

NYC OpenData

Data Quality Standards and Review Process

May 2022 Revision

NYC Open Data Team
Office of Technology and Innovation (OTI)

Table of Contents

Data Quality — Overview	3
Open Data Team Review Process	4
Data Quality Self-Assessment Checklist — Overview	5
Data Quality Self-Assessment Checklist	6
Technical Logistics	7
Data Dictionary	10
Compliance Information	14
Disclosure Considerations	15

This document is meant to serve as a resource for Open Data Coordinators and Data Owners that sets forth data quality standards for publishing on NYC Open Data.

Data Quality — Overview

An important element of making open data accessible to the public is ensuring the data is high-quality and is accompanied by good data documentation. The following is a list of data quality and data documentation standards that all NYC Open Data datasets must follow to ensure completeness, clarity, and overall data quality.

This document is intended as a resource for ODCs and data owners to review and self-assess the quality of their data and data documentation and is also used by the NYC Open Data Team (ODT) for reviewing datasets prior to their publication to NYC Open Data.

Feedback

The ODT at the Office of Technology and Innovation (OTI) welcomes your feedback on the data quality standards - please email us at opendata@oti.nyc.gov with any suggestions how we can improve our standards and review processes.

Updates

The ODT will be making changes to this document, last updated on May 19, 2022. Please check periodically for updates and make sure you are referencing the most up to date version.

Open Data Team Review Process

The ODT reviews each dataset prior to its publication to NYC Open Data to ensure it meets data quality and documentation standards. Upon review of any dataset, the ODT communicates feedback to the agency and may direct the agency to fill data and data documentation gaps, edit necessary information, or gather more input from data owners prior to allowing for the publication of any dataset on NYC Open Data.

In preparing datasets for publication to NYC Open Data, agencies must take into account additional time needed for 1) dataset review by the ODT and 2) for agency updates to the dataset based on feedback received from the ODT. The ODT review period will typically take one week or less, but this can vary based on dataset complexity and the number of datasets being reviewed at the same time. In all cases, the ODT will provide feedback or a feedback timeline within a week of receiving a dataset. Agencies with datasets that meet legislative publication deadlines and deadlines set forth in the annual Open Data Plan should pay particular attention to this additional review time.

Data Quality Self-Assessment Checklist — Overview

The following checklist communicates the ODT's standards for the quality of data and data documentation. Each item within the checklist is linked to additional context and descriptive examples.

ODCs must review and follow this checklist to ensure completeness, clarity, and overall quality for all datasets and data dictionaries **before** passing them to the ODT for publication.

The ODT consults the following checklist when it reviews and shares feedback on datasets received from agencies prior to publication to NYC Open Data.

ODCs should feel free to share this document and the checklist with their data owners and other relevant stakeholders within their agencies who may be working on preparing datasets and completing data documentation.

Data Quality Self-Assessment Checklist

Each item must be checked off for a dataset to be eligible for publication on NYC Open Data.

Technical Logistics

- [Is the dataset “data” as defined by Chapter 5 of Title 23 of the NYC Administrative Code?](#)
- [Is the dataset in a machine-readable, flat, tabular format?](#)
- [Is the dataset one that cannot be consolidated with another dataset on Open Data?](#)
- [Are there no duplicate records within the dataset?](#)
- [Is the data at a consistent level of aggregation, with a single answer to “Each row is a...”?](#)
- [Is the data at the smallest level of aggregation?](#)
- [If applicable, has the dataset been geocoded with all required geospatial reference fields included?](#)

Data Dictionary

- [Is there a data dictionary that accurately reflects the information present within the dataset?](#)
- [Is the “Dataset Name” brief yet descriptive with all acronyms expanded?](#)
- [Is there a thoroughly completed “Dataset Information” tab in the data dictionary?](#)
- [Are all of the dataset column headers present in the data dictionary, exactly as they appear in the original dataset?](#)
- [Does the data dictionary contain an easily understandable explanation for each column present in the dataset?](#)
- [Is there a complete list of “Expected/ Allowed Values” that matches those included in the dataset?](#)
- [Is there an explanation of the significance of any null, missing, or zero values present in the dataset?](#)
- [Are all acronyms, agency-specific/technical terms, abbreviations, and codes defined?](#)

Compliance Information

- [Has it been indicated whether rows are removed from this dataset when the data is updated?](#)
- [Has it been indicated whether this data is also present on a website maintained by or on behalf of the agency?](#)
- [Has it been indicated whether this dataset can be feasibly automated and/or is currently updated automatically on the agency website?](#)

Disclosure Considerations

- [Has the dataset been reviewed for legal and privacy considerations by the agency’s legal office and cleared for publication?](#)
- [Is the dataset owned and maintained by the agency publishing it?](#)

Technical Logistics

Is the dataset “data” as defined by [Chapter 5 of Title 23 of the NYC Administrative Code](#)?

