Table of Contents

Data Structure and Formats

Standard Reference Data

Appendix

Data Standards Reference Handbook (Beta Release)

DataSF operates the City and County of San Francisco's official open data portal. We are documenting standards to make data more useful and consistent across the City at scale.

This document serves several purposes:

1. Introduce more consistency within the open data publishing process
2. Provide a single enduring, open document to help onboard new staff
3. Provide a reference for data publishers and users
4. Clarify departmental stewardship of certain reference data

We lean heavily on existing precedent where available. The scope includes:

1. Formats and data structure
2. Common reference standards (lists) that are useful across datasets and departments

We are not including domain-specific standards here like Open311 or LIVES which have their own documentation and communities. We are also not using this to propose new domain specific standards.

Throughout this guide we reference standard names and lists, please refer to the appendix for reserved column names and an index of reference data.

Data Structure and Formats

This section covers format and structure standards for datasets being shared with others. These standards are designed to make sure that field level information is shared as consistently as possible to minimize:

1. Errors
2. Rework
3. Repetitive questions

> Many thanks to Singapore's Open Data Program for providing a Data Quality Guide for Tabular Data the bulk of which made its way into this chapter with some additions and modifications.

Column Headers

- Only use alphanumeric or these 3 special characters: period (.), dash (-), and underscore (_)
    - Ampersand (&) should be replaced by “and” if needed
- Each must be unique
    - Can’t have two headers called "duration"
- Units of measure should be omitted
    - Units can and should be provided with the data dictionary
- Keep short (less than 30 characters)
    - A full description can and should be provided with the data dictionary

Column Order

- *Unique identifiers* should be in the left-most column if applicable
- *Date and time variables* should be in the first column for time series data
- *Fixed or classified variables* should be ordered with the highest-level variable on the left and most granular variable on the right, for example
    - 311 cases: service_name, service_subtype, service_details
    - Police incidents: category, descript
- *Observed variables* should always be on the rightmost columns, these are measured variables often numeric, for example:
    - Duration
    - Number of Units
    - Number of Stories
    - Year Built
    - People Served

Is anything wrong, unclear, missing?

Leave a comment.

Date and Time

- Based on ISO8601, an international standard for representing date and time. We chose the "extended format" with the hyphens because it is more human readable.
  - Compare 2016-01-01 to 20160101
- All date and time variables must be local time (UTC -8hrs Pacific Standard Time UTC -7hrs Pacific Daylight Savings Time) unless specified.

Date variables

Interval	Column name	Format	Range of values	Example
Annual	`year`	YYYY	YYYY: 1776 onwards	2015
Monthly	`month`	YYYY-MM	MM: 01 to 12	2015-01
Daily	`date`	YYYY-MM-DD	DD: 01 to 31	2015-01-01
Weekly	`week`	YYYY-[W]WW	[W]WW: W01 to W52	2015-W01
Quarterly	`quarter`	YYYY-[Q]Q	[Q]Q: Q1 to Q4	2015-Q1
Half-yearly	`half_year`	YYYY-[H]H	[H]H: H1 or H2	2015-H1

For fiscal periods, prefix “fiscal_” to column name

Interval	Column name	Format	Example
Fiscal, annual	`fiscal_year`	YYYY	2015
Fiscal, monthly	`fiscal_month`	YYYY-MM	2015-01
Fiscal, quarterly	`fiscal_quarter`	YYYY-[Q]Q	2015-Q1
Fiscal, half-yearly	`fiscal_half_year`	YYYY-[H]H	2015-H1

- Fiscal year start date must be indicated in the data dictionary
  - e.g. The fiscal year starts on July 1 and ends on June 30 for the City and County of San Francisco

Date-time and time variables

- ISO 8601 uses 24 hour clock system in hh:mm:ss format (do not use AM or PM)
- e.g. 13:00 is equivalent to 1:00 PM

Type	Column name	Format	Example
Date + time	`date_time`	YYYY-MM-DD[T]hh:mm	2015-01-01T13:00
		or YYYY-MM-DD[T]hh:mm:ss	2015-01-01T13:00:00
Time only	`time`	hh:mm	13:00
		or hh:mm:ss	13:00:00

Specify the timezone if it is not local time (UTC -8hrs Pacific Standard Time UTC -7hrs Pacific Daylight Savings Time):

Type	Column name	Format	Example
Date + time	`date_time`	YYYY-MM-DD[T]hh:mm+hh:mm	2015-01-01T12:00+00:00
		or YYYY-MM-DD[T]hh:mm:ss+hh:mm:ss	2015-01-01T12:00:00+00:00:00

Date and time extracts

In certain cases you may want to provide a single variable representing the number or name of an individual date component, a day, a month, etc. There's no requirement to provide these, but follow this guidance:

Extract	Column name	Type	Range of values
Year	year_num	integer	any valid year
Month	month_num	integer	1 to 12
Month Name	month_name	string	January, February, March, April, May, June, July, August, September, October, November, December
Week of Year	woy_num	integer	0 to 51
Day	day_num	integer	1 to 31 (varies by month)
Day of Week	dow_num	integer	0 to 6
Day of Week Name	dow_name	string	Monday, Tuesday, Wednesday, Thursday, Friday, Saturday, Sunday
Hour	hour_num	integer	0 to 23
Minute	minute_num	integer	0 to 59
Second	second_num	integer	0 to 59

These can often be automatically extracted from a valid ISO-8601 date, for example the open data portal enables querying a dataset with these date extract functions:

- date_extract_d() - extracts the day from a date as an integer
- date_extract_dow() - extracts the day of week as an integer between 0 and 6 (inclusive)
- date_extract_hh() - extracts the hour of the day as an integer between 0 and 23 (inclusive)
- date_extract_m() - extracts the month as an integer
- date_extract_mm() - extracts the minute from the time as an integer
- date_extract_ss() - extracts the second from the time as an integer
- date_extract_woy() - extracts the week of the year as an integer between 0 and 51 (inclusive)
- date_extract_y() - extracts the year as an integer

Durations

Durations can be automatically calculated if you provide a separate start and end period in your dataset. If you also want to provide a duration, please:

- Provide the milliseconds between the start and end period (include the duration unit in

the data dictionary)
  - Milliseconds can be rolled up to other time intervals
- Use duration in your column name but prepend with a useful descriptor, e.g:
  - flight_duration
  - response_duration
  - dwell_time_duration
  - travel_duration
- Do not duplicate any of the duration column names per the guidance on columns

> **Note:** ISO 8601 does have separate guidance on duration formatting, but we find this more cumbersome than just calculating milliseconds between a period for which there are many standard programming libraries.

Is anything wrong, unclear, missing?

Leave a comment.

Text

- UTF-8 encoding should be used
    - This ensures that special characters can be decoded by users
- No line breaks within cells
    - This can break parsing in software like Excel, introducing data integrity issues
    - There are many ways to remove and detect line breaks, but this can vary based on how you're extracting data

Considerations for categorical variables

- Please maintain consistency with canonical and standard reference lists
    - This helps with analysis across departments and data systems
- Common reference lists are provided within this document, including the departmental steward of the list where applicable and links to the data

Character case

Text should be presented in the easiest to interpret/read format where appropriate.

Title case

- Address String
- Categories when either the source system presents them this way or it is easy to interpret from the source

Upper case

- Acronyms - e.g - PSA (Park Service Area)
- States - e.g. CA

Lower case

- Categories when the source system presents them in caps and there's no way to interpret them to title case
- Research suggests lower case as opposed to uppercase is easier to read for humans and just as useful to machines, note exceptions above

Is anything wrong, unclear, missing?

Leave a comment.

Numeric

- No commas
    - e.g. "1000" instead of "1,000"
- No units of measurement
    - Units should be in metadata instead
- Express as full number where possible
    - e.g. "1200000" instead of "1.2" (million)
    - If rounded, indicate in metadata
- No rounding if possible
    - Give raw numbers as far as possible
    - If rounding is needed, try to provide at least 2 decimal places of precision
- Percentages can be expressed as either a proportion out of 1 or 100.
    - e.g. 20% can be expressed as 20 or 0.2
    - The representation of percentages must be consistent throughout your dataset (e.g. among different percentage fields)
    - Agencies must indicate how percentages are expressed in the data dictionary

Is anything wrong, unclear, missing?

Leave a comment.

Location (coordinates)

- Coordinates in EPSG 4326 or EPSG 2227
- Only EPSG 4326 coordinates can be mapped within the open data portal
- Should be represented in two columns
  - EPSG 4326: `latitude` and `longitude` or
  - EPSG 2227: `x_coord` and `y_coord`
  - **Note:** all EPSG 4326 coordinates will be loaded into the open data portal to support mapping and presented in an additional single location column there called `the_geom`. EPSG 2227 coordinates will be represented as the two original columns
- In positive/negative floating point
  - e.g. `latitude`: 37.761146; `longitude`: -122.436235
- EPSG should be indicated in metadata

Is anything wrong, unclear, missing?

Leave a comment.

Location (addresses)

Why valid addresses matter

- Consistent formatting of valid addresses is important for accurately mapping and referencing geographic information
- A poorly formed address could end up mapping to the wrong geographic reference or not at all, reducing the usefulness of the data
- Poorly formed addresses can make cleanup of data labor intensive and result in reporting errors where geography (neighborhoods, census, etc.) is concerned
- Poorly formed addresses could also result in additional costs because of things like:
    - Undeliverable/returned mail
    - Failure to apply benefits to recipients appropriately based on geography
    - Poor routing of vehicles or people in the field

Address formatting

- Addresses should be output with the level of detail relevant to the data
    - e.g. permits can be applied down to the sub-address level
- If providing addresses in a complete string, make sure the addresses are well formed and consistent for easy parsing, for example:
    - 741 Ellis Street, Unit 5, San Francisco, CA 94109
    - 901 Bayshore Boulevard, Unit 209, San Francisco, CA 94124
- When providing multiple addresses within a dataset, prepend your column names with the type of address
    - e.g. address vs. mailing_address (see Registered Businesses dataset)
- Where appropriate, use a valid Enterprise Address System address
    - EAS addresses capture addresses input by DBI staff, see the section on address numbers for more detail

Address elements

Below are some common elements of an address (but not all)

- Not all addresses will have all elements
- Address granularity will be driven by the business need, so not all systems will collect

every element
  - Note: systems can be designed to validate or lookup addresses on entry, minimizing error
- Make sure the individual elements of an address line up with the guidance below
- You can publish addresses as either single strings or break into separate fields

> **Note:** this guidance is provided to promote consistency across the bulk of shared tabular datasets and not as a comprehensive guide to address standards. For a comprehensive standard on addressing, see the Federal Geographic Data Committee (FGDC) United States Thoroughfare, Landmark, and Postal Address Data Standard

Element	Data Type	Definition	Valid Values
From Address Number	Numeric	First part of a range: **1000-**1100 Main Street, San Francisco, CA 94102	For each street centerline on the right side: `rt_fadd` ; on the left side: `lf_fadd`
To Address Number	Numeric	Second part of a range: 1000-**1500** Main Street, San Francisco, CA 94102	For each street centerline on the right side: `rt_fadd` ; on the left side: `lf_fadd`
Address Number Prefix	Numeric	The portion of the Complete Address Number that precedes the Address Number itself: **B**315 Main Street, San Francisco, CA 94102	Official address numbers available through the Enterprise Address System as `address_number`
Address Number	Numeric	The numeric identifier for a land parcel, house, building, or other location along a thoroughfare or within a community: **315**A Main Street, San Francisco, CA 94102	Official address numbers available through the Enterprise Address System as `address_number`
Address Number Suffix	Text	The portion of the Complete Address Number that follows the Address Number itself: 315 **A** Main Street, San Francisco, CA 94102	Official address numbers available through the Enterprise Address System as `address_number_suffix`
Street Name Pre Modifier	Text	A word or phrase in a Complete Street Name that 1. Precedes and modifies the Street Name, but is separated from it by a Street Name Pre Type or a Street Name Pre Directional or both, or 2. Is placed outside the Street	Official list of street names maintained by

Pre Modifier		Name so that the Street Name can be used in creating a sorted (alphabetical or alphanumeric) list of street names.: 315A **Old** Main Street, San Francisco, CA 94102	Public Works
Street Name Predirectional	Text	A word preceding the street name that indicates the directional taken by the thoroughfare from an arbitrary starting point, or the sector where it is located: 315A **East** Main Street, San Francisco, CA 94102	Official list of street names maintained by Public Works
Street Name Pretype	Text	A word or phrase that precedes the Street Name and identifies a type of thoroughfare in a Complete Street Name: **US Route** 101, San Francisco, CA	Official list of street names maintained by Public Works
Street Name	Text	The portion of the Complete Street Name that identifies the particular thoroughfare (as opposed to the Street Name Pre Modifier, Street Name Post Modifier, Street Name Pre Directional, Street Name Post Directional, Street Name Pre Type, Street Name Post Type, and Separator Element (if any) in the Complete Street Name.): 315A **Main** Street, San Francisco, CA 94102	Official list of street names maintained by Public Works
Street Name Posttype	Text	A word or phrase that follows the Street Name and identifies a type of thoroughfare in a Complete Street Name: 315A Main **Street**, San Francisco, CA 94102	Official list of street names maintained by Public Works
Street Name Postdirectional	Text	A word following the street name that indicates the directional taken by the thoroughfare from an arbitrary starting point, or the sector where it is located: 315A Main Street **East**, San Francisco, CA 94102	Official list of street names maintained by Public Works

Street Name Post Modifier	Text	A word or phrase in a Complete Street Name that follows and modifies the Street Name, but is separated from it by a Street Name Post Type or a Street Name Post Directional or both: 315A Main Street **Extended**, San Francisco, CA 94102	Official list of street names maintained by Public Works
Occupancy Type	Text	The type of occupancy to which the associated Occupancy Identifier applies. (Building, Wing, Floor, Apartment, etc. are types to which the Identifier refers.): 315A Main Street, **Apt** 2, San Francisco, CA 94102	There is no complete reference of subaddresses (aka units) at the time. You can refer to Enterprise Address System addresses with units for a partial list.
Occupancy Identifier	Text	The letters, numbers, words, or combination thereof used to distinguish different subaddresses of the same type when several occur within the same feature: 315A Main Street, Apt **2**, San Francisco, CA 94102	There is no complete reference of subaddresses (aka units) at the time. You can refer to Enterprise Address System addresses with units for a partial list.
City	Text	The city the address sits within: 315A Main Street, **San Francisco**, CA 94102	
State Name	Text	The names of the US states and state equivalents: the fifty US states, the District of Columbia, and all U.S. territories and outlying possessions. A state (or equivalent) is "a primary governmental division of the United States." The names may be spelled out in full or represented by their two-letter USPS or ANSI abbreviation: 315A Main Street, San Francisco, **CA** 94102	Recommend using standard abbreviations. Spell out if you can do so without introducing misspellings (e.g using validated entry).
ZIP code	Numeric	A system of 5-digit codes that identifies the individual Post Office or metropolitan area delivery station associated with an address: 315A Main Street, San Francisco, CA **94102**	Note, zip codes are not actually boundaries, but are defined by routes. A list of valid San Francisco zipcodes can be downloaded here.

