Metadata for All Guide

NYC Open Data Team
With Julia Marden, Tiny Panther Consulting
Sharon Lintz and Jo Polanco

Table of Contents

Introduction

NYC Open Data's mission is *Open Data for All*. In order to meet that goal, we must examine our Data Dictionaries— the metadata we provide to all Open Data users. Well-written, user-friendly data dictionaries help users to become self-reliant, able to answer their own questions about a dataset without needing to contact an Open Data Coordinator for clarification.

NYC Open Data partnered with the Metropolitan Library Council (METRO), Pratt Institute, Sloan Foundation, The Mayor's Fund to Advance New York City, Tiny Panther Consulting, New York Public Library, Queens Public Library and Brooklyn Public Library to implement an initiative called *Metadata for All* in the Summer of 2018 to accomplish the following:

- Assess the usability of the metadata for the top 100 most viewed datasets on NYC Open Data
- Develop a new standard for NYC Open Data metadata based on usability criteria
- Create resources for NYC Open Data Coordinators to learn how to develop metadata in accordance with the new standard

This document is a resource for Open Data Coordinators to implement new Metadata for All standards.

Project Background and Research Methods

Metadata for All was led by a team of three data librarians, with support from a broader community of librarians through the Metropolitan New York Library Council (METRO). Data librarians have expertise in *helping data owners to manage and share their data* and *helping users to find and access data*. Engaging data librarians for this initiative is the first step in bringing new support and perspective to NYC Open Data.

During a six (6) month pilot project, the Metadata for All team accomplished the following:

- Assessed usability of metadata of the top 100 most viewed datasets on NYC Open Data
- Developed new **Metadata for All Standards** for NYC Open Data metadata based on usability criteria
- Created **Resources** for NYC Open Data Coordinators to learn how to develop metadata in accordance with the new standard

In order to meet these goals, the team conducted user research and solicited feedback from a number of project stakeholders in the NYC Open Data ecosystem and beyond:

- Led a Data Dictionary Workshop with library and information professionals to new pilot methods for evaluating Data Dictionary usability
- Reviewed the existing 2016 Data Dictionary template and supporting resources
- Conducted focus groups and interviews with Open Data Coordinators and their colleagues at five (5) New York City agencies
- Hosted five (5) Open Data workshops with library patrons across all five of New York City's boroughs
- Interviewed open data teams and librarians in New York City, Pittsburgh, PA, and Boston, MA to learn about other efforts to improve data documentation and usability

Recommendations

We salute the NYC Open Data team and Open Data Coordinators who have met legal requirements to provide a Data Dictionary for every dataset on the open data portal, and for the research conducted in 2016 to develop the first NYC Open Data Data Dictionary template. Given the mandate of Metadata for All to assess all resources for usability, these are the major usability challenges identified with the Data Dictionary template issued in 2016:

1) *NYC Open Data users don't understand what they can do with the data.* Current Data Dictionaries fail to help users assess whether they can use a particular dataset to answer their questions or to form appropriate questions for a dataset. Users identified four elements they need in order to form appropriate questions for data analysis: 1) *a description of one row of data*, with 2) *the necessary context of the agency who creates and maintains that data,* 3) *a complete list of columns or fields* in the dataset, and 4) *a complete list of expected values* for each field.

2) *Open Data Coordinators are unsure how to fill out the Data Dictionary template issued in 2016.* We heard this in interviews with Open Data Coordinators and saw it evidenced among the 100 Data Dictionaries we reviewed on NYC Open Data. We read many Data Dictionaries that left us with more questions than answers. Open Data Coordinators told us 1) *they struggled to define certain agency or City terms*, 2) *they struggled to get necessary information from data owners within their agency*, and at times, 3) *were not certain about what information was being requested in the Data Dictionary template*.

