Data Quality Score

In response to the growing number of datasets on Open Data, in December 2020 the Data Quality Score was launched to provide a numeric indicator for the quality of each dataset on the platform. This is accomplished by 1. Creating a new dataset (https://data.winnipeg.ca/dataset/Data-Quality-Score/73sq-j2qi) to list the score (1 to 100) and rank (gold, silver, bronze) of each dataset, and 2. Creating two new Socrata metadata attributes Quality.Score and Quality.Rank to be populated by the daily refresh system.

As of 2020-12-24, these are the current Score to Rank relations. They are subject to change however.

- 0-70: Bronze
- 71-80: Silver
- 81-100: Gold

Scoring Metrics

Table of all individual used in calculation of the total data quality score. The score is a weighted sum of each metric.

Dataset Column Name	Database Column Name	Weight	Scoring
Quality Level	Quality Level	0.2	Score determined by putting dataset into one of three categories based on the displayed data. • Raw: Data is displayed in the most raw form possible, 100% • Aggregated: Data is aggregated from the source for any reason not related to obscuring personal data. 50% • Filtered: Selected data is filtered out from the source to Open Data for any reason not related to obscuring personal data. 0% Any dataset that has not been put into one of these categories is given a 0% for this category.
Stability	DATASET_STABILITY	0.2	Score is based on the last time a column has changed or been removed from the dataset, with the following formula: 1 - (days_since_change - 31)/(31*6 - 31) This score is bounded with:

			0% at days_since_change < 31 100% at days_since_change > 31*6 If there has never been a changed or deleted column, the score is 100%
Scheduled Update	FRESHNESS_SCHEDULED_UPDATE	0.1	Score is based on the dataset's "Update Frequency" metadata attribute. • Real-time: 100% • Daily: 80% • Weekly: 70% • Monthly: 50% • Yearly: 20% • Census Period: 10% • As Required: 10% null/other: 0%
Dataset Update	FRESHNESS_DATASET_UPDATE	0.2	Score is based off the dataset's scheduled update frequency and the last time the dataset was updated. • If the dataset update time is within the scheduled update time frame, score is 100% • If the dataset update time has missed the scheduled update time frame, the score linearly decays until its missed two update time frames • If the source update time has missed two or more time frames, the score is 0%
Source Update	FRESHNESS_SOURCE_UPDATE	0.2	Score is based off the dataset's scheduled update frequency and the last time the source was updated. • If the source update time is within the scheduled update time frame, score is 100% • If the source update time has missed the scheduled update time frame, the score linearly decays until its missed two update time frames • If the source update time has missed two or more time frames, the score is 0%

			If the dataset has a SOURCE_UPDATE_EXEMPT attribute of Y, score is 100% If the dataset does not contain a column to determine the source update time, score is 0%
Has Metadata	HAS_METADATA	0.1	Score is based on the number of metadata attributes filled out. Each attribute contributes a specific percentage of the score. • Tags: 20% • Steward: 10% • Custodian 10% • ETL Location 10% • Description 10% • Update Frequency 10% • Source Location 5% • Category 5% • Department 5% • Department Group 5% • Canada Open Data Licence 5% • Winnipeg Open Data Licence 5%
Meaningful Column Names	MEANINGFUL_COLUMN_NAMES	0.2	Score is the percentage of the dataset's columns that are deemed "meaningful". This is determined by a column fulfilling at least 1 out of 2 requirements: 1. The column has the metadata attribute "description" filled out 2. The column is composed solely of English words