ZIP+4	Numeric	A 4-digit extension of the 5-digit Zip Code (preceded by a hyphen) that, in conjunction with the Zip Code, identifies a specific range of USPS delivery addresses: 315A Main Street, San Francisco, CA 94102-**1212**	Note, zip codes are not actually boundaries, but are defined by routes. A list of valid San Francisco zipcodes can be downloaded here.

Is anything wrong, unclear, missing?

Leave a comment.

Reference Data Overview

Reference data generally refers to an authoratative list of permissible values to be used in other data. It may also refer to standards of collection methods against different lists as often is the case with demographic information.

Reference data, unlike transactional data, will change less frequently and will often have a controlled process for changes; for example, the addition of official addresses is controlled through the permitting process.

These pages are designed to improve discoverability and documentation of some of the most common references used across city data. This should be useful to data users, but also data publishers as they make decisions about how to disseminate data. Additionally, this can be used by those developing new systems.

The following pages are grouped into several related sections:

1. **General Admin**. Reference lists used in the administration of City business. For example, in the City financial system.
2. **Demographics**. Reference lists used to capture demographic information in systems or on surveys.
3. **Basemap**. References generated or used in the production of basemap data including parcel numbers, street names, and address numbers.
4. **Boundaries**. References that refer to common boundaries like census areas, neighborhoods and supervisor districts.

Reference: General Admin

This section covers any references that are used in the administration of City business that don't fall into the other reference categories. For example, categories used in the financial system of record.

Department Names and Codes

Definition

The City and County of San Francisco is made up of many organizations that perform work and deliver services according to the charter and administrative codes of the City.

These organizations have common names, but also codes that are used in accounting for the work and services performed. The Controller's Office maintains these codes in the Executive Information System (EIS) where department staff maintain records related to spending, revenue and budget among other things. Other enterprise systems use these codes to link administrative data among departments.

Reference

Dataset	Description and Constraints	Reference Columns

Department Code List	These department codes are maintained in the City's Financial System of Record. Department Groups, Divisions, Sections, Units, Sub Units and Departments are nested in the dataset from left to right. Each nested unit has both a code and an associated name.
	The dataset represents a flattened tree (hierarchy) so that each leaf on the tree has it's own row. Thus certain rows will have repeated codes across columns. Data changes as needed.
		Nested (right to left):
		`department_group_code`
		`division_code`
		`section_code`
		`unit_code`
		`sub_unit_code`
		`department_code`

Reference: Demographics

Where standards or references exist for demographic information, we include those here. Currently this covers Sexual Orientation & Gender Identity.

Sexual Orientation and Gender Identity

Below are standards on how to collect sexual orientation and gender identity (SOGI) If your department does not have a standard, you are encouraged to use one of the standards below.

San Francisco Standards

Administrative Code Chapter 104

Chapter 104 of the San Francisco code requires the collection of SOGI data by select departments consistent with Department of Public Health guidelines. View the code for the full set of requirements. Below is an overview.

Description of Standard

Select departments are required to solicit SOGI data consistent with Department of Public Health's Policies and Procedures:

- "Sexual Orientation Guidelines: Principles for Collecting, Coding, and Reporting Identity Data," reissued on September 2, 2014
- "Sex and Gender Guidelines: Principles for Collecting, Coding, and Reporting Identity Data," reissued on September 2, 2014
- or any successor Policies and Procedures

Sexual Orientation Guidelines

Below is a brief and incomplete excerpt - please review the full set of guidelines before using this standard.

> When collecting data on sexual orientation, the following format should be followed:
>
> - Selection of sexual orientation identity should be limited to one answer choice.
> - How do you describe your sexual orientation or sexual identity? (Check one)
>   - a. Straight / Heterosexual
>   - b. Bisexual
>   - c. Gay / Lesbian / Same-Gender Loving
>   - d. Questioning / Unsure
>   - e. Not listed. Please specify: ________________________
>   - f. Decline to answer
> - And for internal use only (not to be listed as an option to the individual):
>   - g. Not Asked
>   - h. Incomplete / Missing data

Sex and Gender Guidelines

Below is a brief and incomplete excerpt - please review the full set of guidelines before using this standard.

> Two questions should be used together to identify sex and gender. You should ask these two questions, together as follows and in this order, to acquire sex and gender demographics about both the person’s present gender identity and his or her history.
>
> 1. What is your gender? (Check one that best describes your current gender identity.)
>   i. (1) Male
>   ii. (2) Female
>   iii. (3) Trans Male
>   iv. (4) Trans Female
>   v. (5) Genderqueer / Gender Non-binary
>   vi. (6) Not listed, please specify____________________
>   vii. Survey forms would include options 1-6. Coding should also allow for options 7 and 8
>   i. (7) Declined / Not stated
>   ii. (8) Question Not Asked
> 2. What was your sex at birth? (Check one)
>   i. (1) Male
>   ii. (2) Female
>   iii. Survey forms would include options 1-2. Coding should also allow options 3 and 4
>   iv. (3) Declined / Not stated
>   v. (4) Question Not Asked

Definitions

"Gender Identity" means a person's gender as designated by that person. A person's gender identity shall be determined based on the individual's stated gender identity, without regard to whether the self-identified gender accords with the individual's physical appearance, surgical history, genitalia, legal sex, sex assigned at birth, or name and sex as it appears in medical records, and without regard to any contrary statement by any other person, including a family member, conservator, or legal representative. An individual who lacks the present ability to communicate his or her gender identity shall retain the gender identity used by that individual prior to losing his or her expressive capacity.

> From Section [3304.1] (c) of the Police Code

"Sexual orientation" shall mean the status of being lesbian, gay, bisexual or heterosexual.

> From Section [12B.1] (c) of the Administrative Code.

Who must Comply

The following departments must comply with Administrative Code Chapter 104. See the code for details on exceptions, the official list of required departments, and other requirements. Other departments may use this standard as helpful.

- Department of Public Health
- Department of Human Services
- Department of Aging and Adult Services
- Department of Children, Youth and their Families
- Department of Homelessness and Supportive Housing
- Mayor's Office of Housing and Community Development.Requirements

These departments are also required to flow down the standard to their contractors and service providers. The code provides more detail on these provisions.

Authority

San Francisco Administrative Code Chapter 104: Collection of Sexual Orientation and Gender Identity Data.

California Standards

At this time we do now know of any California standards. However, CA Government code 8310.8 (cited as the Lesbian, Gay, Bisexual, and Transgender Disparities Reduction Act) requires certain state departments (listed below) to collect SOGI data. It does not however specify how they should collect that data or even that they should do it consistently. These state agencies may flow down SOGI data collection requirements to your department for purposes of state data collection.

> - (a) (1) This section shall only apply to the following state departments:
>   - (A) The State Department of Health Care Services.
>   - (B) The State Department of Public Health.
>   - (C) The State Department of Social Services.
>   - (D) The California Department of Aging.

Race and Ethnicity

Background and Overview

Background

The concepts of race and ethnicity are not concrete. They represent social-political constructs that evolve over time and are subject to the perceptions of self and others. View a timeline of changes in race and ethnicity in the US Census from 1790-2010.

As a result, there is no perfect standard for race and ethnicity. The standardization of race and ethnicity data represents a tension between (1) collecting race and ethnicity data to maximize opportunities to self-identify, self-describe, or place oneself within a group that feels welcoming and right, and (2) collecting data that decision makers and the public can use effectively to advance social justice and civil rights.

The changing nature of society’s understanding of race and ethnicity presents an ongoing challenge to how it is captured.

At this time, San Francisco does not have a standard, required method for collecting race and ethnicity data. As a result, methods vary not only by department but by program or data system. Methods in place may be an artifact of reporting expectations, system defaults or historic decisions.

The purpose of this section is to provide guidance as follows:

- Define a recommended standard given the latest research and testing on race and ethnicity data collection methods and to promote consistent data collection over time
- Provide information on other standards that are available and may be used as alternatives to the recommended standard

Overview of Standards

Below is an overview of the race and ethnicity data standards covered in this section.

Jurisdiction	Title	Who should use
City and County of San Francisco	San Francisco Recommended Standard	Departments should comply with this standard unless they face conflicting requirements. Note that external reporting fields are not requirements. Appendix E provides a rationale for these recommendations.
City and County of San Francisco	Department of Public Health's Ethnicity Guidelines	All new data collection systems purchased or designed for or by the Department of Public Health.
State of California	Racial and Identity Profiling Act of 2015 Regulations	The Police Department is required by state law to use this in the context of collecting data on stops. This standard is not required for other types of data collection, including in the Police Department, and may not be appropriate as it was designed to capture perceived race/ethnicity.
Federal Government	Standards for the Classification of Federal Data on Race and Ethnicity (rev. 1997)	This standard does not face City Departments. This is included for reference as this may flow down to City departments via federal reporting requirements.

City and County of San Francisco

San Francisco does not have a citywide standard. We include a recommended standard and the existing guidance from the Department of Public Health.

San Francisco Recommended Standard

Description of Standard

The purpose of this data standard is to support the consistent collection, maintenance and reporting of data on race and ethnicity across Departments. Consistent race and ethnicity data will:

- Improve our ability to track and compare differences across City services and programs
- Help inform policy and procedural changes to reduce disparities across City services and programs

The categories in this standard come from the Census 2015 National Content Test and like the Census are not genetically, anthropologically, or scientifically based. Instead the categories represent a socio-political construct. The Census 2015 National Content Test consisted of a sample of 1.2 million households making it the largest and most thorough testing and validation of detailed racial and ethnic categories. This standard relies heavily on this research as well extensive testing done by the OMB Tabulation Working Group for the 1997 race/ethnicity standard.

Standard or Guideline

Collection Protocol

1. **Self-identification preferred.** Respect for individual dignity should guide the processes and methods for collecting data on race and ethnicity. Use self-identification when feasible and practical. If self-identification is not feasible or practical at the point of collection, departments should provide a later opportunity for individuals to self-identify.
    i. **Exception.** When collecting data for purposes of understanding bias in perceptions, use perceived race and ethnicity. For example, data collection on stops must use perceived race and ethnicity.
2. **Multiple selections must be allowed.** Respondents or data collectors must be allowed to select more than one response.
3. **Refusal to answer.** If the respondent does not answer the race/ethnicity question, the interviewer may repeat the question and response options. If the respondent fails to respond to the question, the interviewer may infer a response (based upon observation or information provided by another source).
4. **Training.** If staff will be collecting data verbally per this standard, Departments should

develop and implement standard training.

Question Format

Below are formats you should use when collecting race and ethnicity data. The formats below address:

- **Ability to collect multiple values.** Not all systems are able to collect multiple selections for a single field value. Use the formats as follows:
  - Format A. Use this format if your system allows for the selection of multiple values. Most modern systems should be able to accommodate this.
  - Format B. Use this format if your system is unable to select multiple values.
- **Option to collect detailed data.** Under each format option or via subsequent questions, you can collect additional details on subgroups. Each detailed option must roll up into a one of the 7 standard groups (1). See Appendix C for suggested options.

> (1) Refer to 2015 National Content Test Race and Ethnicity Analysis Report. February 28, 2017. Matthews, Kelly et all. Pages 200-282 for roll up guidance

Format A. Multi-Select

Field name (1)	**Race and ethnicity**
Question prompt (2)	• Paper data collection: Mark all that apply • Electronic data collection: Select all that apply
Options and order (3)	• White • Asian • Hispanic, Latino, or Spanish • Black or African American • Middle Eastern or Northern African • Native Hawaiian or Other Pacific Islander • American Indian or Alaska Native
Format	Multi-select checkbox. See Appendix B for examples.

> (1-2) This terminology was tested in the Census 2015 National Content Test.
>
> (3) Order based on population of San Francisco MSA.

Format B: Single Select

If you cannot use a multi-select option, this format consists of the same field collected at least twice as follows.