Given these challenges, the Metadata for All team's recommendation is to replace the Data Dictionary template issued in 2016 with:

1. A **User Guide**, published as a Google Doc on the dataset's Primer Page
2. A new **Data Dictionary** template, published as a Google Sheet on the dataset's Primer Page

The **User Guide** is a narrative introduction to describe the content of a dataset, how it was created, the agency who maintains it, and how users can begin to use the data. The User Guide template helps answer common questions posed by users and clear up common misconceptions about open data. We want users to understand what kinds of questions they can ask with data; to do that, they need the background and context provided by a User Guide.

We have also updated the recommended **Data Dictionary** template to make it easier to complete by Open Data Coordinators, and to standardize metadata for datasets across NYC Open Data. While many new Open Data users will benefit from a long-form User Guide, experienced Open Data users count on accurate and concise Data Dictionaries.

The **Metadata for All Standards** require both forms of documentation, so that users will be able to choose the format they feel most comfortable using and the level of information they need. Below is a summary of the new **Metadata for All Standards** and a look at why the new Data Dictionary and User Guide improve usability for Open Data Coordinators and users alike.

Comparing the 2016-issued Data Dictionary and Metadata for All Standard's User Guide & Data Dictionary

Summary of Metadata for All Standards	2016 issued Data Dictionary	Metadata for All Standard's Data Dictionary and User Guide
Web-friendly format		✔
Clear, consistent language		✔
Defines jargon		✔
Describes one row or geospatial area	✔	✔
Provides agency context		✔
Explains every column or field	✔	✔
Explains all expected values		✔
Describes how data is managed		✔
Describes how data has changed over time		✔
Suggests appropriate questions for use		✔

Summary of Resources

The Metadata for All Team developed the following resources to help Open Data Coordinators in the development of User Guides and Data Dictionaries that meet the Metadata for All Standards:

- **Rubric:** to help you assess whether your User Guide and Data Dictionary are complete and meet metadata usability criteria. You must be able to check all of the boxes on the Rubric in order to meet the **Metadata For All Standards** for user-friendly metadata.

- **Templates:** Google Docs that can be copied and used for every User Guide and Data Dictionary. Each template is already formatted and has suggestions for how to fill out each section.

- **Glossary:** a list of standard definitions for fields that appear in a number of NYC Open Data sets; you can pull from these standard definitions when developing new metadata for your User Guide or Data Dictionary

- **Process Map Worksheet:** a tool for helping you to map out the programs, policies, and agency structures behind the dataset; this will help suss out metadata that you will need to write about in your User Guide

- **Questionnaire:** a set of questions to help you start thinking about the type of information you will need to gather for your Data Dictionary and User Guide

- **Community Worksheet:** an exercise you can conduct with the public, to gather feedback on how well your Data Dictionary and User Guide are meeting their needs

Additional Opportunities

Given the limited scope of the Metadata for All Team's work, not all insights were able to translate into immediate resources or usability criteria. Additional recommendations for NYC Open Data to keep in mind in the continued pursuit of user-friendly metadata include:

1. We recommend continued partnerships with NYC librarians and archivists to help NYC Open Data meet the challenge of writing and assessing data documentation for every dataset on the NYC Open Data portal
2. For improved user-friendly documentation, we encourage NYC Open Data to explore more web-friendly formats. We believe data dictionaries and user guides work better as web pages or woven into Socrata primer pages, rather than as attachments on the Open Data portal.
3. We suggest that NYC Open Data find ways to automate and scale the creation of data documentation. For example, create a content management system (CMS) so Open Data Coordinators can upload documentation through a webform or publish straight to a web page. Make it easier for Open Data Coordinators to write documentation, by curating a database of common terms and definitions.
4. We hope that NYC Open Data will update Socrata's Primer Pages to incorporate more information from the User Guide. For example, add the "one row of data" feature or make

sure a concise description of a dataset appears at the top of the page with links to tutorials or other introductory resources to NYC Open Data.

Special Thanks

Metadata for All Standards

Data Dictionaries and User Guides must meet the following standards to ensure that metadata can be used effectively by a wide variety of audiences. The following standards are also provided in rubric form in the next section.

