NEW YORK STATE

OPEN DATA

DATASET SUBMISSION GUIDE

v3.0

OPEN NY

EMPOWERING CITIZENS WITH OPEN DATA FOR THE DIGITAL AGE

Transparency, Accessibility, Innovation

Quality and Operational Excellence

Together these are the foundation driving our performance

NEW YORK STATE OF OPPORTUNITY. | Office of Information Technology Services

Table of Contents

1. Introduction

The New York State Open Data Initiative

In March 2013, Governor Andrew M. Cuomo signed Executive Order 95 laying the foundation for a more open, transparent, and innovative state. At the same time, Governor Cuomo introduced Open NY, an award-winning initiative of policies, programs, and tools that provide public access to digital data for collaboration and analysis.

Open NY (https://data.ny.gov/) puts tools for transparency, accountability, and innovation directly into the hands of New Yorkers and people all around the world by providing centralized access to machine-readable data to explore, search, download, analyze, reuse, and share in ways never before possible. Impressive in growth and quality content, New York State has become a trusted and respected leader in Open Data, having received national and international recognition for its *Quality by Design* Open Data portal.

As highlighted in the NYS Open Data Handbook:

- New York State is committed to proactively releasing publishable state data.
- New York State is committed to making publishable state data openly and freely available in accessible formats for the public to reuse and consume.
- New York State is committed to publishing high quality data with comprehensive metadata and documentation to foster interoperability and maximize citizen understanding of the data.
- New York State is committed to ongoing and continuous publication of publishable state data.

2. Purpose of this Dataset Submission Guide

Quality, Standardization, Interoperability, Analytics

New York State's focus on data quality has been a hallmark of Open NY. This document is intended to be read together with the NYS Open Data Handbook, (http://ny.github.io/open-data-handbook/OpenDataHandbook.pdf), and includes best practices garnered from lessons learned regarding optimal formatting and documentation. This guide represents a commitment to continuous quality improvement to maximize understanding, and the advancement of standardization to promote interoperability, analysis, and utilization of the data.

Our core ethos, "*Quality by Design*," was born from the recognition that it isn't enough to simply publish raw data. While data can be a catalyst for innovation, value is derived from *quality* data. There is a penchant to ascribe value to the sheer volume of data captured. Realizing meaningful value from data, however, is much more complex. Relevant, timely, consistent, and accurate data are an expectation and not achieved serendipitously. Open Data is about getting data into the hands of all and giving it context - making it easily consumable, understandable, and usable. The guidance herein institutionalizes the processes and protocols to harness the value of data published on data.ny.gov.

Since the launch of data.ny.gov, in March 2013, New York State has intentionally set high standards to maximize the public's understanding and reuse of the data, making sure the data is as accurate and well documented as possible. In furtherance of this objective, Open NY enriches raw content with context to promote research, innovation, and accessibility. The consumption and discovery of data is no longer a passive nor discrete event; government and the public have become active participants. Open NY is a global platform - context provides the diverse community of users with the greatest possible understanding of the data to maximize its discovery, utility, reuse, and derivative value.

Open data is empowering citizens and government with data for the digital age. It is the new infrastructure and digital fuel of the 21st century. Open data is a gateway to advancing data discovery, data standardization, interoperability, and analytics.

3. Components of a Dataset Submission

As has been the requirement since the launch of Open NY in March 2013, each dataset submitted for publication must consist, at a minimum, of the following: a data file, a metadata form, a data dictionary, an overview document, and a signed approval form. Each of these elements is required for each individual dataset prior to publication on Open NY.

Additional documentation may also be required to further end-users' understanding and may include, but is not limited to: survey instruments, data collection tools, a research benefits narrative, supplemental links, studies or reports specific to the data, etc. Supporting documentation maximizes understanding and enhances reuse. Attaching metadata to a data file (such as frequency, coverage, category, and other fields) promotes discoverability and optimizes the ability to develop links with relevant data from different sources.