Field name	Race and ethnicity 1
Question prompt	• Paper data collection: Mark which one that applies • Electronic data collection: Select which one that applies
Options and order	• White • Asian • Hispanic, Latino, or Spanish • Black or African American • Middle Eastern or Northern African • Native Hawaiian or Other Pacific Islander • American Indian or Alaska Native
Format	Radio button. See Appendix B for examples.

Field name	Race and ethnicity 2
Question prompt	If applicable, mark an additional race/ethnicity • Paper data collection: Mark which one that applies • Electronic data collection: Select which one that applies
Options and order	• White • Asian • Hispanic, Latino, or Spanish • Black or African American • Middle Eastern or Northern African • Native Hawaiian or Other Pacific Islander • American Indian or Alaska Native
Format	Radio button. See Appendix B for examples.

Reporting

At a minimum, you should calculate the following estimates when reporting on race and ethnicity data.

- **Each race and ethnicity alone.** This table will provide a Census compatible table that sums to 100%. To create this table, report the following groups:
    - White alone
    - Asian alone
    - Hispanic, Latino, or Spanish alone
    - Black or African American alone
    - American Indian or Alaska Native alone
    - Middle Eastern or Northern African alone
    - Native Hawaiian or Other Pacific Islander alone
    - Two or more races

- **Each race and ethnicity plus some other race.** This table will sum to more than 100%. To create this table, report the following groups:
    - White plus any other race and ethnicity
    - Asian plus any other race and ethnicity
    - Hispanic, Latino, or Spanish plus any other race and ethnicity
    - Black or African American plus any other race and ethnicity
    - American Indian or Alaska Native plus any other race and ethnicity
    - Middle Eastern or Northern African plus any other race and ethnicity
    - Native Hawaiian or Other Pacific Islander plus any other race and ethnicity

Mapping and Transformations

You may need to map your race and ethnicity data for the purposes of matching how this data is reported by other jurisdictions, surveys or even historical data your department may have collected. When doing mapping and transformations, you will have to address three core issues:

1. Mapping to a standard that does not allow for multi-select
2. Mapping to a standard that used two separate questions for race and ethnicity
3. Mapping to a standard that uses different groups or categories

The rules below break out by case depending on the destination system or standard. The mapping tables provide detailed specifications on how to meet these. Appendix F provides more background on these rules. Appendix A provides details on how to do this mapping.

Case 1. Mapping to a combined question format with multi-select options

In Case 1, the only issue that would come up would be different categories. The most common differences should be mapped as follows. If you come across additional ones, feel free to reach out to us for guidance.

1. **Middle Eastern or North African missing.** Map to White as per Census designation. (1)
2. **Native Hawaiian or Other Pacific Islander missing.** Map to Asian. (2)
3. **Any other missing categories missing.** Use ‘Other’ or ‘Some Other Race’ or ‘Unknown’ when available.

> (1) 2015 National Content Test Race and Ethnicity Analysis Report. February 28, 2017. Matthews, Kelly et al. Pages 200-282.
>
> (2) Tabulation Working Group. December 15, 2000. Provisional Guidance on the Implementation of the 1997 Standards for Federal Data on Race and Ethnicity Ch. 5 Section B.1 p 88.

Case 2. Mapping to a combined question format with single-select option

Our standard allows for multi-selection. If you have to report to an external system that only allows one value, use the following rules for records with multiple selections. Appendix A provides details on how to do this mapping:

1. **Missing categories.** Refer to Case 1 rules if your categories do not match.
2. **More than 1 selected, "Hispanic, Latino, or Spanish" selected.** If one of the values is Hispanic, report the respondent as Hispanic regardless of what other selections are made. For example, if someone selects Hispanic and Asian, you would map them to Hispanic.
    i. If the destination standard does not have Hispanic, Latino, or Spanish as an option use the other response to report it.
3. **More than 1 selected, "Hispanic, Latino, or Spanish" NOT selected.** Apply "Largest Group other than White" rule. Map the respondent to the largest of the group as represented in the San Francisco Bay Area general population unless that race is White. For example, if someone selects White and Asian, report them as Asian.The order from largest to smallest is determined using population estimates for race and ethnic groups (when available) for the San Francisco Metropolitan Statistical Area (see Appendix D):
    i. White
    ii. Asian
    iii. Hispanic, Latino, or Spanish
    iv. Black or African American
    v. Middle Eastern or North African
    vi. Native Hawaiian or Other Pacific Islander
    vii. American Indian or Alaska Native
4. **Exceptions to 2 and 3.** If an option for multi-race exists, map multi-selections to that option.

Case 3. Mapping to a separate question format with multi-select option

Some external standards will separate race and ethnicity into two separate fields, with ethnicity designated for Hispanic, Latino, or Spanish, and still allow for multiple selections under the race field. Use the following rules in this case.

1. **Missing categories.** Refer to Case 1 rules if your categories do not match.
2. **“Hispanic, Latino, or Spanish” selected.** Record ethnicity as Hispanic, Latino, or Spanish or equivalent and:
    i. If other race/ethnicities selected, record under race
    ii. If no other selected, record as Unknown or Other
3. **More than 1 selected, “Hispanic, Latino, or Spanish” NOT selected.** Record each selection in the destination standard using the Case 1 rules as needed.

Case 4. Mapping to a separate question format with single-select option

Like Case 3, race and ethnicity are two separate fields, with ethnicity designated for Hispanic, Latino, or Spanish. However, you may only select one option under the race field. Use the following rules in this case. Appendix A provides details on how to do this mapping.

1. **Missing categories.** Refer to Case 1 rules if your categories do not match.
2. **“Hispanic, Latino, or Spanish” selected.** Record ethnicity as Hispanic, Latino, or Spanish or equivalent and:
    i. If another race/ethnicity selected, record that under race. If more than 1 additional race/ethnicity selected, use rule 3 below.
    ii. If no other selected, record as Unknown or Other
3. **More than 1 selected, Hispanic, Latino, or Spanish NOT selected.** Apply “Largest Group other than White” rule. Map the respondent to the largest of the group as represented in the San Francisco Bay Area general population unless that race is White. For example, if someone selects White and Asian, report them as Asian.The order from largest to smallest is determined using population estimates for the race alone values (when available) for the San Francisco Metropolitan Statistical Area (see Appendix D):
    i. White
    ii. Asian
    iii. Hispanic, Latino, or Spanish
    iv. Black or African American
    v. Middle Eastern or North African
    vi. Native Hawaiian or Other Pacific Islander
    vii. American Indian or Alaska Native
4. **Exception to 3.** If an option for multi-race exists, map multi-selections to that option.

Definitions

Race and ethnicity data collections should include the following minimum categories and definitions.(1)

> (1) Definitions from Census 2015 National Content Test.

Category	Definition
American Indian or Alaska Native	The category "American Indian or Alaska Native" includes all individuals who identify with any of the original peoples of North and South America (including Central America) and who maintain tribal affiliation or community attachment. It includes people who identify as "American Indian" or "Alaska Native" and includes groups such as Navajo Nation, Blackfeet Tribe, Mayan, Aztec, Native Village of Barrow Inupiat Traditional Government, Nome Eskimo Community, etc.
Asian	The category "Asian" includes all individuals who identify with one or more nationalities or ethnic groups originating in the Far East, Southeast Asia, or the Indian subcontinent. Examples of these groups include, but are not limited to, Chinese, Filipino, Asian Indian, Vietnamese, Korean, and Japanese. The category also includes groups such as Pakistani, Cambodian, Hmong, Thai, Bengali, Mien, etc.
Black or African American	The category "Black or African American" includes all individuals who identify with one or more nationalities or ethnic groups originating in any of the black racial groups of Africa. Examples of these groups include, but are not limited to, African American, Jamaican, Haitian, Nigerian, Ethiopian, and Somali. The category also includes groups such as Ghanaian, South African, Barbadian, Kenyan, Liberian, Bahamian, etc.
Hispanic, Latino, or Spanish	The category "Hispanic, Latino, or Spanish" includes all individuals who identify with one or more nationalities or ethnic groups originating in Mexico, Puerto Rico, Cuba, Central and South American, and other Spanish cultures. Examples of these groups include, but are not limited to, Mexican or Mexican American, Puerto Rican, Cuban, Salvadoran, Dominican, and Colombian. The category also includes groups such as Guatemalan, Honduran, Spaniard, Ecuadorian, Peruvian, Venezuelan, etc.
Middle Eastern or Northern African	The category "Middle Eastern or North African" includes all individuals who identify with one or more nationalities or ethnic groups originating in the Middle East or North Africa. Examples of these groups include, but are not limited to, Lebanese, Iranian, Egyptian, Syrian, Moroccan, and Algerian. The category also includes groups such as Israeli, Iraqi, Tunisian, Chaldean, Assyrian, Kurdish, etc.
Native Hawaiian or Other Pacific Islander	The category "Native Hawaiian or Other Pacific Islander" includes all individuals who identify with one or more nationalities or ethnic groups originating in Hawaii, Guam, Samoa, or other Pacific Islands. Examples of these groups include, but are not limited to, Native Hawaiian, Samoan, Chamorro, Tongan, Fijian, and Marshallese. The category also includes groups such as Palauan, Tahitian, Chuukese, Pohnpeian, Saipanese, Yapese, etc.
White	The category "White" includes all individuals who identify with one or more nationalities or ethnic groups originating in Europe. Examples of these groups include, but are not limited to, German, Irish, English, Italian, Polish, and French. The category also includes groups such as Scottish, Norwegian, Dutch, Slavic, Cajun, Roma, etc.

Who must comply

Departments should comply with this standard unless they face conflicting requirements. Note that external reporting fields are not requirements. Your data can be transformed to meet external reporting fields if they are different from this standard. Review the section on transformations and mapping.

Authority

San Francisco Administrative Code Chapter 22D: Open Data Policy Section 22D.2(b)(7).

Appendices

Appendix A. Mapping Crosswalk

The mapping to other data standards crosswalk provides crosswalks from the San Francisco Recommended Standard to 4 different reporting options that do not allow the preservation of a respondents multiple race/ethnicity designations:

- Mapping to a combined question format with single-select option (Case 2)
  - Variation A: Without option of ‘Two or More Races’
  - Variation B: With option of ‘Two or More Races’
- Mapping to a separate question format with single-select option (Case 4)
  - Variation A: Without option of ‘Two or More Races’
  - Variation B: With option of ‘Two or More Races’

Appendix B. Example Question Formats

Please view this google form for example question formats for implementing Format's A and B.

Appendix C. Detailed Categories

Under this standard, departments have the discretion to collect additional detail on subgroups within each category as long as the values roll up into one of the seven values in this standard.

Below we provide the main categories with detailed subgroup options from two sources:

- The Census 2015 National Content Test
- An analysis of San Francisco MSA race and ethnicity estimates

Departments should only collect detailed subgroup data to the degree it is useful for delivering, providing or evaluating programs and services. For example, a department may want additional subgroup detail for one category but not for others. Many departments may find that the seven main categories are sufficient. Appendix B includes example question formats when collecting detailed data.

For additional roll up guidance: Refer to 2015 National Content Test Race and Ethnicity Analysis Report. February 28, 2017. Matthews, Kelly et all. Pages 200-282.

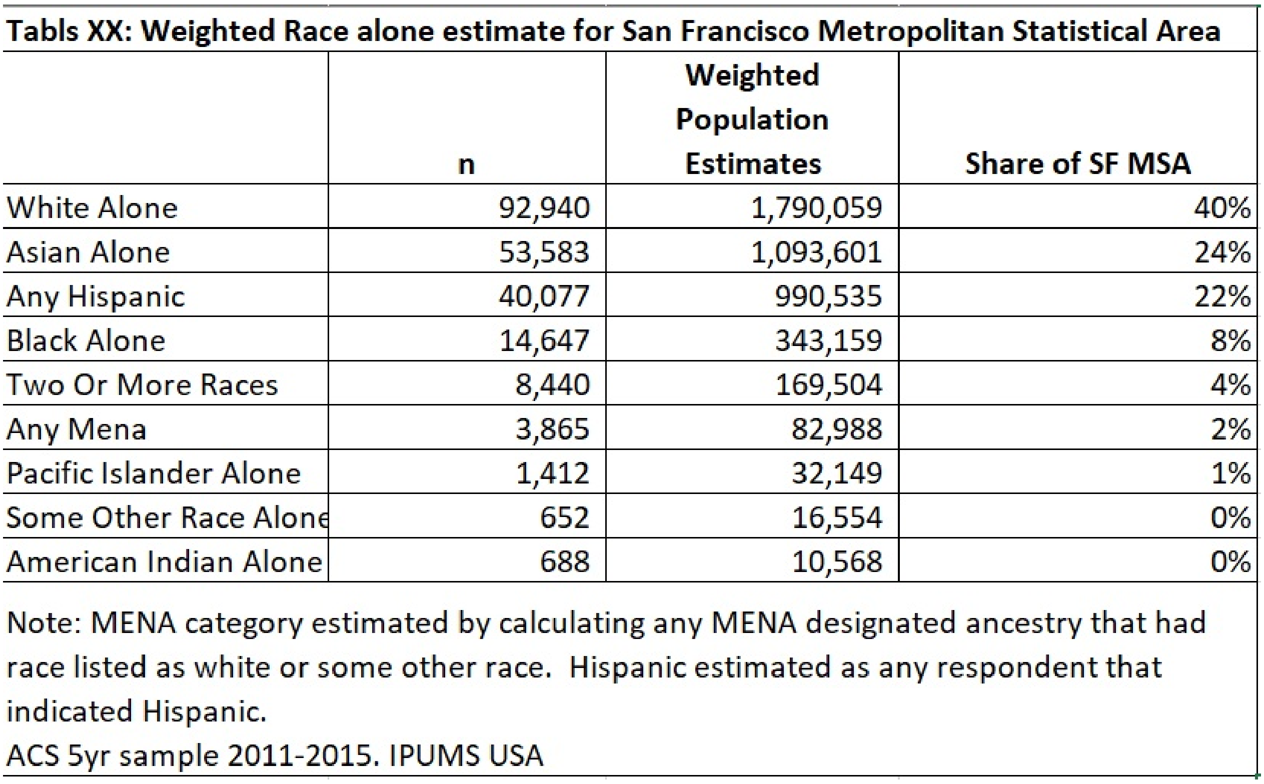
Main Category	**Census 2020 Categories Based on US Population**	**Detailed Categories Based on SF MSA distribution (1)**
White	• German • Irish • English • Italian • Polish • French • Write in	• Irish • German • English • Italian • Russian • French • Scottish • Portuguese • Polish • Swedish • Norwegian • Write in
Hispanic, Latino, or Spanish	• Mexican or Mexican American • Puerto Rican • Cuban • Salvadoran • Dominican • Columbian • Write in	• Mexican or Mexican American • Salvadoran • Guatemalan • Nicaraguan • Puerto Rican • Spaniard • Peruvian • Honduran • Cuban • Columbian • Write in
Black or African American	• African American • Jamaican • Haitian • Nigerian • Ethiopian • Somali • Write in	• African American • Nigerian • Ethiopian • Jamaican • Eritrean • Haitian • Somali • Write in
Asian	• Chinese • Filipino • Asian Indian • Vietnamese • Korean • Japanese • Write in	• Chinese • Filipino • Asian Indian • Vietnamese • Korean • Japanese • Taiwanese • Thai • Laotian • Cambodian

