## What is a Public Dataset?

A “dataset” is a collection of tabular information where each row represents a record and each column represents details about that record. Columns may contain numbers, text, or dates, for example.

Datasets may exist in a dedicated database system such as Oracle, MS SQL Server, or Postgres; they may be part of a larger software application system; or they may be a standalone file in one of these common formats: csv (text file where each record is a new line and columns are separated by commas), excel, json, xml, geojson, kml, and esri formats.

### Examples of Public City Datasets

|  |  |
| --- | --- |
| **Transactional Data** | ***311 Service Requests***: each record represents a service request and the columns contain information on when it was made, the type, and which agency it was directed to. Other examples: ***DCA License Applications*** and ***DOB Job Application Filings***. |
| **Inventory Data** | ***DCP’s PLUTO*** contains information on all tax lots in the city. Each record is a tax lot and the columns represent information about the lot, such as location, number of buildings, and zoning. Other examples*:* ***DPR’s Street Tree Census*** and ***DCAS Managed Public Buildings***. |
| **Operations Data** | ***HPD’s Housing Maintenance Code Violations***: each record represents a violation issued by HPD. Other examples: ***DSNY’s Recycling Diversion and Capture Rates*** and ***Vacant Lots Cleaned***. |

### Identifying New Datasets

All public data must be on the Open Data Portal by the end of 2018. Here are some ideas for finding new datasets that might not already be on the Open Data Portal or listed in the agency compliance plan:

* **The Mayor’s Management Report (MMR) performance indicators**: What data underlying MMR indicators or other KPIs can be released on the Open Data Portal? Is it possible for those metrics to be calculated using data already on the Open Data Portal? If not, what data can be released to make this possible?
* **Data on Websites**: The law requires that all data on websites maintained by or on behalf of your agency also be on the Open Data Portal. Please review your agency’s websites.
* **FOIL responses**: Any datasets released through FOIL should also be considered for release on the Open Data Portal. Please work with your agency’s FOIL officer. This can also cut down on future FOIL requests.

### Best **Practices**

* **Datasets should be as granular as possible**. Each row should represent a single record or occurrence, not an aggregation of data.
* **Include as much of the source dataset as possible**. Exceptions may include columns that contain private or sensitive information.
* **Include as much historical data as possible.** Go as far back as there are records in the same schema. If there exists historical data in a different schema, release it as a separate dataset.
* **Data dictionaries should be understandable to the average New Yorker**. They should contain information such that someone not familiar with the agency’s operations can understand how and why the data is collected and what each column represents. Any acronyms or technical word should be defined. The more understandable the data is, the less time you and your agency will spend responding to questions.