1) Data Dictionaries and User Guides should empower users to **easily access and understand metadata**:

   - Title the dataset to be descriptive and accurate following the format: User Guide | {Name of Dataset}, Data Dictionary | {Name of Dataset}
   - Use clear, concise and standardized language throughout the User Guide and Data Dictionary. All jargon and acronyms should be defined. Users should be able to understand the data without any specialized knowledge of your agency. Don't force your users to search for any terms.
   - Follow the User Guide and Data Dictionary template, so that information is presented in a visually cohesive and easy-to-read format.
   - Providing all information requested.

2) The User Guide helps **users understand how the dataset can be used to ask and answer questions.** In order to use the dataset, the user must understand what the data represents, and how the data was collected and published. To meet Metadata for All Standards, a User Guide must:

   - Clearly, concisely, and accurately, describe one row or point of data.
   - Give appropriate Agency context that explains how and why data was collected.
   - Sufficiently describe dataset quality and/or limitations (missing data, time range, etc.)
   - Suggest 2-3 analytic questions to help users get started.
   - Summarize all columns or attributes, and describe expected values (plain text data dictionary and/or codebook.)
   - Sufficiently describe data management (geocoding, columns added or removed, frequency of updates.)

3) The Data Dictionary empowers users to **quickly assess the meaning of each column or attribute.** Each column in the Data Dictionary helps the user understand how that data is collected, managed, and published -- critical information for deciding whether to use that aspect of the data in an analysis. A Data Dictionary must:

- Add context to every column name or attribute.
- List all codes or values in Expected Values or provide a link to relevant information in the User Guide.
- Define all Agency and City terms or acronyms in easy-to-understand language.
- Ensure data types are accurate and match Socrata's allowed values when possible.

Metadata for All Rubric

The Metadata for All Rubric was developed to help you assess whether your User Guide and Data Dictionary are complete and meet metadata usability criteria. You must be able to check all of the boxes on the Rubric in order to meet the **Metadata For All Standards** for user-friendly metadata.

Dataset Name: {Name}{Link to dataset primer}

		✔	Notes:
Overall Usability	**User can easily access and understand metadata documentation.**		
	Title is descriptive and accurate. Follows format User Guide {Name of Dataset}, Data Dictionary {Name of Dataset}, etc.		
	Language is clear, concise and standardized when possible.		
	Information is presented in a visually cohesive and easy-to-read format. (Follows User Guide and Data Dictionary Templates)		
	User Guide and Data Dictionary templates are complete and accurate (all elements filled in).		
User Guide	**User understands how dataset can be used to ask and answer questions.**		
	Guide clearly, concisely, and accurately, describes one row or point of data.		
	Guide gives appropriate Agency context in order for user to understand how and why data was collected.		
	Guide sufficiently describes dataset quality and/or limitations (missing data, time range, etc.)		
	Guide suggests 2-3 analytic questions to help users get started.		
	Guide summarizes all columns or attributes, and describes expected values (plain text data dictionary and/or codebook)		
	Guide sufficiently describes data management (geocoding, columns added or removed, frequency of updates.)		
	User can quickly assess meaning of each column or attribute. Supplemental document to User Guide.		
	Description adds context to every column name or attribute.		
	All codes listed in Expected Values or provided via link in the		

Data Dictionary	User Guide.	
	All Agency and City terms or acronyms defined in accessible language.	
	Data types are accurate and match Socrata's allowed values when possible.	

How to Write a User Guide and Data Dictionary

Templates and examples can be found on the [Open Data Coordinator Resource page](#) in the "Other Documents" section under "Metadata for All".

Step 1: Using a Google Account that is dedicated for work purposes, make a copy of the User Guide and Data Dictionary Templates. If you are unable to access Google Drive via your work computer please get in touch with the Open Data Team so that we can work with your IT to get you access.

Rename your files:

User Guide | Name of Dataset
Data Dictionary | Name of Dataset

Step 2: Follow Examples.

We've assembled example Data Dictionary and User Guides for datasets from three different agencies. Read through each to make sure you understand what we're asking for in each section.