		Write In
American Indian or Alaska Native	American Indian Alaska Native Central or South American Indian Write in	American Indian Alaska Native Central or South American Indian Write in
Middle Eastern or North African	Lebanese Iranian Egyptian Syrian Moroccan Algerian Write in	Iranian Armenian Arab Lebanese Palestinian Turkish Egyptian Israeli Yemeni Algerian Write in
Native Hawaiian or Other Pacific Islander	Native Hawaiian Samoan Chamorro Tongan Fijian Marshallese Write in	Native Hawaiian Samoan Chamorro Tongan Fijian Marshallese Write in

> (1) Determined by analyzing weighted population counts for either ancestry, tribe (American Indian or Alaska Native), detailed hispanic information (Hispanic) or detailed race information (Asian) information for respondents in SF Metropolitan Statistical Area. Each main race/ethnicity category was analyzed in isolation for all respondents who identified as that category (either alone or in combination with another main race/ethnicity category) using IPUMS provided flags, except for MENA which currently has no flag. MENA was determined by finding the weighted population rank of MENA valid ancestry values. Detailed Categories assigned to Main Categories based Census 2020 proposed mapping (see page 200-282 at 2015 National Content Test Race and Ethnicity Analysis Report. February 28, 2017. Matthews, Kelly et al).

Appendix D. San Francisco MSA Race and Ethnicity Estimates

The table below provides a weighted population estimate by race and ethnicity.



Tabls XX: Weighted Race alone estimate for San Francisco Metropolitan Statistical Area

	n	Weighted Population Estimates	Share of SF MSA
White Alone	92,940	1,790,059	40%
Asian Alone	53,583	1,093,601	24%
Any Hispanic	40,077	990,535	22%
Black Alone	14,647	343,159	8%
Two Or More Races	8,440	169,504	4%
Any Mena	3,865	82,988	2%
Pacific Islander Alone	1,412	32,149	1%
Some Other Race Alone	652	16,554	0%
American Indian Alone	688	10,568	0%

Note: MENA category estimated by calculating any MENA designated ancestry that had race listed as white or some other race. Hispanic estimated as any respondent that indicated Hispanic.
ACS 5yr sample 2011-2015. IPUMS USA

Appendix E. Rationale for Recommended Standard

In the absence of a citywide standard, we relied on the following to inform a citywide recommended standard:

- The results of the Department of Public Health's research that resulted in department wide race and ethnicity guidelines released in 2011
- The large scale, random assignment testing conducted by the US Census in 2010 and 2015 to compare alternative question formats (see overview of research)

Combined or separate questions

A standard for race and ethnicity must address a key design choice: should race and ethnicity be asked as separate (one question for ethnicity, i.e. Hispanic or Latino, and another for race) or combined questions?

Repeated testing by the Census showed that a combined question format yielded data of the highest quality. This is consistent with DPH's recommendation to use a combined question format.

Below is an excerpt from the US Census 2015 National Content Test (NCT):

> “The 2015 NCT research demonstrates that a question format that combines race and ethnicity into one question results in more accurate reporting and dramatically lower item nonresponse compared to the two separate questions on Hispanic origin and on race. In addition, with a new combined question design approach which employed multiple detailed checkboxes to help collect the reporting of detailed groups, the NCT research successfully demonstrated how an innovative approach could collect data for myriad groups across our nation’s diverse population. By combining the race and Hispanic origin questions into 84 one question on race/ethnicity, the research has shown that Hispanics can better find themselves among the race and ethnicity categories.” (Census, 2015)

In addition, responses to combined question format can be mapped to any external reporting requirements that are structured using separate questions.

As a result, the recommended standard for San Francisco combines race and ethnicity into a single question using terminology and language tested in the 2015 National Content Test. To address data mapping concerns, the standard also provides guidance and tools for external reporting and data mapping.

Inclusion of a new category, MENA

The Census tests also explored including a category for Middle Eastern North African (MENA), a group that historically is included in the “white” category. The results concluded that the Census should include MENA:

> “The NCT research findings show that the use of a distinct MENA category elicits higher quality data; and people who identify as MENA use the MENA category when it is available, whereas they have trouble identifying as only MENA when no category is available.” (Census, 2015)

As a result, the recommended standard includes a category for MENA using terminology and language tested in the 2015 National Content Test. The standard also provides guidance and tools for external reporting and data mapping.

Census decision to not make changes for 2020

Despite the results of testing related to the topics above, the Census is not making changes for the 2020 Census. This decision is controversial (this article provides some background on the decision). Despite this decision, we are moving forward with the recommended standard because:

- San Francisco data collection does not operate under the same climate as federal decision making

- The combined question format and inclusion of MENA has generated better response rates and better quality data in repeated testing
- Our one example of a local standard (DPH's race/ethnicity) uses combined as a result of their extensive process of analysis and community engagement
- The existing federal standard already provides for a method for collecting using both combined and separate formats
- External comparisons can be mapped and most reporting already requires mapping the census data to obtain accurate comparisons for Hispanic, Alone

Multiple Selection must be allowed

A study from the Census ranked California as 2nd highest state for those selecting two or more races in the 2010 census. The census has historically captured race/ethnicity information via options that allowed the respondent to select more than one. Likewise the 1997 OMB Race & Ethnicity standard calls for the use of multiple selection.

Any modern data system is capable of capturing multiple selections. For older systems limited to single select, it is possible to capture the equivalent information via two or more instances of a single select question.

Detailed Race/Ethnicity subgroups left to discretion of departments

This standard should not be interpreted as discouraging or limiting the collection of detailed race/ethnicity information. For certain purposes it is desirable to collect more detailed race/ethnicity sub-group information. Different departments and offices will have different sub-groups that are relevant to their work or may be needed for internal or external reporting (ex. detailed asian ethnicities).

Similar to the Department of Public Health's 2011 race and ethnicity guidelines, collection of detailed race/ethnicity information is permitted as long as the values can be rolled up into one the the 7 values in this standard.

Appendix F. Background on Mapping and Transformation Rules

The San Francisco Recommended Standard provides rules for mapping and transforming the standard to external or historical data collection methodologies. Below we provide background on the Largest Group other than White rule.

Most state and federal reporting systems request data ‘as is’ via electronic transfer and will handle aggregation (and the associated decisions) themselves. The reasoning behind this mirrors the reasoning for this standard; it provides the reporting agency with the most detailed data available as well as ensuring a consistent aggregation method across the various state and local jurisdictions.

Challenge: how to map multi-select to a single value. In cases where the department has to perform the aggregation several challenges appear when aggregating from multi-select racial/ethnicity categories to often a single race/ethnicity value. For example if a respondent selected White and Black, or Asian and Hispanic as their races, which option do you report?

Federal working group identified multiple methods. Considerable thought and testing went into such questions during the shift to allowing multiple race selections in the 1997 OMB Race Ethnicity Standard. In 2000 the OMB Tabulation Working Group released guidance on best practices for transforming multi-race data to single race reporting standards. They presented options that ranged in complexity with each containing pros and cons.

Deterministic whole assignment methods should be used. The federal working group identified two main approaches:

- Deterministic Whole Assignment methods which are fixed rules for assigning race/ethnicity values
- Probabilistic and Fractional Assignment methods which rely on statistical estimation

The probabilistic and fractional assignment methods are much more complex to implement and to explain, particularly on a local scale. Given a review of the options and in consultation with experts, we recommend using Deterministic Whole Assignment methods. We identified three options suited to the purposes of the this standard. The three Deterministic Whole Assignment methodologies for when there are 2 or more races selected are:

- Smallest Group. The smallest of the 2 races in the general population is the one reported.
- Largest Group other than White. The largest of the 2 races in the general population is the one reported unless that race is white.
- Largest Group. The largest of the 2 races in the general population is the one reported.

The table below provides examples to illustrate the methods using the makeup of the population in San Francisco. Given the unique demographics of San Francisco, the preferred option is ‘Largest Group other than White’ to ensure adequate representation by non-white groups.

Race and ethnicity 1	Race and ethnicity 2	Smallest group	Largest group other than White	Largest group
White	American Indian or Alaska Native	American Indian or Alaska Native	American Indian or Alaska Native	White
White	Asian	Asian	Asian	White
White	Black or African American	Black or African American	Black or African American	White
White	Hispanic, Latino, or Spanish	Hispanic, Latino, or Spanish	Hispanic, Latino, or Spanish	White
White	Middle Eastern or Northern African	Middle Eastern or Northern African	Middle Eastern or Northern African	White
White	Native Hawaiian or Other Pacific Islander	Native Hawaiian or Other Pacific Islander	Native Hawaiian or Other Pacific Islander	White
Asian	American Indian or Alaska Native	American Indian or Alaska Native	Asian	Asian
Asian	Black or African American	Black or African American	Asian	Asian
Asian	Hispanic, Latino, or Spanish	Hispanic, Latino, or Spanish	Asian	Asian
Asian	Middle Eastern or Northern African	Middle Eastern or Northern African	Asian	Asian
Asian	Native Hawaiian or Other Pacific Islander	Native Hawaiian or Other Pacific Islander	Asian	Asian
Asian	White	Asian	Asian	White

Department of Public Health’s Ethnicity Guidelines

Description of Standard

Below is an excerpt from the DPH guidelines:

> “These guidelines were developed by SFDPH Community Programs epidemiologists, researchers, and analysts who share concerns regarding the collection, coding, reporting, interpretation, and use of social identity indicators. To monitor health outcomes and intervene on behaviors that are the underlying causes of disease and injuries, SFDPH must be able to incorporate changing definitions, relevance, and boundaries that individuals, communities, programs and/or institutions use to identify themselves and others.
>
> These guidelines address the following key issues concerning race and ethnicity:
>
> 1. Desire for consistency in grouping or categorizing of race and ethnicity data across time and data regimes.
> 2. Need for flexibility to accommodate many different existing data collection practices.
> 3. Lack of clarity in the meaning and use of terms defining race and ethnicity.”

Standard or Guideline

The full guidelines include details on how to collect and report the data. Below are excerpts:

> A single set of common mutually-exclusive core ethnicity categories that are aligned with state and federal minimum reporting categories should be used.
>
> Persons who select more than one ethnicity should be given the opportunity to also select their primary ethnicity.
>
> Ethnicity data should be minimally reported by these core categories and definitions.

Definitions

- African American/ Black. A person having origins in any of the black ethnic groups of

Africa
- Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia (including Philippines), or the Indian subcontinent
- Native Hawaiian or Other Pacific Islander (NHOPI). A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands
- Native American. A person having origins in any of the original peoples of North America, Central America, or South America
- Latino/a. A person having origins in Mexico, Central America, South America, Puerto Rico, or Cuba
- White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa
- Multi-ethnic. A person having origins in more than one of the other core categories specified.

> “Other” should not be an option under the Core categories, for all ethnicities fall under one of the above seven options.

Who must comply

> “All new data collection systems purchased or designed for or by the Department of Public Health that will be used to track the ethnicity of patients, clients, participants, or other cohorts must have the ability to track ethnicity in accordance with these guidelines. Additionally, reporting of collected data should also adhere to these guidelines whenever possible, recognizing third party reporting requirements may be in conflict.”

Authority

San Francisco Department of Public Health, Central Administration, “Principles for Collecting, Coding, and Reporting Social Identity Data – Ethnicity Guidelines (COM3)”.

State of California

We know of only one California standard that faces a City Department. If you know of others, regardless of whether or not they apply to City Departments, please contact someone at DataSF.

Racial and Identity Profiling Act of 2015 Regulations

Description of Standard

Under the California Racial and Identity Profiling Act of 2015 (AB 953), state and local law enforcement agencies must collect data regarding stops of individuals, including perceived demographic information on the person stopped. They must report this data to the California Attorney General's Office.