Figure 1, below, provides a brief description of each component and their purpose/importance, with each fully explained and detailed in the sections that follow.

Figure 1 - Components of a Dataset Submission

Components of a Dataset Submission		
Component	**What is it?**	**Why is it important?**
Data File	Tabular or geo-spatial data. Data must be machine-readable, and formatted according to uniform	Data is one of the State's most strategic assets. Data in an accessible format allows it to be readily used and re-used by citizens, businesses,

	technical standards for import to data.ny.gov	researchers, journalists, developers, government, etc., to process, trend, innovate, and inform.
Metadata Form	Metadata provides important structural and contextual information about the data; it describes characteristics and attributes of the data (e.g., who, what, where, why, how).	Metadata makes finding content and data faster and easier. Metadata facilitates data discovery and linkage across relevant and different data sources.
Data Dictionary Document	The Data Dictionary defines/explains the individual columns that comprise each dataset.	The Data Dictionary is critical in that it clearly and consistently defines the columns and the characteristics of the data elements contained within the columns. A consistent, standardized vocabulary helps provide the end-user with the means to fully understand the information contained within the dataset.
Overview Document	The Overview document provides key information about the data. It contextualizes the data and explains what it represents.	The Overview document provides the background and context of the data. It allows end-users unfamiliar with an agency and/or its data a comprehensive understanding of the data, the agency division which collects and maintains it, the data methodology, statistical and analytical issues, and any limitations regarding the use of the data.
Approval Form	A completed approval form signed by all required parties must be submitted with each dataset prior to being published onto data.ny.gov. Agencies may determine additional internal approvals and signatures are required, and should include such additional persons in their review and sign off process.	Agencies are responsible for certifying that their content has been approved for publication by appropriate agency personnel, including confirming compliance with all laws, rules, and regulations related to confidentiality, privacy, security, intellectual property rights, and the Freedom of Information Law (FOIL).
Additional documentation as determined by Agency or as required to maximize understanding of the data	This may include a survey instrument, data collection tool, study or report specific to the data, explanatory documentation for complex datasets, description of research benefits, etc.	Additional documentation can be useful to end-users (e.g., researchers, others) to, for example, explain how the data might be utilized, as well as aid with interpretation and additional understanding of complex data.

4. 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.

4.1. 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.

4.2. 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).

4.3. Data File Format

Only comma-separated value (CSV) and tab-separated value (TSV) 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. The TSV standard is available from IANA (Internet Assigned Numbers Authority) at https://www.iana.org/assignments/media-types/text/tab-separated-values.

Important: Carriage return/line feed characters are record termination characters in CSV and TSV files and must exist once and only once at the end of each data record. Carriage return/line feed characters embedded in source fields such as multiline addresses or comments must be removed or converted to some other separating character such as a space, comma, semi-colon, etc.

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

4.4. Use Vertical Rather Than 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

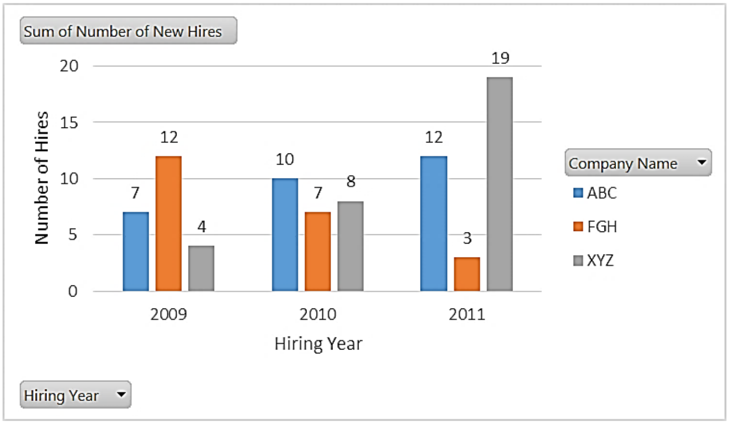
Company Name	2009 New Hires	2010 New Hires	2011 New Hires
ABC	7	10	12