> 311 (DoITT/311): Example of how to create usable metadata for a big and messy dataset. 311 data is shifting everyday, but our User Guide and Data Dictionary help users keep track of the changes, understand how the data is structured, and aware of what types of changes they can expect.
>
> Neighborhood Tabulated Areas (DCP): Geospatial data can be challenging for a new data user, but our User Guide and Data Dictionary break down complicated concepts with clear accessible language.
>
> Active Vehicles (TLC): TLC datasets reflect the huge operational challenges they meet to regulate the City's taxis and for-hire vehicles. Our User Guide and Data Dictionary give users insight into the processes, the policies, and the agency behind the data.

Step 3: Draft your User Guide and Data Dictionary.

We recommend filling out the Data Dictionary first, since this information will inform parts of your User Guide. Before beginning the User Guide, gather background information about your agency's internal related business practices. You can use a **Process Map** to visualize all the steps your agency takes to collect and manage your data. Consult the data owners of the dataset and other colleagues who might offer unique perspectives on the data. Try using the **Questionnaire** to help you brainstorm the right questions to ask about your dataset. Act as an investigative

reporter and treat the User Guide like your final report.

Step 4: Edit: Use Clear, Consistent Language.

When filling out the templates, write for a public audience who might not be familiar with the inner workings of your agency or institution. Think about what you needed to learn when you started your job. Consider all the programs, initiatives, or vocabulary you had to learn, that is specific to your field or workplace. Take advantage of the **Glossary** to help you define common terms like Borough or Community Board. Run language past someone outside your department to see they have any questions.

Step 5: Review your work. Use the **Rubric** or ask a colleague to use the Rubric to make sure that you have met all the Metadata for All Standards.

Step 6: Share your documents with the Open Data Team for publication. Using the Google Doc settings, share the links to your User Guide and Data Dictionary with the Open Data Team for publication. Be sure to change the sharing setting to *Anyone with the link* ***can view***.
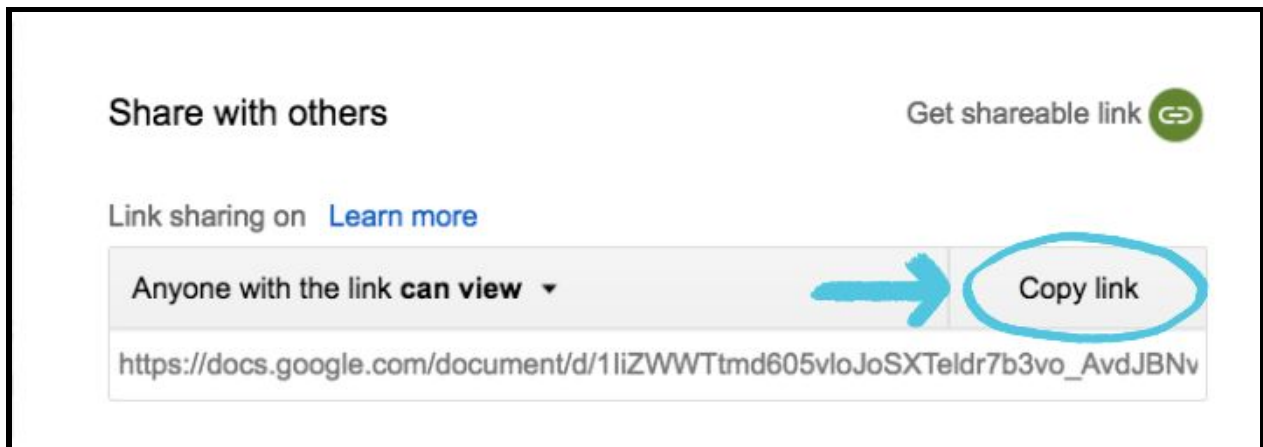
Additional Resources

In this section, we provide resources to help you incorporate user research into the development of Metadata for All User Guides and Data Dictionaries.

1. You can use the **Interview Questions** before you begin writing the User Guide, to help you brainstorm or facilitate a discussion with data owners at your agency.

2. The **Process Map** exercise can also help you brainstorm and draft an explanation of all the agency programs, policies, and people who influence a dataset.