As part of this law, the California Attorney General’s Office issued regulations that detail how stops data must be collected. This data standard includes data on race and ethnicity. Below is the race and ethnicity excerpt from the state regulations. The full standard is available online.

> **Caution:** This data standard only applies to the Police Department in the context of stops data. As a result, the data standard requires perception of race and ethnicity and the standard reflects this in the categories used. This is due to the purpose of the data collection, including to identify potential bias. In contrast, other standards rely on self-identification, which typically leads to different categories

Standard or Guideline

Below is an excerpt from the standard.

> "Perceived Race or Ethnicity of Person Stopped" refers to the officer's perception of the race or ethnicity of the person stopped. When reporting this data element, the officer shall make his or her determination of the person's race or ethnicity based on personal observation only. The officer shall not ask the person stopped his or her race or ethnicity, or ask questions
> or make comments or statements designed to elicit this information.
>
> When reporting this data element, the officer shall select all of the following data values that apply:
>
> 1. Asian
> 2. Black/African American
> 3. Hispanic/Latino(a)
> 4. Middle Eastern or South Asian
> 5. Native American
> 6. Pacific Islander
> 7. White
>
> Example: If a person appears to be both Black and Latino(a), the officer shall select both "Black/African American" and "Hispanic/Latino(a)."

Definitions

- "Asian" refers to a person having origins in any of the original peoples of the Far East or Southeast Asia, including for example, Cambodia, China, Japan, Korea, Malaysia, the Philippine Islands, Thailand, and Vietnam, but who does not fall within the definition of "Middle Eastern or South Asian" or "Pacific Islander."
- "Black/African American" refers to a person having origins in any of the Black racial groups of Africa.
- "Hispanic/Latino(a)" refers to a person of Mexican, Puerto Rican, Cuban, Central or South American, or other Spanish culture or origin, regardless of race.
- "Middle Eastern or South Asian" refers to a person of Arabic, Israeli, Iranian, Indian, Pakistani, Bangladeshi, Sri Lankan, Nepali, Bhutanese, Maldivian, or Afghan origin.
- "Native American" refers to a person having origins in any of the original peoples of North, Central, and South America.
- "Pacific Islander" refers to a person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands, but who does not fall within the definition of "Middle Eastern or South Asian" or "Asian."
- "White" refers to a person of Caucasian descent having origins in any of the original peoples of Europe and Eastern Europe.

Who must comply

The Police Department in the context of collecting data on stops. This standard is not required for other types of data collection and may not be appropriate as it was designed to capture perceived race/ethnicity.

Authority

State of California Government Code Title 2 Section 12525.5.

Federal Government

Standards for the Classification of Federal Data on Race and Ethnicity (rev. 1997)

Description of Standard

The Office of Management and Budget sets standards for the collection of race and ethnicity data used for federal government purposes.

The current OMB definition is from 1997 per the Standards for the Classification of Federal Data on Race and Ethnicity (rev. 1997). The standards represent minimum requirements; agencies can, and do, go beyond these minimum standards but they must be able to aggregate data to the OMB's defined categories.

These standards

> "were developed in cooperation with Federal agencies to provide consistent data on race and ethnicity throughout the Federal Government. Development of the data standards stemmed in large measure from new responsibilities to enforce civil rights laws. Data were needed to monitor equal access in housing, education, employment, and other areas, for populations that historically had experienced discrimination and differential treatment because of their race or ethnicity. The standards are used not only in the decennial census (which provides the data for the "denominator" for many measures), but also in household surveys, on administrative forms (e.g., school registration and mortgage lending applications), and in medical and other research. The categories represent a social-political construct designed for collecting data on the race and ethnicity of broad population groups in this country, and are not anthropologically or scientifically based."

OMB initiated a process in 2016-17 to revisit the standard due to limitations with the existing standard. While a notice came out in March of 2017, we do not know of any additional steps.

Standard or Guideline

Below is an excerpt containing the bulk of the standard. Read the full standard online.

> This classification provides a minimum standard for maintaining, collecting, and presenting data on race and ethnicity for all Federal reporting purposes. The categories in this classification are social-political constructs and should not be interpreted as being scientific or anthropological in nature. They are not to be used as determinants of eligibility for participation in any Federal program. The standards have been developed to provide a common language for uniformity and comparability in the collection and use of data on race and ethnicity by Federal agencies.
>
> The standards provide two formats that may be used for data on race and ethnicity. Self-reporting or self-identification using two separate questions is the preferred method for collecting data on race and ethnicity. In situations where self-reporting is not practicable or feasible, the combined format may be used.
>
> Respondents shall be offered the option of selecting one or more racial designations. Recommended forms for the instruction accompanying the multiple response question are "Mark one or more" and "Select one or more."
>
> In no case shall the provisions of the standards be construed to limit the collection of data to the categories described above. The collection of greater detail is encouraged; however, any collection that uses more detail shall be organized in such a way that the additional categories can be aggregated into these minimum categories for data on race and ethnicity.

Definitions

The minimum categories for data on race and ethnicity for Federal statistics, program administrative reporting, and civil rights compliance reporting are defined as follows:

- American Indian or Alaska Native. A person having origins in any of the original peoples of North and South America (including Central America), and who maintains tribal affiliation or community attachment.
- Asian. A person having origins in any of the original peoples of the Far East, Southeast Asia, or the Indian subcontinent including, for example, Cambodia, China, India, Japan, Korea, Malaysia, Pakistan, the Philippine Islands, Thailand, and Vietnam.
- Black or African American. A person having origins in any of the black racial groups of Africa. Terms such as "Haitian" or "Negro" can be used in addition to "Black or African American."
- Hispanic or Latino. A person of Cuban, Mexican, Puerto Rican, Cuban, South or Central American, or other Spanish culture or origin, regardless of race. The term, "Spanish origin," can be used in addition to "Hispanic or Latino."
- Native Hawaiian or Other Pacific Islander. A person having origins in any of the original peoples of Hawaii, Guam, Samoa, or other Pacific Islands.

- White. A person having origins in any of the original peoples of Europe, the Middle East, or North Africa.

Who must Comply

This standard does not face City Departments. Below is the compliance requirement at the federal level:

> “The new standards will be used by the Bureau of the Census in the 2000 decennial census. Other Federal programs should adopt the standards as soon as possible, but not later than January 1, 2003, for use in household surveys, administrative forms and records, and other data collections.”

Authority

Executive Office of the President, Office of Management and Budget (OMB), Office of Information and Regulatory Affairs.

Reference: Basemap

A basemap is most often associated with a visual representation of base geography (streets, buildings, parks, etc.) upon which other elements may be mapped. The base layers on that map help the user orient themselves within space.

In this section, we lay out some component basemap pieces that form core reference data. The underlying data can be used in more than just developing a visual reference map.

- We start with an overview of how the pieces fit together. Understanding this can help you when linking and referencing data across multiple department datasets.
- Then for each basemap component we provide:
    - A definition
    - Visual illustration of the concept
    - Authority under which it is collected
    - Primary or authoritative uses
    - Accepted values
    - And summary of supporting reference data

Basemap Overview

A location reference (addressing) data model

Each component is described in this section individually, but there are important relationships among them.

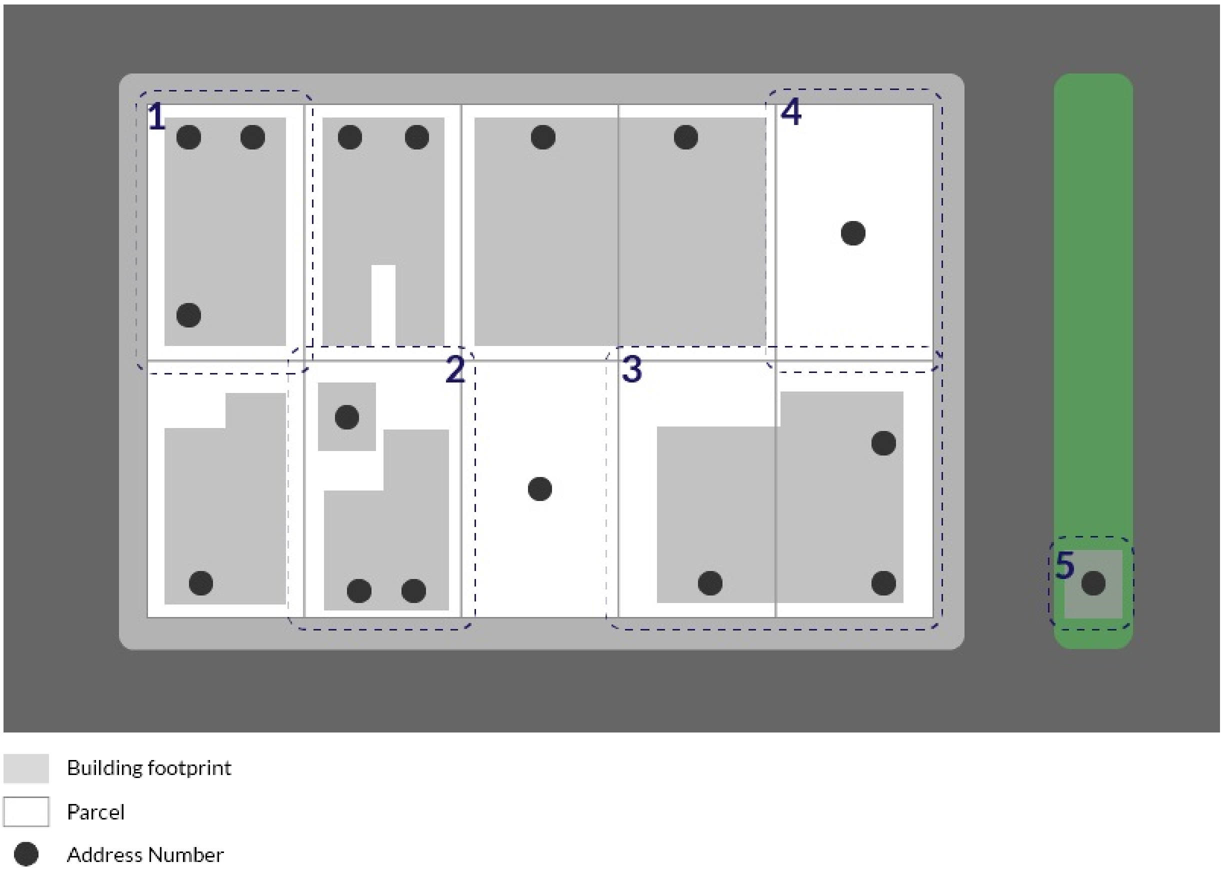
Let's start with three core components:

- **Parcels**. The most common unit of reference for City data, the parcel defines the physical extent of land ownership. It is the outcome of a regulated land subdivision process.
- **Address Numbers**. As an outcome of permitting, new address numbers are assigned to each entry from the street per rules specified in the Building Codes.
- **Building Footprints**. Building footprints represent a physical structure in 2D extents. These are not formally digitized and added to a reference during the development process.

> **Note:** at the time, because buildings are not updated as development occurs, there will be missing data.

Illustration

The following illustrates the relationship among the components.



1. **1 Parcel, 1 Building.** This is often the case in areas like the Financial District or other densely populated office districts.
2. **1 Parcel, Many Buildings.** This occurs in neighborhoods with accessory dwelling units or detached buildings.
3. **Many Parcels, 1 Building.** This occurs when a building is subdivided into different ownership.
4. **1 Parcel, No Buildings.** While rare, this does happen in the case of parking lots, vacant lots and some parks.
5. **No parcels, 1 Building.** Buildings can be built in the right of way (e.g. on a median) where parcels don't exist. This happens rarely.

In all cases, a single address can never be associated with multiple buildings or multiple parcels.

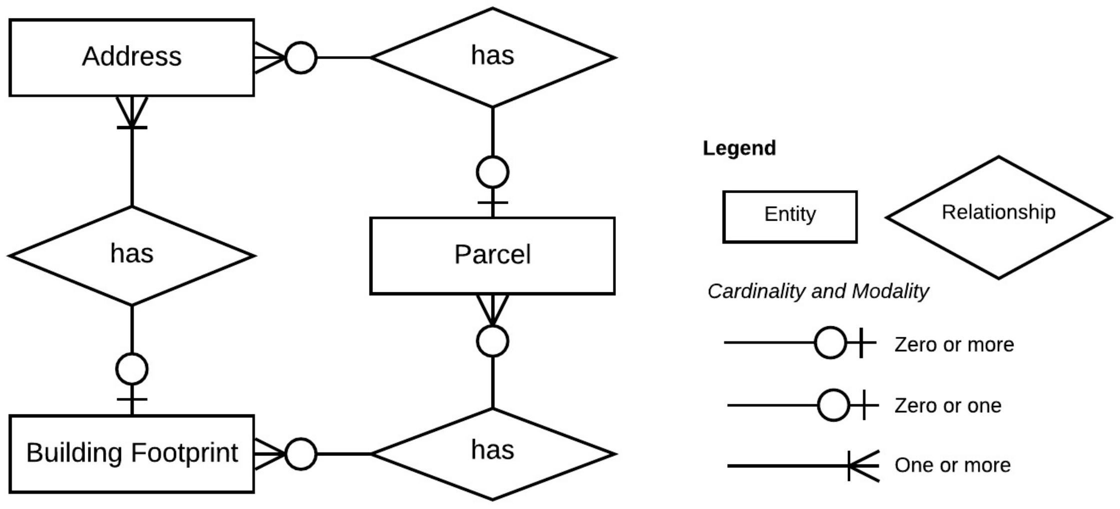
Relationship table and conceptual diagram

The table and diagram below explain the relationship among the 3 core components above.