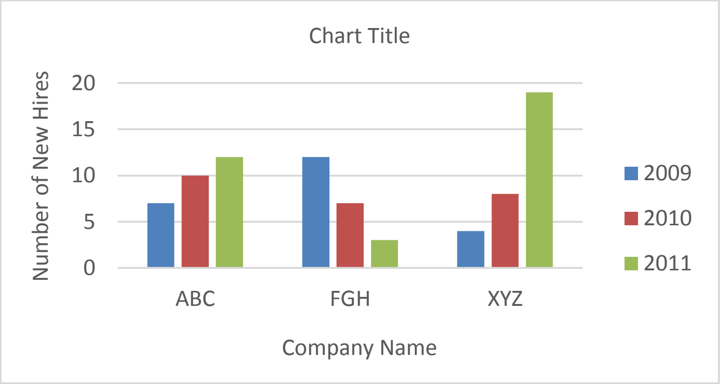
Acceptable – vertical data orientation - additional years can be appended to the data file as needed

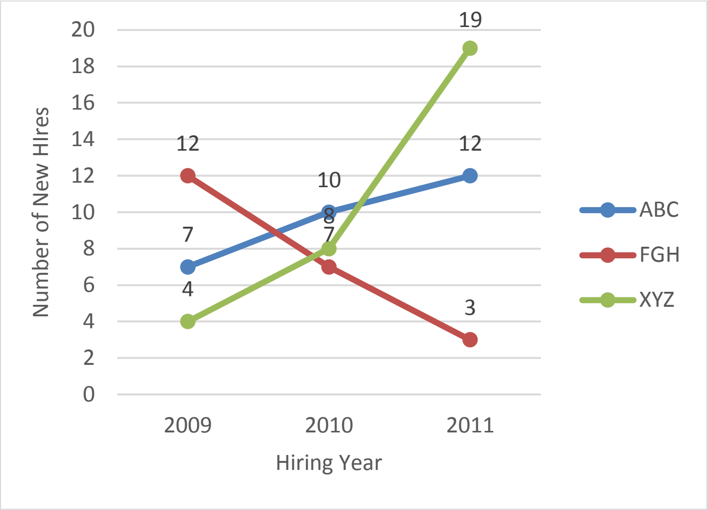
Company Name	Hiring Year	Number of New Hires
ABC	2009	7

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

4.5. 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 Name	**Contact Name**	**Contact Title**
ABC	John Smith	Project Director

4.6. 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

CountyofResidence1YearAgoPopulation1YearandOver_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

4.7. 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

4.8. 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.

Not Acceptable – text included in numeric field

Company	Year	Work Related Injury	Number of Cases
ABC	2009	Skin Disorders	27
ABC	2009	Respiratory Conditions	unknown
ABC	2009	Poisoning	0
ABC	2009	Eye irritations	N/A

Acceptable – text excluded in numeric field

Company	Year	Work Related Injury	Number of Cases
ABC	2009	Skin Disorders	27
ABC	2009	Respiratory Conditions	
ABC	2009	Poisoning	0
ABC	2009	Eye irritations	

If a blank field has a significance that should be represented in the visualizations of a dataset, then the data owner should consider creating a separate column in the data that clearly identifies how it should be represented.

For example, if data on water quality samples contains no values in the test results column because the result of the water quality test was below the detectable limit of the test, then that should be clearly indicated in the data or stated in the metadata. If the data owner wants those test results to display in a visualization of the dataset, then the data owner may want to create an additional column that indicates the test results were below the detectable limit.