3. Use the **Community Engagement Worksheet** to test how well your new User Guide and Data Dictionary help users form appropriate questions to ask with your data.

4. Save yourself time by using one of the standard definitions provided in our **Glossary**. (Or add one of your own to help other Open Data Coordinators!)

Additional Resource: Questionnaire

Use these questions to jumpstart your thinking process on how to develop valuable metadata, context around a dataset, to the public. For some datasets (larger, more popular, more complex), you may want to ask these questions directly to data owners within your agency. For other datasets, you may be able to answer these questions on your own. You can use your answers to fill out Data Dictionaries, User Guides and check the accuracy and completeness of your work.

For the User Guide

1. What does this data describe directly? What does it <u>not</u> describe?

   Good Example: The 311 dataset describes how, when, and where people are making 311 service requests. It does not measure the degree or frequency of problems like noise or trash, just the level of complaints.

2. How is data collected and maintained by your agency?

   Is data collection automated, if so how? Is data collection overseen by a particular role or department? Is data collected by survey, by field staff, by digital forms, etc? How often is data collected and/or updated? Are there people on staff who regularly check data for quality and completeness? Is data quality or completeness regulated by any laws or internal policies?

3. What business practices correspond to this data?

   Does your agency have any forms, surveys or data gathering processes that correspond to this data? Is data collection governed in any way by internal policies, City Laws, lawsuits or other legal guidelines?

4. What type of internal reports do you run with this data (or the internal version of this dataset)? What questions do these data answer? How do those reports support agency goals and workflow?

   Good Example: We report on how many wheelchair accessible vehicles are on the road. This data tells us if we're meeting accessibility requirements.

   Bad Example: We run Report #XL%K!#A.

   Do other agencies use this data, and if so how?

Good Example: At the Department of City Planning, we maintain Neighborhood Tabulation Area shapefiles that can be used by many other agencies to inform their work.

If this data is compiled exclusively for external public use, what datasets did this data come from and how is that internal data used?

Good Example: We made this data available after receiving a request from the public. However DSNY is moving away from tracking recycling bins separately from all other kinds of waste. For more accurate data that reflects our work, see THESE OTHER DATASETS: <link>, <link>, <link>.

5. Who are the data owners for this data? Why?

   Good Example: The Director of Finance owns this dataset because she oversees how money is being spent on this particular program.

   *Bad Example: Stephanie owns this dataset because she runs the #*IABC@D report and she's the only one who understands it.*

6. What do you or the data owners wish was better about this data? Select all that apply.

   a. More fields?
   b. More complete rows?
   c. Longer date range?
   d. Less redundancy?
   e. More clarity around what each column or field means?

7. What are your future plans for this data?

8. Describe how this dataset differs from the internal version? Select all the apply

   a. Is it anonymized?
   b. Are there more or fewer columns or fields?
   c. Have values been edited or transformed for quality?
   d. Has the data been aggregated?
   e. There is no internal version of this dataset.

   Please give any details you can about each selection:

9. What values are missing from this dataset and why?

For the Data Dictionary

1. Which fields are required?

2. Which fields are provided by the agency and which are added by DoITT?

3. Does the range of values for each column match the Expected Values in the Data Dictionary?

4. Does your agency have internal data standards or codes that need to be translated or explained?

5. Have the columns or fields been updated or changed over time?

6. Have the rows been updated or changed over time?

Additional Resource: Process Map Worksheet

Use this exercise to spark your thinking or facilitate a conversation about how your agency's data corresponds to your agency's internal business practices.