In order for a dataset to be eligible for publication on NYC Open Data, it must adhere to the legal definition of Data, that is “final versions of statistical or factual information.”

A partial set of records or segments of a larger dataset are examples of datasets that are not appropriate for publication on NYC Open Data. Instead, the larger, complete dataset would be eligible for publication.

To ensure that a dataset is a complete final version, it can be helpful to determine if it can be included as part of a larger dataset. Agencies should not publish a subset of information (for example, registered properties from May 20, 1996 - December 15, 2002) when there exists a dataset that contains more information (for example, all registered properties from 1924 to present).

Is the dataset in a machine-readable, flat, tabular format?

A tabular dataset is a flat file structured into rows and a set number of columns. It must not include merged cells, highlighted or color-coded values, multiple tables in one dataset (e.g. multiple tabs in Excel), or unsupported data types.

A good way of determining if a dataset is in a flat format is if each row represents the same thing (for example, every row is a property). Agencies should generally remove rows that contain calculated values (for example, totals or averages of a subset of the data). Otherwise, datasets where different rows contain different things should be evaluated to be split into multiple datasets (for example, a dataset of properties and a related dataset of detailed property maintenance). Please see the checklist question [Is the data at a consistent level of aggregation with a single answer to “Each row is a...”?](#) below for more detail.

Is the dataset one that cannot be consolidated with another dataset on NYC Open Data?

Agencies are expected to consolidate multiple datasets on the same topic whenever possible. This includes data collected across different years or geographies.

For example, datasets such as “AgencyName Properties Brooklyn” , “AgencyName Properties Manhattan”, and “AgencyName Properties Bronx” can be combined into one dataset named “AgencyName Properties” with an additional “Borough” column.

Similarly, datasets such as “AgencyName Properties 2019”, “AgencyName Properties 2020”, and “AgencyName Properties 2021” should be consolidated into one dataset named “AgencyName Properties 2019 to Present” with an additional “Year” column. Every subsequent dataset that continues to pertain to “AgencyName Properties” will then be added to the one dataset.

Are there no duplicate records within the dataset?

Agencies are expected to remove all repetitive or duplicate rows throughout a dataset. Equivalent records at varying levels of aggregation qualify as duplicate rows and should be consolidated to the smallest level of aggregation. Please see the checklist question [“Is the data at a consistent level of aggregation with a single answer to “Each row is a...”?”](#) below for more detail.

Is the data at a consistent level of aggregation, with a single answer to “Each row is a...”?

For a dataset to be published to NYC Open Data, each row must represent the same thing. Agencies must ensure that datasets include information at a singular consistent level of aggregation by specifying the unit of analysis in the data dictionary field “Each Row is a...” (for example “each row is a registered property”).

The data dictionary field “Each Row is a...” should not provide any additional context or detailed information about the data. This additional detail should instead be communicated through the data dictionary’s “Dataset Description” field. For example, “each row is a property registered between the years of 2007 and 2019 as used for the AgencyName database on property maintenance” is not necessary to determine the unit of analysis or level of aggregation of the data and should instead be included in the data dictionary’s “Dataset Description” field.

If it seems difficult to provide a clear answer to the data dictionary field “Each Row is a...” as different rows contain different units of analysis (for example, each row is either a registered property or a date when property maintenance took place”), agencies should split the dataset into multiple datasets (for example, a dataset of properties and a related dataset of detailed property maintenance).

Information repeated through the inclusion of multiple different levels of aggregation qualifies as duplicate rows and should be removed. Specifically, data included at both a singular and summary level should be published solely at the smallest level of aggregation. All records that include larger levels of aggregation should be removed from the dataset. For example, a dataset with rows that indicate total property visits by district, total property visits by borough, and total property visits by city provides the same information through varying levels of aggregation and should be consolidated into a dataset with the smallest level of aggregation, in this case, total property visits by district.