Example of Water Quality Sample Dataset, **Acceptable and Properly Formatted,** with added column

Site	Chemical	Test Result	Qualifier
Sample Site	PCB		< 0.1 ug/l

4.9. 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

4.10. 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 Codes[1]. 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

[1] FIPS 6-4 was replaced with ANSI INCITS-31-2009. The United States Census Bureau provides information about the standard here: http://www.census.gov/geo/reference/ansi.html along with lookup tables here: http://www.census.gov/geo/reference/codes/cou.html

4.11. 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

4.12. 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

4.13. 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
```

4.14. 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

4.14.1. *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

4.14.2. 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

4.14.3. 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

4.15. Date Fields

4.15.1. 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.

4.15.2. 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

4.16. 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 4.17 below.

4.17. 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

4.18. 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

4.19. 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.

4.20. 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 – Includes decimal degree latitude and longitude

Facility Name	Street Number	Street Address	City	Zip	Latitude	Longitude

XYZ Medical Care Center	803	Genesee Street	Rochester, NY	14611	42.65749	-73.75566

Acceptable – formatted for mapping

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

4.21. Geographic Coordinates

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

4.22. 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.

4.23. 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.

4.24. 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.

4.24.1. Blank 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 or TSV.

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

4.24.2. 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

4.24.3. 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 or TSV 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

4.24.4. 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

4.24.5. 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
ABC	1/3/13	1/12/13	1/20/13

Depending upon an agency's preference, this may also be highly vertical, especially if the inclusion of additional data concerning status (see example below) is being planned. This will serve to minimize manual intervention for refresh.

Highly Vertical Acceptable Alternative

Project Name	Status	Status Date
ABC	Started	1/3/13
ABC	Design Complete	1/12/13
ABC	Development Started	1/15/13
ABC	Development Complete	1/20/13

4.25. Special Guidance for Data Export Programmers

4.25.1. 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).

4.25.2. 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.

5. Requisite Documentation

As previously noted, "*Quality by Design*" was born from the recognition that it isn't enough to simply publish raw data. Open NY enriches raw content with context to provide the diverse community of users with the greatest possible understanding of the data to maximize its discovery, utility, analysis, reuse, and derivative value. Open Data is about getting data into the hands of all and giving it context - making it easily consumable and understandable. Documentation provides context to data. Context helps removes ambiguity and aids in interpretation.

Each data file submitted for publication on Open NY must include the following accompanying documentation:

- Metadata Document
- Data Dictionary Document
- Overview Document
- Signed Approval Form

Moreover, additional documentation may also be required where it furthers understanding of the data (e.g., a survey instrument, data collection tool, research benefits narrative, supplemental links, studies or reports specific to the data, and more).

5.1. Metadata Form

As noted in the Open Data Handbook, the metadata scheme allows data publishers to classify selected contextual fields or elements within their dataset as well as adhere to common Meta attributes identified platform-wide empowering data end-users to build automated discovery mechanisms at a granular-level. Using a common metadata taxonomy allows Open NY to convey and increase discoverability of high-value datasets.