A process map illuminates all the steps or decisions in a typical operation. The Metadata for All Team used process maps to illustrate all the internal practices that influence or correspond to the creation and publication of open data.
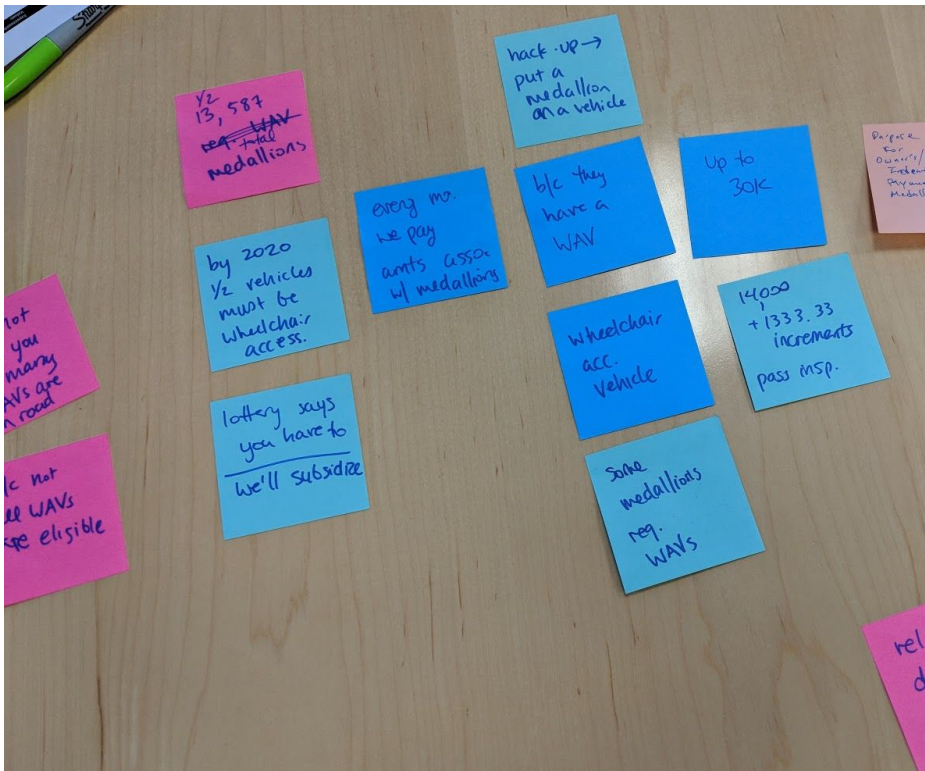
Example: This Process Map exercise helped us make connections between the TLC's Wheelchair Accessible Vehicle program and the [Taxi Improvement Fund (TIF) Medallion Payments](#) dataset.

Task One

● Grab a Data Dictionary and some sticky notes in two different colors (Color A and Color B).
● Identify what business processes relate to this dataset. Is there a corresponding form? Policy? Regulation?
● Write each step in the process down on a separate sticky note in Color A.

Task Two

- Identify who at your agency could tell you about each business practice outlined on each sticky note. Write their names or positions down, one on each sticky note in Color B.

Task Three

- Arrange all of the post-its to tell the story of how the dataset was created or reflects the work done at your agency

BONUS: Brainstorm a list of people in the community who might be interested to hear this story

Additional Resource: Community Engagement Workshop

You can learn a great deal about how to improve your Data Dictionaries and User Guides by engaging with your community. Get in touch with your agency's public relations / communications / community engagement unit to see how you can partner to get feedback on your datasets and metadata. You'll need the following to conduct this workshop:

1. *A room that can fit 10-15 people booked for 2 hours; the room should have tables and chairs. If the room has computers that is an added bonus as the workshop participants can look up the dataset (if it's already published) directly online.*
2. *A list of constituents to invite to the workshop*
3. *A draft of an email invitation and reminders; send the email invitation at least 2 weeks before the workshop*
4. *A way for constituents you are trying to engage to sign up or RSVP for the workshop*

In the workshop use the example worksheet below, which was originally written for the 311 dataset, to understand how well the public is able to use your data. In order to use this worksheet:

1) *Rewrite the questions so that they apply to datasets from your agency*
2) *Hand out printouts of the worksheet along with a printout or link to your User Guide and Data Dictionary. Ask users to complete the questions using the User Guide and Data Dictionary as a resource.*
3) *Facilitate a conversation in which you get feedback on the User Guide and Data Dictionary. How well did these resources help users come up with good data analysis questions?*

Worksheet Example for the 311 Service Request Dataset

Exercise One: Trash Talk

Your trash hasn't been picked up this week and you want to know what's going on. Is this a trend across NYC? Are other neighborhoods getting better service?

