicable

If you have numeric data, please include the min and max expected values. For example, if the field is a percentage, include that and note the allowable range. If the expected value is on a likert scale, include something like values range from 1 to 5, 1 being strongly agree and 5 being strongly disagree.

### Define Methodologies where Appropriate

If you have a more complicated methodology than can be explained in the general data description, please describe it in your data dictionary. For example, surveys that involve observation and special collection methods should be described in appropriate detail so the user understands the nature of the data. It is also appropriate to reference existing methods documentation via a link if you’ve already documented on a website for example. The main purpose is to make sure the user can easily ascertain how the data was created.

# Appendices

## Appendix A. Data Catalog Templates

The data catalog includes drop down menus and other data tips to make it as easy as possible to complete your inventory. The data inventory template is available on the open data portal.

* Excel: Data Catalog Template

Data Catalog: Fields and Descriptions

The data catalog templates include the following fields. The spreadsheet version includes data validation, including controlled vocabulary and other data checks.

|  |  |
| --- | --- |
| **Field** | **Description** |
| Department/Division | What department collects and manages the data as an official record? |
| Dataset name | Brief descriptive name for the dataset |
| Brief description of data | Include a brief description of the dataset. What is the purpose? What is it used for? Include key data fields if possible. |
| Data Steward (Business knowledge) - Name | Who manages the data and is responsible for making changes to the data? Who understands what the dataset includes and can answer questions about it? |
| Data Steward (Business knowledge) - Email | Please enter the email of the Data Steward. |
| Data Custodian (Technical knowledge) - Name | If applicable, include a contact who manages the technical execution of the database (IT professional). |
| Data Custodian (Technical knowledge) - Email | Please enter the email of the Data Custodian. |
| Data source | What information system or database contains the data? Or what shared server or shared drive contains the data? |
| Start date | How far back does the data go? Use format (MM/DD/YYYY) |
| End date | When does the data end? Use format (MM/DD/YYYY). If the data is still being updated, use "Current". |
| Geographic coverage | Does this data cover the whole county or a subset such as Supervisorial District? |
| Geographic granularity | What is the lowest level of geography in the data? For example, if the data is collected by address, it would be Street Address. |
| Frequency of data change | At what rate does the information in the dataset change? Daily, weekly, monthly, yearly? |
| Number of records | How many records or entries does the dataset include? |
| Format | What format is the data in? e.g. excel, SQL, oracle database, pdf, word, etc. |
| Existing ETLs | Are there existing database connections or extractions? Your Data Custodian may be aware of any existing processes? If you do not know, simply state ‘Unknown’. |
| Existing publication | Is this data already published (made publicly available) in some form or another, e.g. a report. |
| Link to existing publication | If you answered yes to the previous question, please include a link to the report or document. |
| Priority/value | What is your sense of the relative value in publishing this data?   * High - Existing and ongoing requests for this data; this data addresses pressing information needs or pain points (internally or externally); or we have heard compelling examples of how this data could be used * Medium - This data may be useful for other departments or for people external to the county; we occasionally receive requests for this information; or we have heard some examples for how this data could be used * Low - This data has unclear value for either the public or other county departments; we have never received requests for this data; or we have never heard a use case for this data |
| Priority/value comments | Any details or comments to add to your response on priority? |
| Technical challenges | What technical challenges, if any, do you anticipate in publishing this data? |
| Data Classification | How would you classify this data?   * Public - this data could be publicly disseminated without any concerns * Protected - this data is protected by law or regulation and can only be shared or accessed internally and per organizational procedures; OR this information includes individually identified information * Sensitive - in its raw form, this data poses security concerns, could be misused to target individuals or poses other concerns. |
| Protected - Details | If you marked “Protected” for Data Classification, please indicate what law(s)/regulation(s) protect this data. |
| Sensitive - Details | If you marked “Sensitive” for Data Classification, please describe your concerns |
| Data Quality | Do you have concerns about the quality (accuracy) of this dataset? |
| Data Quality - Details | Please describe your concerns with the quality of this dataset. |

## Appendix B. Definitions

|  |  |
| --- | --- |
| **Term** | **Definition** |
| Dataset | Contents of a single database table, worksheet or defined view; data is provided as a single combination of unique rows (or records) and corresponding columns (or fields) describing that row  Example - Database: A database may contain several data tables - each data table constitutes a dataset. However, you could also create new datasets by combining data from different tables into a new table. |
| Data Schema or Standard | Specification that defines the structure of the data (i.e. required data elements and types and supporting definitions) |
| Data source | Technology or system that stores data, including databases, named spreadsheets, information systems, business applications, etc. |
| ETL | Extract, Transform, Load - three database functions that are combined into one tool to pull data out of one database and place it into another.   * Extract - process of reading data from a database. * Transform - process of converting the extracted data from its previous form into the form it needs to be in so that it can be placed into another database. (rules, lookup tables, combining with other data) * Load - process of writing the data into the target database. |
| Geospatial data | Data related to the position of things in the real world, including boundaries or locations |
| Metadata | Descriptive information about a dataset |
| Tabular | Data that is presented in columns or tables |
| Taxonomy or Category | Methodology by which items or datasets are classified or grouped under a similar theme or topic |

## Appendix C. Resources & Credits

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

|  |  |
| --- | --- |
| **Title** | **Attribution** |
| [DataSF Guidebook: Data Coordinators Edition](http://datasf.org/resources/) | City and County of San Francisco |
| [Open Data Field Guide](https://www.socrata.com/open-data-field-guide/) | Socrata |
| [New York State Open Data Dataset Submission Guide v3.0](https://data.ny.gov/download/c3zp-wr9j/application/pdf) | New York Office of Information Technology Services |
|  |  |