What is the data "About?" Metadata answers this question by describing a number of characteristics or attributes about the data. Metadata is often described as data about data. It provides structure and contextual information about the data. Metadata should include a description of what the data represents, its granularity, posting frequency, keywords, information about origin, linked data, geographic location, time series, and other relevant indices that allow the public to determine the applicability of the data for any particular use. Metadata provides an abstract of the data for easy reference.

Why is metadata important? Metadata helps facilitate the discovery of information and minimizes misinterpretation. It can help locate and retrieve relevant data from different sources. It provides structure and organization for ease of data identification, search, and integration of data with common attributes. In the same way that an on-line library catalog helps you find books, articles and media, or an on-line product catalog helps you shop for the best product, metadata helps your users discover and begin to understand your data.

An on-line card catalogue organizes information so that it can be identified and can be grouped with the same or related subjects. It includes key information to inform the reader such as: Author, Title, Year of Publication, Place of Publication, Publisher, Number of Pages, Genre, Library of Congress #, etc., all in furtherance of making information discoverable and accessible.

Open NY adheres to core components of the Dublin Core standard[2] for metadata (http://www.dublincore.org/documents/dces/). The ability to search and find information is enhanced by the adherence to metadata standards required with each dataset. In addition, metadata is linked to subject categories which provides for more precise searching and document management. Adoption of the Dublin Core, together with standards for Open Data, maximizes adaptability and interoperability.

The Open NY Metadata Form requires the following fields to be completed and submitted with the data file:

- Dataset Name / Title
    - The name of the dataset as it will appear in the data catalogue on the Open NY website. Use under 200 characters, do not include New York State or NYS in the title, and it should end with the temporal component (i.e., date range or beginning YYYY - For example, Electricity Use 1980-2001; Electricity Use beginning 2002). We reserve the right to edit the name you submit.
- Dataset Description
    - Short description that explains the purpose of the Dataset and the data within. Similar to an abstract which succinctly summarizes the details of the dataset.
- Category
    - The general category within which the dataset should be included in on the site (Categories are: Economic Development, Education, Energy & Environment, Government & Finance, Health, Human Services, Public Safety, Recreation, Transparency, and Transportation). Other categories may be added in the future
- Tags / Keywords
    - Generally up to five (5) lower-case keywords (or phrases) about the dataset used for searching purposes. Please separate terms or phrases with commas, e.g. “power generation, energy, electricity, utility, solar.” Use tags and keywords which will assist in the discoverability of the data and drill down hierarchical. Examples:

[2] The Dublin Core Metadata Initiative (DCMI) is incorporated as a non-profit organization hosted at the National Library Board of Singapore. See: http://wiki.dublincore.org/index.php/User_Guide.

    - energy, solar energy, photovoltaic; or
    - tax, real property, exemptions; or
    - restaurant, food inspection, violation, public health, environmental health;

  If determined useful to discovery and search, more than five keywords or phrases can be applied:

    - alternative fuel, station, biodiesel, natural gas, ethanol, electric, hydrogen.

- Data Provided By
  - The name of the Agency that owns the data.
- URL to Source Link (Dataset Program Web Page)
  - The URL to the primary program web page. Programs may have multiple web pages in addition to the primary landing page and such links can be included under 'Additional Resource.'
- Organization
  - The Program/Division within the Agency that owns the data
- Time Period
  - The timeframe of data available in the associated data file.
    - Proper formatting requires identifying the first date which the published dataset displays such as: Beginning 2005.