More examples of data with consistent versus inconsistent levels of aggregation can be seen below:

Consistent Level of Aggregation			Inconsistent Level of Aggregation	
Location	Year	Quarter	Location	Reporting Period
Outside NYC	2021	Q1	Europe	Cumulative (2021)
Queens	2021	Q2	New Jersey	2021 Q2
Brooklyn	2020	Q4	United States	2020 Q4
Bronx	2021	Q4	Brooklyn	Cumulative (2021)
<p>The data in this table is included at a uniform level of aggregation throughout (by borough and quarterly) and at the smallest level of aggregation at which the data is collected.</p>			<p>The data in this table is included at different and inconsistent levels of aggregation as it contains both row-level data and summary-level data. Data of differing orders of magnitude should not appear in the same table.</p>	

	<p>In this case, the data should be disaggregated to the smallest, most granular level (by borough and quarterly).</p>
--	--

Is the data at the smallest level of aggregation?

Agencies should create datasets that include information only at a consistent, granular level of aggregation. For example, geography information included at the borough level should be disaggregated to the Census tract. Similarly, registration information collected monthly should have each record indicated by a new month, not quarter or year. All rows that include larger levels of aggregation should be removed. Please see checklist questions [“Is the data at a consistent level of aggregation, such that there is a single answer to “Each row is a...”?”](#) above for more detail.

If there is an issue of privacy that restricts the data from being further disaggregated, the reasoning must be communicated to the ODT. See [“Has the necessary legal and/or privacy office authorization for publishing the dataset been received, if applicable?”](#) below for more detail.

If applicable, has the dataset been geocoded with all required geospatial reference fields included?

Any dataset containing street addresses is expected to be geocoded and contain Borough, Postcode, Latitude, Longitude, Community Board, Council District, BIN (Building Identification Number), BBL (Borough Block Lot), Census Tract, and NTA (Neighborhood Tabulation Area) columns. Any agency requiring further technical guidance here can receive assistance from the ODT.

Data Dictionary

Is there a data dictionary that accurately reflects the information present within the dataset?

Every dataset published on NYC Open Data must possess an accompanying data dictionary that explains every column within the dataset and includes four key sections: primer page and internal information, dataset information, column information, and dataset revision history.

Is the “Dataset Name” brief yet descriptive with all acronyms expanded?

Any dataset published on NYC Open Data is expected to have a clear, unique name that describes the dataset as well as differentiates it from similar datasets.

Acronyms in the dataset title must be expanded and abbreviations should not be used. For example, a dataset name such as “BIN Prop. Violations” would be written as to “Building Identification Number (BIN) Property Violations”

Datasets that require additional information to clarify ambiguity, such as “Properties”, should include the agency abbreviation or full agency name in the dataset name to provide necessary distinction (i.e. AgencyName Properties). Otherwise, including the name of the agency in the dataset name should be avoided as it is already recorded in other parts of the primer page and data dictionary.

Additionally, information about the dataset time frame should be included in the dataset name. If the dataset will be regularly updated, the starting year should be included in the dataset title (i.e. “AgencyName Properties: 2009 to Present”). If the dataset pertains to a specific year, cannot be consolidated with similar datasets, and will not be updated in the future, agencies should include the specific year in the dataset title (i.e. “AgencyName Historic Properties 1897”).

Is there a thoroughly completed “Dataset Information” tab in the data dictionary?

The data dictionary’s “Dataset Information” tab must provide a main overview of the dataset and include all necessary context, such as what the dataset contains, how often changed data is published to this dataset, the definitions of key terms, how the data was collected, etc.

A complete, usable “Dataset Information” tab should provide answers to all of the following questions:

1. What information does the dataset contain?
2. What is the publishing and data change frequency of the dataset?
3. Why was the data collected? (e.g. what were the legal policies, including Executive Orders or policy directives, that required collection?)
4. How was the data collected?
5. How can the data be used?
6. What are the unique characteristics or limitations of the dataset?
7. For geospatial data, what is the coordinate system/ projection being used?

Dataset Description Details:

Where relevant, the “Dataset Description” field in the “Dataset Information” tab must include links to agency websites, external sources, and suggested links to other datasets for additional context. For datasets that contain mappable spatial information, the map or related visualization should be linked to the dataset in this field.