From	To	Relationship	Notes
Address Number	Parcel	An address number is related to 0 or 1 parcel	An address number may only occasionally fall in the right of way where there is no parcel
Address Number	Building	An address number is related to 0 or 1 building	In some cases an address will be assigned to a lot with no physical structure
Parcel	Address Number	A parcel has 0 or many address numbers	When a parcel is first created through subdivision, it may have no addresses associated with it yet
Parcel	Building	A parcel has 0 or many buildings	A parcel doesn't have to have a building on it
Building	Address Number	A building has 1 or many address numbers	Per Building Code, once a building is approved, it will have at least 1 entrance address if not more*
Building	Parcel	A building is in 0 or many parcels	Buildings may actually exist in the roadway (e.g. a public works toolshed) and not sit on a parcel at all. Most buildings sit within 1 or many parcels.

> ***Note:** The relationship between buildings and address numbers is conceptual at the time. Staff create address points in the Enterprise Addressing System (EAS) within the parcel but not in reference to the building. In cases where there is one building on one parcel, the address point may fall within the building footprint, but there's not an explicitly modeled relationship across all buildings.

CONCEPTUAL RELATIONSHIP DIAGRAM



Relationship to streets

Each of these components relates to one or more streets. A street centerline has a unique identifier called a Centerline Node Network ID.

- A building or parcel will relate to at least one street segment
  - Parcels or buildings with streets on either side will have 2 related segments
  - Corner buildings and parcels will also relate to two segments
  - Depending on the size and shape of the parcel or building, they could be fronted on several or all sides by street segments
  - Most parcels or buildings on the interior of a block will relate to a single street segment
- An address number can only be associated with a single street segment
  - When the City assigns an address number they do so along a street segment within an allowed range

A note on historical data

Streets, parcels, buildings, addresses all change over time. Historical City data could reference things that don't currently exist. In each of the next pages, references include both current and historical where available. You can also browse the reference index.

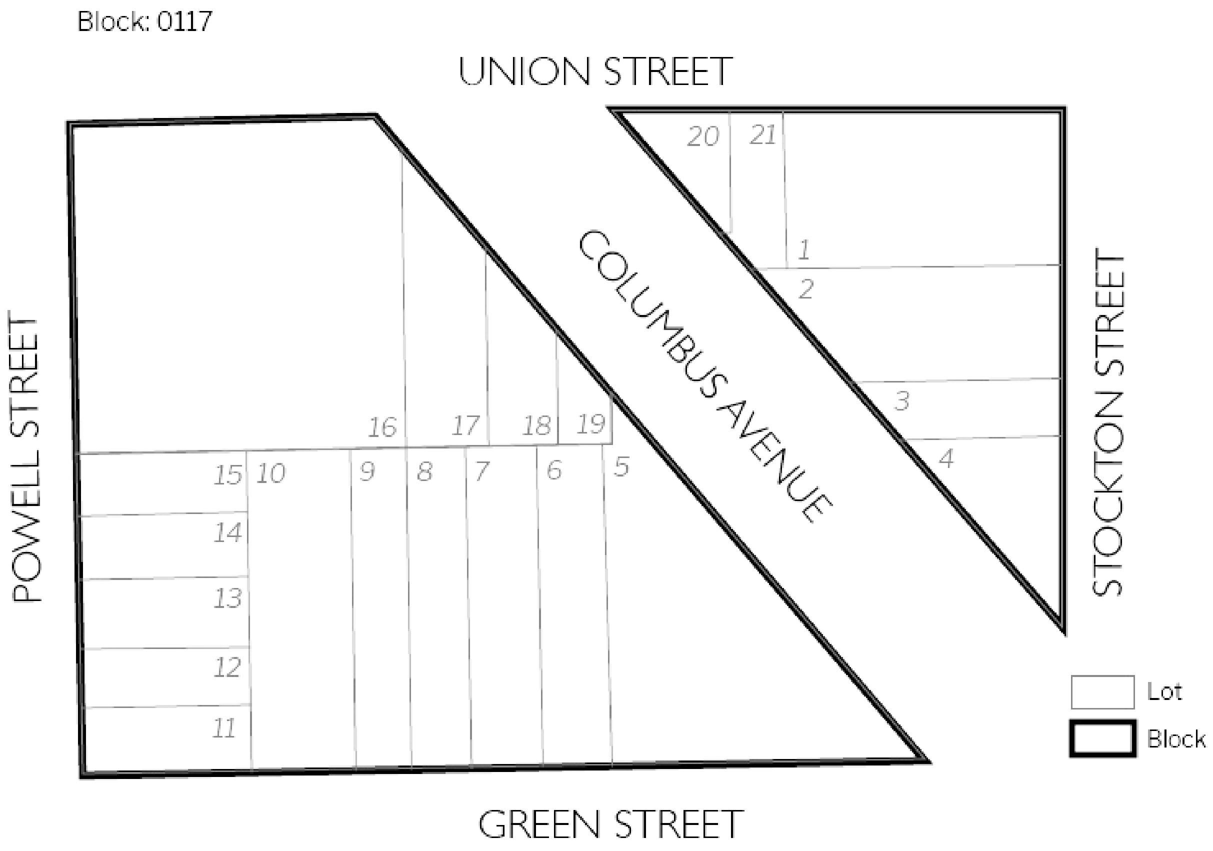
Parcels

Definition

- A parcel is a piece of land or a lot (real property) identified by a unique Assessor Parcel Number (APN)
- The APN is comprised of a **block number** and a **lot number**
    - Block number format: 4 numerical digits + 1 optional letter character (0012A)
    - Lot number format: 3 numerical digits + 1 optional letter character (037B)
    - Blocks are groupings of lots which are usually contiguous and usually bounded by streets or other features on all sides
        - Blocks can be discontiguous and split by other blocks or streets
    - The City is broken up into over 6,000 blocks and over 200,000 individual lots

> **Note:** You will see reference to `mapblklot` in some City data. This is to reference a 1:M relationship of vertical parcels to a base parcel; e.g. condo or timeshare lots.
>
> The practice of representing a vertical lot digitally is to duplicate and "stack" the base parcel for each vertical lot in the building, assigning each a unique `blklot` number. The `mapblklot` is the reference to the base APN. So `blklot` will be unique, while `mapblklot` will duplicate across vertical lots.

Illustration



- Block 0117 above is bounded:
    - On the North and South by Union and Green Streets
    - On the East and West by Stockton and Powell Streets
- Columbus Avenue bisects it, but both sides are still part of the same block
- The block is subdivided into lots numbered from 001 through 021
- A full Assessor Parcel Number would be the concatenation of the block and lot
    - Blocks are 4 digits with an optional letter suffix - 117 becomes 0117
    - Lots are 3 digits with an optional letter suffix - 4 becomes 004
    - The full APN for lot 4 in block 117 is 0117004
- These are recorded in paper maps in the Office of the Assessor Recorder and digitized

Authority

- Recordation of final parcel maps happens with the Office of the Assessor-Recorder
- Before recordation, subdivision maps are approved by the County Surveyor, the Public Works Director and the Board of Supervisors
    - More information about the subdivision process and related codes on the Public Works website

Use

- Assessor Parcel Numbers are used to tie deeds and legal records to property
- Assessor Parcel Numbers used to assess and collect taxes on land and improvements
- As a common administrative identifier for a number of processes like permitting

Accepted values

- Must be provided in a dataset as 2 separate fields:
    - Block as `blk` or `block` or `block_num` - must have 4 numeric digits and an optional letter suffix
    - Lot as `lot` or `lot_num` - must have 3 numeric digits and an optional letter suffix
- When representing the fully qualified APN as a single field:
    - Name the column either `apn` or `assessor_parcel_number` or `blklot` or `block_and_lot`
    - Concatenate the block and lot values together
    - Do not separate the block and lot number with space or other characters
        - 0585012D instead of 0585/012D
    - Do not prepend with additional text like `APN` or `Block and Lot Number`
- Current parcels and corresponding identifiers in the current subdivision parcels below
- Historic parcels and corresponding identifiers in the recorded parcel geography below (note limitations)

Reference Datasets

Dataset	Description and Constraints	Block Column	Lot Column	APN Column
Current Subdivision Parcels	These are the current active recorded parcels. The geography can be used as reference but should not be used for anything requiring precision.	`block_num`	`lot_num`	`blklot`
Recorded Parcel Geography with Transaction Date History	These are the current and historic parcels with recorded dates. Historic parcels only go back to about 1995 with some exceptions. Useful for tying historic administrative records to a location. The geography can be used as reference but should not be used for anything requiring precision.	`block_num`	`lot_num`	`blklot`
San Francisco Assessor Blocks	Just the blocks without lots	`block_num`	N/A	N/A

Is anything wrong, unclear, missing?

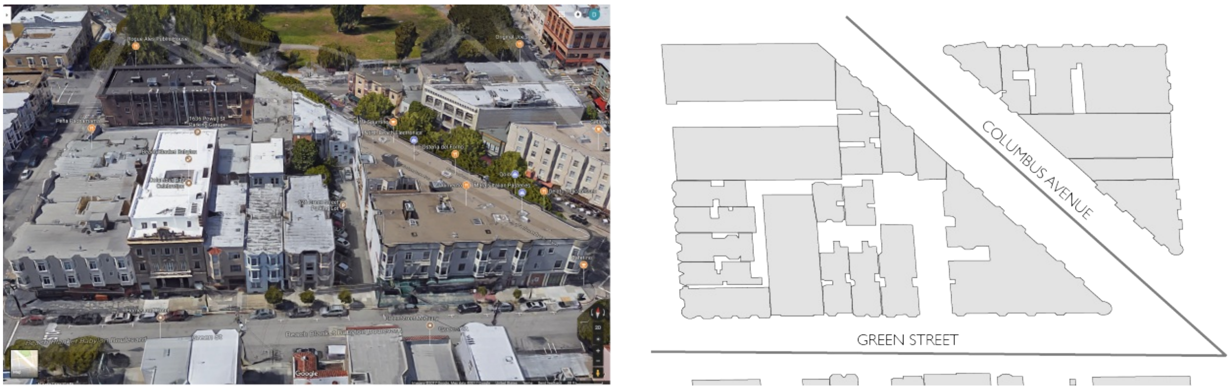
Leave a comment.

Building Footprints

Definition

- The extent of a building in 2 dimensional space
- Includes a unique identifier and other information derived from LIDAR (e.g. max height)

Illustration



- On left: Oblique view of Green St facing north between Columbus and Powell (Imagery: © 2017 Google; Left Panel Map Data: © 2017 Google)
- On right: building footprints for the same block

Authority

- SFGIS in the Department of Technology manages data collection and processing from LIDAR
  - LIDAR data is provided by a third-party and is updated every ???
  - From this data, SFGIS derives the footprints and assigns unique identifiers as well as additional derived statistics about the building (e.g. min, max and median height)
- Information about buildings is captured by other departments including Building Inspection, SF Environment, SF Planning and the City's Real Estate Division among others.
  - Building footprints do not include administrative data about a building
  - They can be related to administrative data spatially and via unique identifiers

Use

- To relate other administrative records to a structure
- To clarify among administrative datasets what specific structure is being referenced
- To improve the addressing model so that address numbers reference a building, not just a parcel

Accepted values

- Footprints are not currently updated as new buildings are constructed
- For those buildings constructed before 2010, you can use the unique identifier `sf16_bldgid`

Reference Datasets

Dataset	Description and Constraints	Reference Columns
Building Footprints	The footprint extents are collapsed from an earlier 3D building model provided by Pictometry of 2010, and have been refined from a version of building masses publicly available on the open data portal for over two years. The building masses were manually split with reference to parcel lines, but using vertices from the building mass wherever possible. These split footprints correspond closely to individual structures even where there are common walls; the goal of the splitting process was to divide the building mass wherever there was likely to be a firewall. An arbitrary identifier was assigned based on a descending sort of building area for 177,023 footprints. The centroid of each footprint was used to join a property identifier from a draft of the San Francisco Enterprise GIS Program's cartographic base, which provides continuous coverage with distinct right-of-way areas as well as selected nearby parcels from adjacent counties.	`sf16_bldgid` unique identifier for footprint `mblr` for reference to property identifiers including parcels and right of way

Is anything wrong, unclear, missing?

Leave a comment.