    - May include whether it is fiscal year (FY), tax year, etc.
    - Do not include end dates unless it is a closed dataset, or a historically static dataset as these would need to be updated each time the dataset is refreshed.
    - For snapshots such as Park locations, Farmers' Markets indicate with the word "Current."
- Posting Frequency
  - How often the Dataset will be refreshed. Valid values are Daily, Twice weekly, Weekly, Monthly, Quarterly, Semi-annually, Annually, Biannually, Decennially, As Needed, Static - Not Updated.
- Contact e-mail information
  - The e-mail address the viewers of the data can use to ask questions about the dataset. The preference is this not be an individual's mailbox due to staffing changes. Optimally, this should be a monitored shared mailbox or distribution list. If your agency/program already has a generic mailbox that is monitored for answering inquiries, please use that e-mail address.
- Coverage
  - The coverage area included in the dataset (e.g., Statewide).
- Granularity
  - The lowest levels of granularity available within the data file (e.g., town, city, county, DOT region, hour, toll plaza, etc.).
- Define any limitations
  - Description of any limitations of the Dataset or any exclusions.

- URL(s) to additional resources (optional)
    - URLs to additional resources that may be useful to an end-user

5.2. Data Dictionary

The data dictionary creates consistency in understanding and definition. As with metadata, the key is to define elements so that the public has sufficient information to process and understand the described data without ambiguity. The more information that can be conveyed in a standardized format, the more valuable data becomes. Each dataset prepared for submission must include a data dictionary as a separate document. The purpose of the data dictionary is to ensure each column that is included in the data is defined in terms easily understood by the public. All of the detail necessary to describe each of the columns in the data file should be included in the data dictionary. Remember, end-users will not be as familiar with your data as you are.

The Data Dictionary is comprised of the following three components (described below) for each column of the data file:

1. Column Name
2. Data Type
3. Data Description

1. Column Name

- **The Column Names provided in the data dictionary *must match the column names provided in the dataset*.**
- If a term used in a column name is very broad and cannot be altered, the data dictionary is critical in aiding understanding and interpretation. That is, a column name of "Category" will have to be defined so the public knows this means for example "classification of a state park facility."
- Use plain text and fully expand any acronyms/abbreviations. Do not simply repeat the column name as the definition of the column, e.g. a column with a name like "EBT" needs to be labeled in plain English (Electronic Benefit Transfer) and explained, and should not be defined by its name simply as e.g., "EBT".

2. Data Type

Data Type describes the kind of data the column contains. When specifying data types, use common terminology ("Text", "Numeric ", "Date", "Percent", "URL") rather than database-specific terminology such as "Varchar", "Float", etc. Below is a table of common data types and the typical information that kind of column contains.

Data Type	Typical Information Contained

Text	Alphanumeric values such as last name, address, description, etc.
Numeric	Money, quantities, measures.
Date	Full date values.
Time	Time values.
Date & Time	Full date with time.
Percent	Values that represent a percentage.
URL	Hyperlinks to other on-line resources.

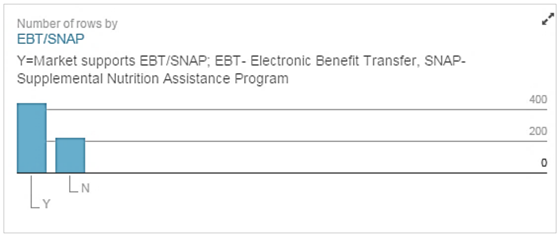
3. Data Description

- Clarity of definition; be as descriptive as possible. Be as descriptive and as granular as possible. For example, if a column name is "Region," and the data populating the column lists numbers 1-9, the column should be defined as the specific state agency region (e.g., DOT Region, OMH Region, DEC Region, ESD Region). In addition, included in the definition should be a legend which corresponds the numbers in the column with a region (e.g., 1 = Hudson Valley, 2= Capital Region, etc.).
- ***Define blank cells - leave nothing ambiguous.*** If blank cells appear in the dataset, the data dictionary description should acknowledge such and provide explanation to preclude misinterpretation.