Publishing Frequency Details:

The “Publishing Frequency” field in the “Dataset Information” tab must match the intended use and nature of the data, where processes and tools are in place to support the updating and be one of the following options:

- Hourly
- Daily
- Weekly
- Every weekday
- Every 2 weeks
- Monthly
- Every 2 months
- Quarterly
- Every 4 months
- Every 6 months
- Annually
- Every 2 years
- Every 3 years
- Every 4 years
- Every 5 years
- Every 10 years
- Historical Data
- As needed
- To Be Determined
- Not Applicable

Agencies should select “Historical Data” if there is no intent to publish this data in the future. Datasets that were once regularly published can be made historical if their underlying systems have been deprecated. Any one-time release of data will be ‘historical’ as soon as it is published.

Data that is collected and reported annually as separate historical datasets (such as “AgencyName Properties 2019”, “AgencyName Properties 2020”, and “AgencyName Properties 2021”) should be consolidated into one dataset (such as AgencyName Properties 2019 to Present”) with an annual publishing frequency and additional “Year” column. If columns were removed, changed, or measured differently over time, datasets should still be consolidated and specific adjustments indicated in the data dictionary.

Agencies should select “As needed” if the agency intends to publish this data in the future at an as-needed, irregular cadence or a regular cadence that does not match any of the options listed above. The specific frequency must then be explained in the ‘Frequency Details’ field of the data dictionary. Any agency requiring further guidance here can receive assistance from the ODT.

For geospatial data, does the “Additional geospatial information” field indicate the coordinate reference system or projection being used?

For any datasets with geospatial data, agencies must specify the coordinate reference system or projection used within the data dictionary’s “Additional geospatial information” field.


Are all of the dataset column headers present in the data dictionary, exactly as they appear in the original dataset?

In the data dictionary, agencies must include a list of every column name in the order they appear in the dataset. If there is not a fixed number of columns that can be explained (e.g. only a quarter of the rows have entries in the first three columns or columns D and E are merged at multiple points throughout the dataset), this may indicate that the dataset is not in a flat, tabular form.

The column names in the data dictionary must be written identically to the column headers in the dataset. If the original column name includes an acronym, technical term, or abbreviation (such as BBL or Lat), agencies must expand the name in the appropriate “Column Description” field.

All of the dataset column headers must be listed in the same order in the data dictionary as they appear in the dataset. It is strongly recommended that columns are ordered following best practice: the primary key should be the first column and all geocoded or otherwise related columns should be grouped together. A good way to ensure the primary key is the first column is asking whether the dataset’s entries contain the answer to “Each row is a...?”

Example of well versus poorly organized dataset columns:

	Property ID	Property Type	Month	Year	Address	Latitude	Longitude
	325	Helios	October	2019	20 W 34th St, I	40.7486538	-73.9853043
	328	Cloudster	September	2020	405 Lexington	40.7519331	-73.9754538



Month	Property ID	Address	Year	Latitude	Property Type	Longitude
October	325	20 W 34th S	2019	40.7486538	Helios	-73.9853043
September	328	405 Lexington	2020	40.7519331	Cloudster	-73.9754538

Does the data dictionary contain an easily understandable explanation for each column present in the dataset?

Agencies must provide a description for every column present in the dataset in the data dictionary “Column Description” field. The data dictionary “Column Description” field must include the expansion of any acronyms or abbreviations found in the Column Name, incorporation of plain language words that make technical terms more accessible to the general public, and additional information beyond a restating of the column header name.

A complete, usable “Column Description” field should provide answers to all of the following questions:

- 1. What information does the column contain?*
- 2. How does the information in this column relate to other columns in the dataset, if applicable?*
- 3. What do any of the acronyms, technical terms, or abbreviations included in the Column Name mean?*
- 4. What are the unique characteristics or limitations of the entries found in the column?*

Is there a complete list of “Expected/ Allowed Values” that matches those included in the dataset?

To ensure a complete data dictionary, agencies must specify in the data dictionary field “Expected/Allowed Values” the range and/or format of the possible values for each column in the dataset.