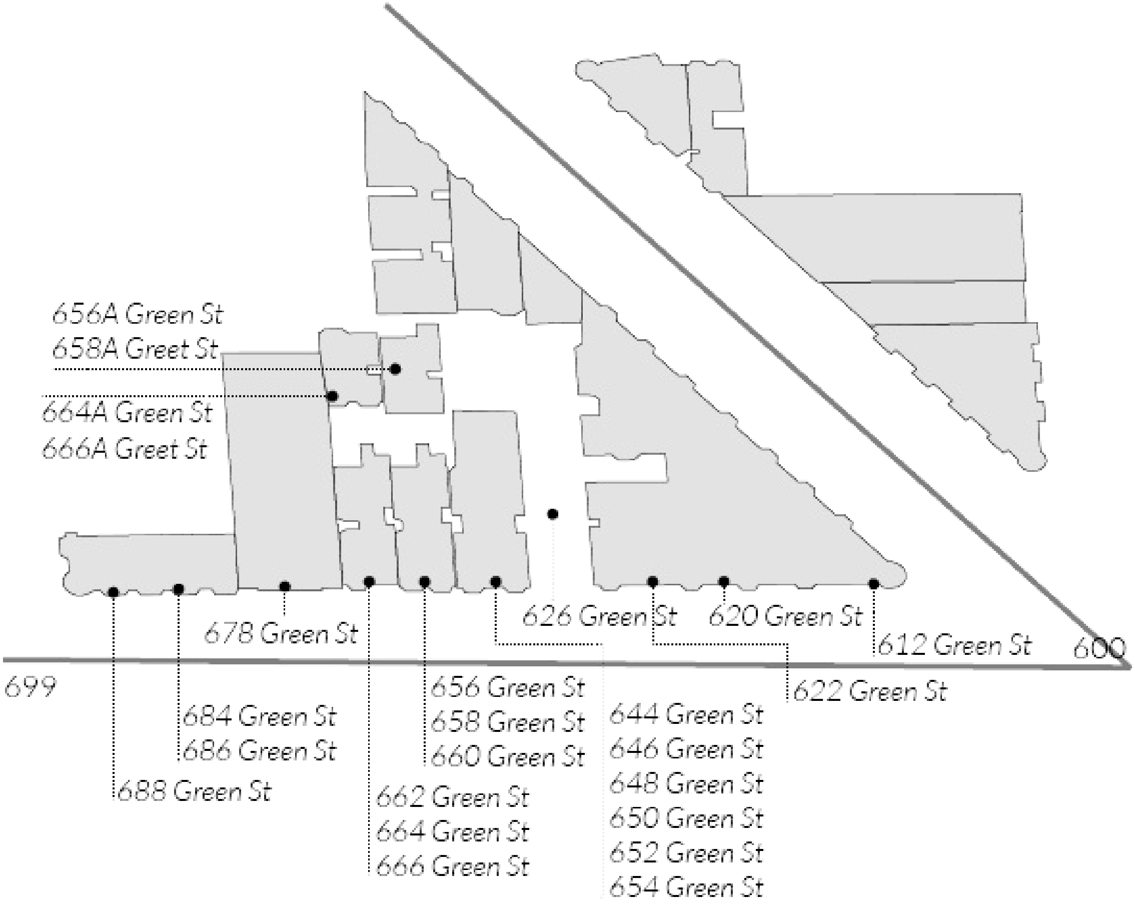
Address Numbers

Definition

Per Administrative Bulletin 035 (AB-035) in the San Francisco Building Codes:

> All primary entrances from the street to all buildings and all direct entrances from the street to separate tenant spaces or dwelling units shall be numbered

Illustration



- Illustration of right side of Green Street between Columbus Ave and Powell St
- 100 valid address numbers on this segment from 600 to 699
  - Even adddresses on right, odds on left
- Each address corresponds to an entrance from the street
  - Note buildings at the rear of the building facing the street have entryways from the street (e.g. 656A, 658A, 664A, and 666A)
  - Numbers can be assigned where there is no building, but they must be associated

with a parcel
- e.g. the parking lot at 626 Green St

Authority

- The official street numbers are assigned by the Department of Building Inspection Building Official prior to permits for new structures according to the procedure in AB-035

Use

- To identify addresses where precision is a requirement
- As a location identifier for a number of citywide business processes including noticing, permitting, business registrations, etc.

Accepted values

- Street numbers are assigned according to rules laid out in AB-035, these specify:
  - The start and end point of address assignment
  - How many addresses are allocated between intersections and where that differs
  - Where even and odd numbers are assigned
- Authorized City staff enter address numbers in the Enterprise Addressing System according to these rules

> **Note on Units:** The City records unit numbers for condos to support tying property records for deeds, property taxes and other business processes. There is no formal requirement to record the units in rental buildings.

Reference Datasets

Dataset	Description and Constraints	Street Number Column
Addresses - Enterprise Addressing System	The EAS is the system of record for DBI when assigning official addresses. Associated coordinates are most often associated with the center of a parcel or close to it, rather than at the door or entry. This still allows associations, but it means that in certain cases a building footprint cannot be spatially matched via intersection or "point in polygon" with it's address(es).	`address_number`
Addresses with Units - Enterprise Addressing System	Same general limitations as the Addresses dataset above, but also includes sub-addresses like units. Unit numbers are formally referenced for condos because the City records these for the purposes of tying deeds and other property records to a specific unit and owner. Rental units are not formally recorded by the City.	`address_number`

Is anything wrong, unclear, missing?

Leave a comment.

Street Names

Definition

- The official name assigned to a segment of street or right-of-way that is legislated through the subdivision process and/or Board of Supervisors
  - Street names are generally established when streets are created as a result of the development / subdivision of land codified in the City's subdivision codes
  - Renaming streets can be initiated by members of the public or the Board of Supervisors according to the process documented by Public Works

> **Note:** The above only applies to city-owned public streets

Illustration

- Above is the street sign for Jack Kerouac Alley (formerly Adler Alley). On the street sign, both names are present for five years following a name change.

Authority

- New street names assigned during the development / subdivision of land
  - Recordation of final parcel maps including new streets happens with the Office of the Assessor-Recorder
  - Before recordation, subdivision maps are approved by the County Surveyor, the Public Works Director and the Board of Supervisors
  - Part of the process defined in the City Subdivision Codes
- Renaming of streets requires:

  - Petition with signatures submitted to Public Works for review with a submittal fee
  - The resolution referred to the Clerk of the Board of Supervisors
  - A Public Hearing at the Land Use and Economic Development Committee
  - Board of Supervisors approval
  - Mayor's signature

Use

- For official base maps to label the streets properly
- As a component part of a full address (see address formatting guidance)
- To validate against user submitted address data (e.g. in a form online)

Accepted values

- Official street names are maintained in the City's Official Basemap updated by Public Works staff
- The full list of valid City street names is available in the street names dataset

Reference Datasets

Dataset	Description and Constraints	Reference Columns
Street Names	Contains a list of officially valid street names contained in the City's Basemap	`fullstreetname` composed of `streetname` & `streettype` & `postdirection`
San Francisco Basemap Street Centerlines	A geographic reference of the all basemap streets including a number of street components like the valid name	`streetname` composed of `street` & `st_type`

Is anything wrong, unclear, missing?

Leave a comment.

Street Suffix Abbreviations

Definition

- A street suffix is a word that follows the name of the street describing its type (e.g. Street, Avenue, Road)
- Suffix abbreviations are shortened forms standardized by the United States Postal Service (USPS), these are the ones the City uses as well

Authority

- The USPS sets standards for addresses for consistency across the delivery of mail
- These are documented in USPS Publication 28: Postal Addressing Standards

Use

- When writing or recording a short form of a full street name
  - 1500 Market Street to 1500 Market St

Accepted values

- Standard street suffix abbreviations available online under Publication 28, Appendx C1
- Best not to encode with a period at the end
  - e.g. **ST** or **St** not St.

Is anything wrong, unclear, missing?

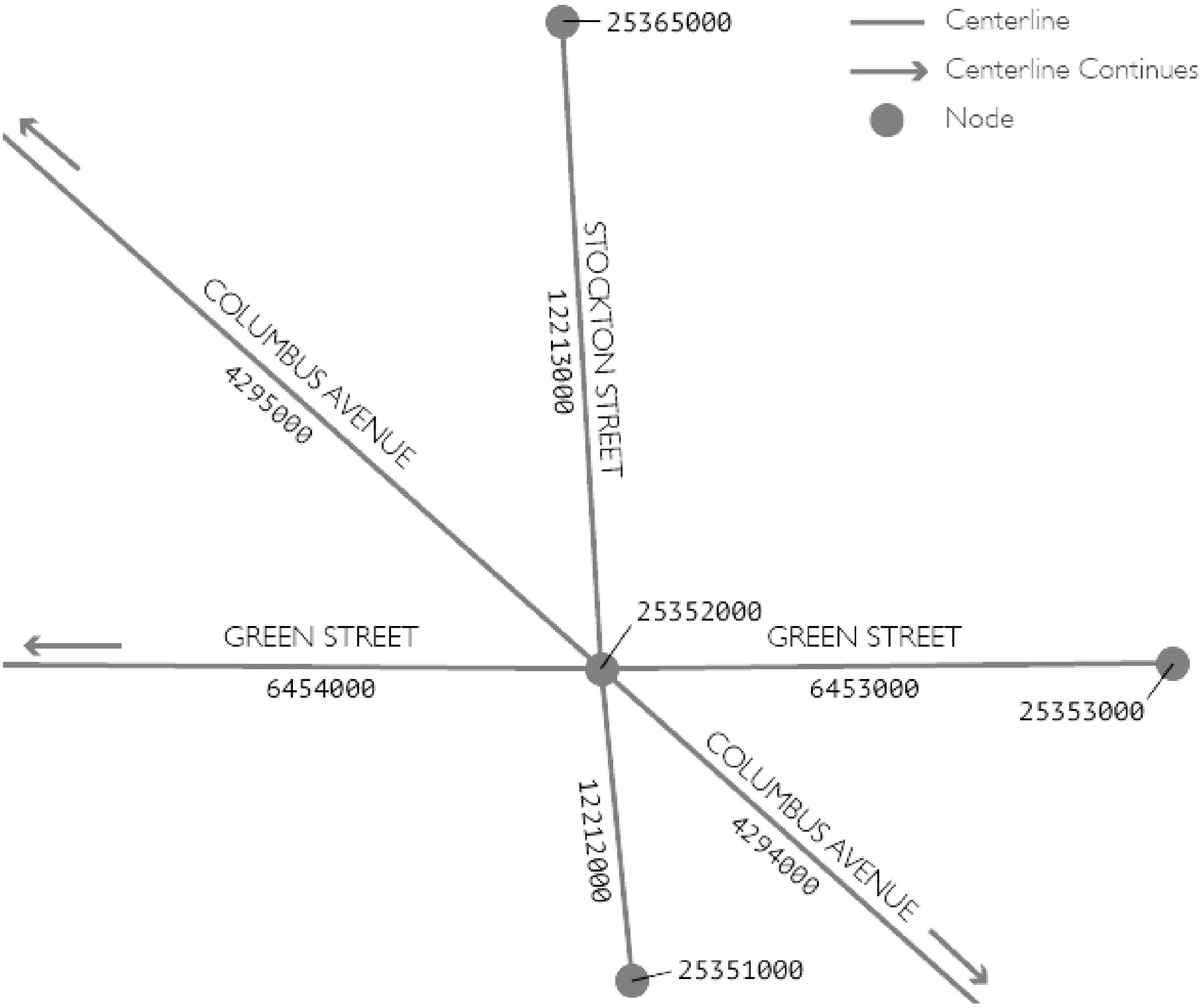
Leave a comment.

Street Centerlines and Nodes

Definition

- Street centerlines are lines that represent a network of streets
  - They are aligned generally to the center of a street
  - They are meant to model the street network and thus have no width or area
  - They have a length component
- Street nodes are the endpoints of a street centerline and represent intersections
  - A node shared among multiple intersecting street segments is an intersection
- Each node and centerline segment will have a unique Centerline Node Network (CNN) identifier
- The collection of Centerline Node Network identifiers are collectively known as "CNNs"

Illustration



- Shows 3 streets (Stockton, Green and Columbus) at a point of intersection
- Each segment sits between two nodes
    - A segment ends where it intersects with another segment OR at the physical end of a street (a dead end)
    - Some segments will start and end at the same node
- Each segment and node has a CNN identifier pictured above
- Segments share the same node where they intersect
    - Node ID 25352000 in the middle is shared by 6 segments

Authority

- The management of streets falls to different jurisdictions within the City
    - Public Works manages and maintains the majority of streets within the City
    - The remaining are managed and maintained by other entities like Caltrans, Presidio Trust National Park and Parks & Recreation, a summary of miles of streets by jurisdiction is available on the open data portal
- Basemap data including streets from various jurisdictions is maintained by Public Works

Use

- Centerline Node Network IDs (CNNs) are referenced in many datasets throughout the City (including but not limited to permits and inspections, project management and asset management systems)
- Used to enhance data by adding location attributes, allowing disparate datasets to be mapped as well as compared for analysis
- To model the transportation network

Accepted Values

- Every centerline and node will have a unique Centerline Node Network (CNN) identifier
    - `cnn` as a number
    - `cnntext` as a text string
- CNN IDs (CNN) may be used in secondary columns as reference
    - For example:
        - `f_node_cnn` and `t_node_cnn` to indicate from and to nodes
    - When referencing a CNN, include clear definition in the data dictionary, and include `cnn` in the column name
- Valid IDs are in the reference datasets below

Reference Datasets

Dataset	Description and Constraints	Reference Columns
List of Streets and Intersections	A list of street segments and intersections sorted by street name and ascending address number. This data set is based on the City's GIS basemap and contains CNN id numbers for each record.	`cnn` as number For segments: `from_cnn` and `to_cnn` define the node IDs at each end
San Francisco Basemap Street Centerlines	A geographic reference of the all basemap streets including centerline node network identifiers and jurisdictions	`cnn` as number `cnntext` as text `f_node_cnn` as the starting (from) node ID `t_node_cnn` as the ending (to) node ID
Street Segment and Intersection (CNN) Change Log	A list of Street Segment and Intersection (CNN) changes including new, dropped, realigned, divided and split records.	`oldcnn` as number `newcnn` as number

Is anything wrong, unclear, missing?

Leave a comment.

Reference: Boundaries

Common boundary references are used in numerous City datasets. This section distills some of the most common references. These include:

- Census
- Neighborhoods
- Supervisor Districts
- Zoning Use Districts

Census Boundaries

Census data is available from the Federal Census Bureau. For certain City administrative datasets, we assign census boundaries to make linking these to Census data easier.