- For example – "Blank means data was not provided," or "Blank indicates the value was not reported." This is very different than reporting a zero which is a value and defined as: "Zero (0) indicates a zero value was reported."

See section 4.6 herein for optimal naming of columns.

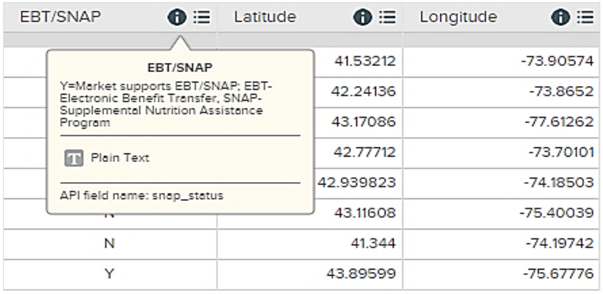
Succinct but descriptive data definitions and column data labels are critical for several reasons:

1. Data dictionary definitions will automatically populate the legends and/or descriptive headers in a variety of visualizations. For example:



Number of rows by
EBT/SNAP
Y=Market supports EBT/SNAP; EBT- Electronic Benefit Transfer, SNAP-Supplemental Nutrition Assistance Program

2. Data dictionary definitions populate fly outs which are accessed through the information icon to understand the tabular data within the column without having to leave the page for the data dictionary.



EBT/SNAP	Latitude	Longitude
	41.53212	-73.90574
	42.24136	-73.8652
	43.17086	-77.61262
	42.77712	-73.70101
	42.939823	-74.18503
N	43.11608	-75.40039
N	41.344	-74.19742
Y	43.89599	-75.67776

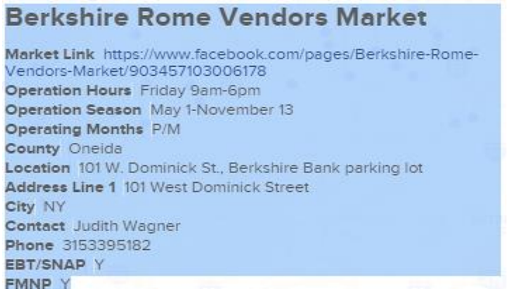
EBT/SNAP

Y=Market supports EBT/SNAP; EBT-Electronic Benefit Transfer, SNAP-Supplemental Nutrition Assistance Program

Plain Text

API field name: snap_status

3. Column Names are displayed in a variety of ways including, but not limited to:
   - Headings of tabular data
   - Map fly outs
   - Chart fly outs
   - Various components of visualizations (e.g., charts, graphs, cards, etc.)



Berkshire Rome Vendors Market

Market Link https://www.facebook.com/pages/Berkshire-Rome-Vendors-Market/903457103006178
Operation Hours Friday 9am-6pm
Operation Season May 1-November 13
Operating Months P/M
County Oneida
Location 101 W. Dominick St., Berkshire Bank parking lot
Address Line 1 101 West Dominick Street
City NY
Contact Judith Wagner
Phone 3153395182
EBT/SNAP Y
FMNP Y

The data dictionary document should be saved as and named using the convention [agency acronym]_[short dataset name with no spaces]_DataDictionary.pdf, for example:

DTF_IncomeTaxComponents_DataDictionary.pdf

5.3. Overview Document

Context is as important as the data itself. It should provide the "who, what, where, how, and why" of the data story. Context helps explain the data. Context is critical to inform decisions and analyses based upon the data. Context aids in the interpretation of the data. Context enables consumers of data to extract maximum value from the data.

The overview document should clearly indicate the name of your agency at the top, and be a separate one to two page PDF document that comprehensively explains the dataset in detail, provides background and context of the data, explains the data collection process (methodology), statistical and analytical issues, and any limitations to data use. Adhering to continuous quality improvement, the overview document must now contain, at a minimum, the following elements:

- NAME OF AGENCY
- TITLE OF DATASET
- General Description
    - Include mission of agency
    - Include role and responsibilities of the Program/Division which owns the data
    - Include a description of the data that is presented (and when referencing dates or quantities use language that will accommodate growth (e.g., beginning, more than) and will not require updates when the data is refreshed.
    - Include why the data is collected: cite statute, regulation, policy, historical context, etc.
- Data Collection Methodology
    - Describe data collection methodology (e.g., how the data is collected, generated). Include any specific tools or survey instruments if applicable. Do not cite name of specific database from which data was extracted.
- Statistical and Analytical Issues
    - May include a description of the data analysis, validity, and reliability gleaned from the data collection. May include a description of the sampling and the analysis upon which the data displayed was based (i.e., figures in this dataset are based upon estimates derived from.…).
- Limitations
    - Include possible methodological limitations (e.g., self-report, coverage area limitations)

Note: 1-2 pages is general guidance for simple datasets. That said, many times more complex datasets necessitate multiple pages in order to provide the end-user with the greatest degree of understanding of the dataset (also see section: Additional Documentation 5.4). This is an opportunity to explain the dataset to end-users who are unfamiliar with the data, your agency, and why the agency maintains such data. As datasets are updated at different required frequencies, (daily, monthly, quarterly, annually, etc.), the overview document should be general enough that it will not need updating each time the data is refreshed, unless the data has significantly changed. For example, citing specific years and dollar amounts.