If the column includes any of the following data types, it necessitates a list of “Expected/ Allowed Values”:

- *dates or times*
 - *agencies should denote the format for the timestamp data, e.g. MM/DD/YYYY or MM/YYYY, 12 or 24-hour time, etc.*
- *numeric data*
 - *agencies should specify the unit of measure of the column, e.g. thousands, millions, \$ value, miles, feet, etc.*
- *a finite list of possible values*
 - *agencies should delineate the expected values as well as how to interpret entries that fall outside those anticipated values e.g. if the Column Name is ‘Borough’ then the entry should be “expected values are solely Brooklyn, Bronx, Manhattan, Queens, Staten Island, or Outside of NYC”*
- *lists or joint values*
 - *agencies should explain how such values are delineated and identify the character used to separate them, e.g. “entries that indicate joint agency jurisdiction are written as independent agencies separated by a backslash (\)”*

Is there an explanation of the significance of any null, missing, or zero values present in the dataset?

The data dictionary must include comprehensive information that specifies how a user should interpret zero, null, missing, or otherwise non-intuitive values present in the dataset. This information should be included in the “Column Information” tab of the data dictionary.

For example, agencies should specify if a zero value indicates that there was no recorded observation or that the observed value was indeed zero. If an entry in a column is lacking data, the agency should provide an explanation or additional context to aid the user in understanding why such data is missing.

If there are empty entries due to columns being solely applicable to a subset of rows (such as a “Quarter” column being inapplicable to rows containing rows with data at an annual level), the data should be disaggregated to the smallest, most granular level and only then published to NYC Open Data. Please see checklist questions [“Is the data at a consistent level of aggregation, such that there is a single answer to ‘Each row is a...?’](#)” and [“Is the data at the smallest level of aggregation?”](#) above for more detail.

Alternatively, if entries are null because columns were added over time, the specific adjustments and additions made must be indicated in the data dictionary, in the “How is this data collected?” field in the “Dataset Information” tab and/or in the “Revision History” tab.

Are all acronyms, agency-specific/technical terms, abbreviations, and codes defined?

To ensure a complete data dictionary, agencies must define any terms, acronyms, codes, or non-intuitive language found in the column values that were not previously defined in the dataset description or column description. This information should be included in the “Additional Information” field in the data dictionary.

Agencies should not assume that an acronym or industry-term is common knowledge. Examples of terms that need to be defined/ given context include, but are not limited to:

- *FHV as an abbreviation for “For-Hire Vehicle” or BBL as an abbreviation for “Borough Block Lot”*
- *the borough that corresponds to each digit (place and value) of the Building Identification Number (BIN)*
- *the corresponding numerical score for a letter grade*
- *any of the terms listed in the NYC Open Data [Glossary of Terms](#)*

Compliance Information

Has it been indicated whether rows are removed from this dataset when the data is updated?

Agencies must specify whether existing rows of the dataset are removed or not when the dataset is updated. If values in the dataset change but no rows are removed as the dataset is updated, then the answer here is no.

Has it been indicated whether this data is also present on a website maintained by or on behalf of the agency?

Agencies must indicate whether the dataset is also present on the agency's website. If the answer is yes, the URL to the agency website where that dataset is featured must also be included in the data dictionary in the field "Is this data also present on a website maintained by or on behalf of the agency?"

Has it been indicated whether this dataset can be feasibly automated or is currently updated automatically on the agency website?

If the dataset is also present on the agency website, agencies must clearly indicate whether it is automatically updated in the data dictionary in the field "If this data is also present on a website maintained by or on behalf of the agency, is that website data updated automatically?"

Additionally, agencies must specify whether the dataset can be feasibly automated on NYC Open Data for future updates in the field "Can this dataset be feasibly automated?" in the data dictionary.

Disclosure Considerations

Has the dataset been reviewed for legal and privacy considerations by the agency's legal office and cleared for publishing?

ODCs should consult with their Agency Privacy Officers (APOs) to determine whether the agency's public datasets include identifying information before such data is made publicly available. If the dataset contains identifying information, the ODC and Agency Privacy Officer should collaborate so that any publication of identifying information on NYC Open Data complies with the requirements of the Identifying Information Law and any other applicable laws and regulations. Agencies must be mindful of all restrictions, possible unintended effects, necessary anonymization, and private or sensitive information related to dataset publication.

Is the dataset owned and maintained by the agency publishing it?

The dataset must be owned by the agency attempting to publish it on NYC Open Data.

In the circumstance where the agency does not possess ownership of the dataset or all the necessary rights to be able to publish the dataset, the appropriate permission must be secured from either the ODT, respective government agency, or sourcing institution.