Use your Data Dictionary and your imagination to come up with at least three (3) questions that will help you understand the state of trash pick-up in NYC.

1.

2.

3.

Choose one (1) question from your above list. What columns can you use to help you answer your question? How would you filter those columns?

Exercise Two: Please Won't You Be My Neighbor?

You're moving to another area of NYC. How can you use 311 Complaints to learn more about your new neighborhood?

Identify three (3) issues you might look up about that neighborhood, using the 311 dataset?

1.

2.

3.

Choose one (1) question from your list above. What columns can you use to help you answer your question? How would you filter those columns?

Exercise Three: Check Your Gut

What do you think the top 311 complaint has been in your neighborhood during the past month?

What columns can you use to help you answer your question? How would you filter those columns?

Additional Resource: Glossary

The terms below appear in many datasets across NYC Open Data. These standardized definitions will save you time in writing Data Dictionaries and User Guides and help standardize metadata across NYC Open Data.

Boro Code / Borough Code: A single digit identifier, indicating a particular borough. 1 = Manhattan, 2 = Bronx, 3 = Brooklyn, 4 = Queens, 5 = Staten Island. Boro codes may appear as a single digit, or may comprise part of a larger city code, such as a Borough Block and Lot (BBL) number or Business Identification Number (BIN).

Borough: New York City's boroughs are five county-level administrative divisions, with each one also being a state county. (Manhattan - New York County; Bronx - Bronx County; Brooklyn - Kings County; Queens - Queens County; Staten Island - Richmond County). NYC's boroughs have existed since the consolidation of the city in 1898, all five boroughs together make up New York City

Building Identification Number (BIN): A unique 7-digit number assigned to every known building by the Department of City Planning (DCP), the first digit of which is the borough code. BINs allow city agencies to process and match building-related data easily and in a consistent manner.

Borough Block Lot (BBL) number: A combination of three numeric codes -- a 1-digit borough number, a block number (up to 5 digits) and a lot number (up to 4 digits) -- designated and modified by the Department of Finance (DOF). BBLs are used by various city agencies to identify real estate for taxes, zoning, construction, and other purposes.

Community Board / Community District: Community Boards are volunteer-run neighborhood organizations that work on all kinds of issues pertinent to their communities, often working with elected city officials in doing so. Community Board members, who can be as young as 16, are appointed by Borough Presidents and City Council representatives. They serve two-year terms. There are 59 Community Boards in NYC, each of which represents a unique geographical area, known as a Community District. Find your own Community District here. *Note: as a column header, "Community Board" (which refers to a group of Board members) is often used to mean "Community District" (the geographical area that a Community Board governs) instead.*

City Council District: The City Council is the lawmaking body of NYC, on equal footing with the Mayor in terms of governing power. Besides legislating, the Council has sole approval power over the City budget, and is the final decision-maker in land use matters. There are 51 City Council members in total, each representing a unique geographical area, called a Council District. To find your own Council District, go here.

Federal Information Processing Standard Code (FIPS): A five-digit code (2-digit state prefix + 3 digit county code) which uniquely identifies counties and county equivalents in the United States.

Neighborhood Tabulation Areas (NTAs): NTAs are small area boundaries, created by the Department of City Planning (DCP) to aggregate population projections in a small area. Each NTA approximates a minimum population of 15,000. While NTAs were initially created to support PlaNYC, the thirty-year (2000-2030) sustainability plan for NYC, NTAs are now also being used to present data from the Decennial Census and American Community Survey. NTA boundaries and their associated names do not definitively represent cultural neighborhood boundaries.

Public Use Microdata Areas (PUMAs): PUMAs were created by the Department of City Planning (DCP) to aid city agencies in administering public services, and to support PlaNYC, the thirty-year (2000-2030) sustainability plan for NYC. They're also used for disseminating American Community Survey (ACS) estimates. PUMAs have a minimum population of 100,000, are aggregated from census tracts, and are either approximate Community Districts, or combinations of Community Districts.