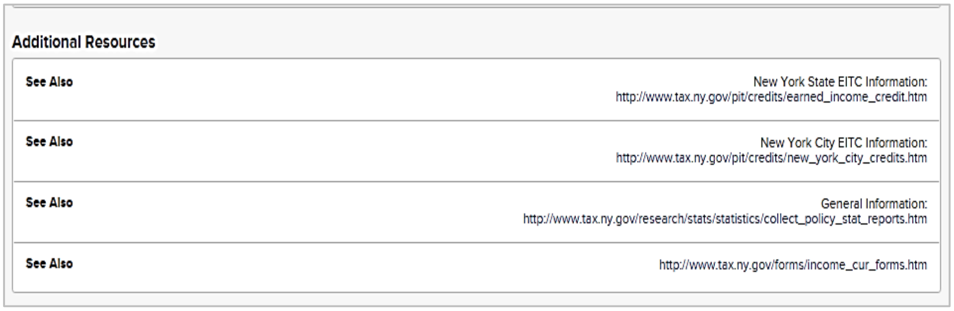
When indicating limitations of data use, take care not to include terms that are contrary to the Open Data Terms of Service (https://data.ny.gov/download/77gx-ii52/application/pdf) especially regarding re-use or re-distribution of the data or requirements for citing the agency in derivative works. The limitations section is intended to explain known data issues that could cause misunderstanding or misinterpretation, such as missing years of data, holes in geographic coverage, or changes to processes that resulted in significant differences in the way data is reported or recorded.

The overview document should be saved as and named using the convention [agency acronym]_[short dataset name with no spaces]_Overview.pdf, for example:

DTF_IncomeTaxComponents_Overview.pdf

5.4. Additional Documentation

As discussed, additional documentation should be included to aid in the understanding of complex data, enhancing its reuse and analysis. Additional documentation may include, but not be limited to the following: survey instruments, data collection tools, research benefits narrative, supplemental links, studies or reports specific to the data, links to additional resources or attachments such as studies, maps, data collection tools, and more.



Additional Resources

See Also	New York State EITC Information: http://www.tax.ny.gov/pit/credits/earned_income_credit.htm
See Also	New York City EITC Information: http://www.tax.ny.gov/pit/credits/new_york_city_credits.htm
See Also	General Information: http://www.tax.ny.gov/research/stats/statistics/collect_policy_stat_reports.htm
See Also	http://www.tax.ny.gov/forms/income_cur_forms.htm

5.5. Required Publishing Approvals

As noted in the Open Data Handbook http://ny.github.io/open-data-handbook/OpenDataHandbook.pdf, in identifying Publishable State Data, agencies should include analyses from their executive and program staff, data coordinators, FOIL officers, data stewards/IT, public information officers, security and privacy officers, and legal counsel.

In furtherance of Quality by Design, and per the guidelines of the Open Data Handbook, agencies engage in an internal process to identify potential datasets eligible for public release. Participating agencies are required to engage in an internal review process and obtain approvals for the datasets which the agency wishes to commit to the Open Data Website. Prior to, and up through publication of any dataset, a comprehensive analysis is undertaken by each agency to review data content, quality, and accuracy of such dataset as well as an assessment of the metadata, overview documents, and data dictionaries. Agencies are responsible for driving towards increasing data content quality and accuracy, and are responsible for ensuring compliance with all security, privacy, confidentiality laws, rules, and regulations, as well as any Intellectual Property Rights requirements and status under the NYS Freedom of Information Law (including whether data may lawfully be withheld under FOIL's limited exceptions). Once such reviews are complete, if the data is assessed as publishable, it is published to data.ny.gov.

For each individual dataset, at a minimum, agencies must receive explicit approval and sign-off from the individuals listed on the Approval Form (Head of Agency or designee, Counsel, Data Owner, Data Coordinator). Such standardized approval form provided by ITS must be completed and signed prior to dataset publication. Agencies may also determine additional signatures are required, and should include such additional persons (e.g. Public Information Officer) on the Approval Form.

6. Conclusion

These guidelines serve to underscore the significance of having a core set of principles that establish, within a dynamic and innovative environment:

- Policies, rules, and procedures to facilitate consistency;
- Standards that drive quality, performance, & data management best practices; standards that provide a structure that will help bridge disparate vocabularies, syntax, data sources, sectors;
- A strong foundation and effective framework that optimizes data and creates opportunity, business value, and insight.