For census boundary IDs we present the full ID starting with State ID and going down to the most granular ID represented by the field (e.g. tract, block or block group). The full IDs are presented as strings, not numbers. You can learn more about geographic boundaries and identifiers on the Census website. The full IDs are constructed in the following order:

```
State FIPS Code (2 digit) > County FIPS code (3 digit) > Tract ID (6 digit) > Blockgro
up ID (1 digit) > Block ID (4 digits, but first digit is the same as Blockgroup ID)
```

On City datasets with a Census geography column, we only represent the ID for the most granular geography appropriate to the data. For example, if we publish down to the Census block, we don't include a separate column for blockgroup or tract. One can derive these from the full ID because of the nesting relationship mentioned above.

Census Boundary	Example ID	Label
State	06	California
County	06075	San Francisco County, California
Census Tract	06075010100	Census Tract 101, San Francisco County, California
Census Blockgroup	060750101001	Block Group 1, Census Tract 101, San Francisco County, California
Census Block	060750101001000	Block 1000, Block Group 1, Census Tract 101, San Francisco County, California

Neighborhoods

The City's Open Data Program provides the Analysis Neighborhoods as the primary neighborhood district boundary on automated datasets. We also provide other neighborhood boundaries when appropriate.

The table below includes:

- the name and link to each of the neighborhood districts
- the human readable column name used on the open data portal
- the application programming interface (API) name
- the shortname used when there are character limits (e.g. in shapefile formats)
- the number of districts included in the dataset
- a quick link to download a CSV of just the boundary names (without geometry)

Dataset	Column Name (Human Readable)	API Name	Short Nar
Analysis Neighborhoods	Neighborhooods - Analysis Boundaries	neighborhoods_analysis_boundaries	NBHDANA
Neighborhood Groups	Neighborhoods - Group Boundaries	neighborhoods_group_boundaries	NBHDGRI
SF Realtor Neighborhoods	Neighborhoods - Realtor Boundaries	neighborhoods_realtor_boundaries	NBHDSFF
SFFind Neighborhoods	Neighborhoods - SFFind boundaries	neighborhoods_sffind_boundaries	NBHDSFF

Note: Datasets published before we codified this practice may not reflect the above. We are actively improving existing datasets on a rolling basis. Please consult the data dictionary and other related documentation under the dataset's About tab. If it's still unclear, contact DataSF, and we'll be happy to help.

Supervisor Districts

Definition

- There are 11 members of the Board of Supervisors, each representing a geographic district.

Illustration

Other Fields	Supervisor District
...	1
...	2
...	3
...	4
...	5
...	6
...	7
...	8
...	9
...	10
...	11

Use

- Primarily used for reporting by supervisor district

Accepted values

- Column name should be `Supervisor District` or `supervisor_district`
- Values between 1 and 11 (integer)
- Acceptable ways to indicate no district include:
    - `null` meaning the field has no value
    - `-1` or `0`

- Indicate how no district is represented in your data dictionary
- For example, not all 311 cases have a location and won't have an associated district

Reference Datasets

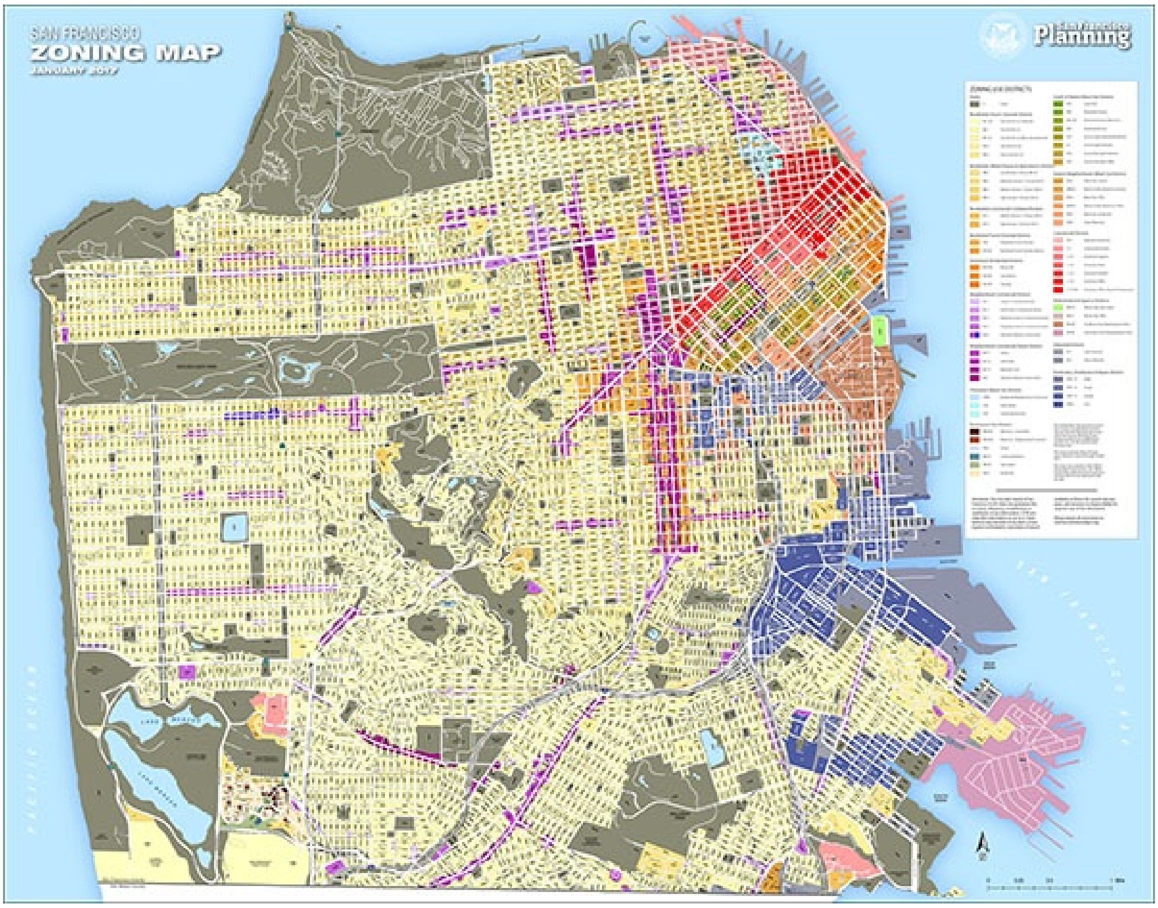
Dataset	Description and Constraints	Reference Columns
Current Supervisor Districts	Supervisor Districts as of the 2012 redistricting	`supervisor` - number of district (integer 1 through 11)

Zoning Use Districts

Definition

- Zoning regulations govern how land can be used in various geographic areas called "zoning use districts" (also known as "zoning," "zones" or "use districts").
- Zoning regulations may:
  - govern sizes and shapes of buildings
  - limit the number of units or apartments that can exists on a property
  - require the accommodation of car parking off of the street
  - set controls on planting street trees under certain circumstances
  - specify how late a business can remain open at night

Illustration



- Each part of the City is divided into zones that correspond to regulations in the Planning Code
- Get a higher resolution PDF version of the map above provided by Planning

Authority

- Zoning regulations are set out in the San Francisco Planning Code and modified through legislation
- The Planning Department enforces zoning compliance

Use

- For understanding what is permitted, conditional and not permitted when building in San Francisco

Reference

Reference	Description and Constraints	Reference Columns
Zoning Districts	The Zoning Districts are a component of the Zoning Map which in turn is a key component of the San Francisco Planning Code.	`url` links to the district definition in the planning codes `zoning` is the district code
Planning Code	The official Zoning Map can be found in the San Francisco Planning Code on the links under ZONING MAPS on the left navigation column).	N/A

Appendix: Reserved Column Names

The following column names should be used only if they adhere to the definitions in this guide:

- analysis_neighborhood
- date
- date_time
- fiscal_half_year
- fiscal_month
- fiscal_quarter
- fiscal_year
- half_year
- latitude
- longitude
- month
- quarter
- supervisor_district
- time
- week
- x_coord
- y_coord
- year
- zip_code

Is anything wrong, unclear, missing?

Leave a comment.

Appendix: Reference Data Index

Below is a table of the reference datasets mentioned in this document. View all the reference data below in the open data portal.

Dataset	Description and Constraints	Reference Columns	Page(s)
Addresses - Enterprise Addressing System	The EAS is the system of record for DBI when assigning official addresses. Coordinates are most often associated with the center of a parcel or close to it, rather than at the door or entry. This still allows associations, but it means that in certain cases a building footprint cannot be spatially matched via intersection or "point in polygon" with it's address(es).	`address_number`	Address Numbers
Addresses with Units - Enterprise Addressing System	Same general limitations as the Addresses dataset above, but also includes sub-addresses like units. Unit numbers are formally referenced for condos because the City records these for the purposes of properly tying deeds and other property records to a specific unit and owner. Rental units are not formally recorded by the City.	`address_number`	Address Numbers
	The footprint extents are collapsed from an earlier 3D building model provided by Pictometry of 2010, and have been refined from a version of building masses publicly available on the open data portal for over two years. The building masses were manually split with reference to parcel lines, but using vertices from the building mass wherever		

Building Footprints	possible. These split footprints correspond closely to individual structures even where there are common walls; the goal of the splitting process was to divide the building mass wherever there was likely to be a firewall. An arbitrary identifier was assigned based on a descending sort of building area for 177,023 footprints. The centroid of each footprint was used to join a property identifier from a draft of the San Francisco Enterprise GIS Program's cartographic base, which provides continuous coverage with distinct right-of-way areas as well as selected nearby parcels from adjacent counties.	`sf16_bldgid` unique identifier for footprint `mblr` for reference to property identifiers including parcels and right of way	Building Footprints

Department Code List	These department codes are maintained in the City's Financial System of Record. Department Groups, Divisions, Sections, Units, Sub Units and Departments are nested in the dataset from left to right. Each nested unit has both a code and an associated name.
	The dataset represents a flattened tree (hierarchy) so that each leaf on the tree has it's own row. Thus certain rows will have repeated codes across columns. Data changes as needed.
		Nested (right to left):
		`department_group_code`
		`division_code`
		`section_code`
		`unit_code`
		`sub_unit_code`
		`department_code`
			Department Names and Codes

Current Subdivision Parcels	These are the current active recorded parcels. The geography can be used as reference but should not be used for anything requiring precision.	`block_num` `lot_num` `blklot`	Parcels
	A list of street segments and intersections sorted by street name and ascending	`cnn` as number	

List of Streets and Intersections	address number. This data set is based on the City's GIS basemap and contains CNN id numbers for each record.	For segments: `from_cnn` and `to_cnn` define the node IDs at each end	Street Centerlines and Nodes
Recorded Parcel Geography with Transaction Date History	These are the current and historic parcels with recorded dates. Historic parcels only go back to about 1995 with some exceptions. Useful for tying historic administrative records to a location. The geography can be used as reference but should not be used for anything requiring precision.	`block_num` `lot_num` `blklot`	Parcels
San Francisco Assessor Blocks	Just the blocks without lots	`block_num`	Parcels
San Francisco Basemap Street Centerlines	A geographic reference of the all basemap streets including centerline node network identifiers and jurisdictions and street names by segment	`cnn` as number `cnntext` as text `f_node_cnn` as the starting (from) node ID `t_node_cnn` as the ending (to) node ID `fullstreetname` composed of `streetname` & `streettype`	Street Centerlines and Nodes & Street Names
Street Names	Contains a list of officially valid street names contained in the City's Basemap	`fullstreetname` composed of `streetname` & `streettype`	Street Names
Street Segment and Intersection (CNN) Change Log	A list of Street Segment and Intersection (CNN) changes including new, dropped, realigned, divided and split records.	`oldcnn` as number `newcnn` as number	Street Centerlines and Nodes

Appendix: Contributing

All of this documentation is open source and available to edit on GitHub. If you see something that you can contribute, submit a pull request with your edits! To make this easy you can click the *"Edit this page"* link at the top of the web docs.

The docs are all written in GitHub Flavored Markdown. If you've used GitHub, it's pretty likely you've encountered it before. You can become a pro in a few minutes by reading their GFM Documentation page.

Organizing Files

You'll notice that the GitHub Repo is in a logical structure. Each of the major sections is a folder. For example the 'Reference: Basemap' pages are in the folder `basemap` in the top level of the repository.

Some of the chapters are split into multiple sections to help break up the content and make it easier to digest. You can easily see how chapters are laid out by looking at the `SUMMARY.md` file. This convention helps keep chapters together in the file system and easy to view either directly on github or gitbook.

Table of Contents

You'll find the table of contents in the SUMMARY.md file. It's a nested list of markdown links. You can link to a file simply by putting the filename (including the extension) inside the link target.

Introduction Page

This is the root README.md file. It's intent is to give the reader an elevator pitch of what this document is about is and why we think it is useful.

Send a Pull Request

So that's it. You make your edits, keep your files and the Table of Contents organized, and send us a pull request.

Enjoy the Offline Docs

Moments after your edits are merged, they will be automatically published to the web, as a downloadable PDF, .mobi file (Kindle compatible), and ePub file (iBooks compatible).

Appendix: Acknowledgements

Many thanks to Singapore's Open Data Program for providing a Data Quality Guide for Tabular Data the bulk of which made its way into the chapter Data Structure and Formats with some additions and modifications.

Appendix: License