



Business Location and Type

Last Updated: 4-5-2018

There is no single source or publicly available dataset by the state of Colorado that contains current business point of sale locations and industry types. The [Business Entities](#) dataset is a great example of challenging oneself in the esoteric exercise of envisioning what data could possibly be available. At any given point in time, businesses are opening and closing, rebranding and relocating, and all around being difficult to track consistently. Not to mention variations that occur in simple but vital things like 'hours of operation', 'website address', 'stock of inventory,' etc., etc. On the other side of the equation is the classification of industry types. Things often seem clear at the most broad categories like 'Commercial', 'Industrial', 'Agriculture', etc., but what happens when the lines blur between those categories? Subcategories of course! And what happens with sub-categories? 'Distributions', 'Outliers', and 'Fuzzy Boundaries'. There are also compelling cases for looking at regions to see what businesses have recently closed, this too can add to the complexity of the database. So when one goes looking for "any data sets that relate business name/entity ID to type of business with current status," it can be understood that the answer to this question is "no" and "kind-of, but it's not exactly straight-forward." This is what makes working with Public Data both challenging and rewarding. Adding value to Public Data means adding value to data as a 'Secondary User.' The state collects business license registration data as a *secondary effort* to their primary function in registering new businesses. The has made starting a business and registering very easy, but in doing so does not force the business to declare an industry type. While a boon to some and frustrating to others, this is a classic case that government collects data for internal use, which often time is not exactly what a data scientist or programmer wants to see in the attributes of the data. Let the fun begin!

Publicly Available Datasets with Business Location or Type

The problem of identifying the 'point of sale' for each individual business as records in a single dataset is challenging for many datasets. For example, "Assumed Business Location" is geocoded and is referenced in the location field, however, location in the database is not always guaranteed to be the physical location of a business in every case. Some businesses may

choose to have mail sent to a separate location, some can be registered in CO but have an address outside of the state, some other businesses (wholesale for example) don't have a public access entry point, and finally, there are also businesses that have operations spread across several addresses.

Business Entities in Colorado.

The Colorado Secretary of State office provides the status of the over 1.6 million unique business IDs registered with the State of Colorado back to the 1800s in the [Business Entities in Colorado](#) dataset. There is a very rough average of 2,200 new businesses added in the State of Colorado each week. The Secretary of State has never been required to collect the name of the business owner, and if a business has an alias (ie Doing Business As) then this information can be found in the separate dataset, [Trade Names in Colorado](#). Business location is referenced primarily by the Principal Address information, but, again, location is not guaranteed to be physical location of business in every case.

Employee Counts by Industry.

The Colorado Department of Labor and Employment publishes the number of individuals employed in each County by for most of the 1,083 NAICS detailed industries for the nation in the [Employee Counts by Industry](#) dataset. The North American Industry Classification System (NAICS) is the standard used by Federal statistical agencies in classifying business establishments for the purpose of collecting, analyzing, and publishing statistical data related to the U.S. business economy. NAICS was developed and adopted in 1997 to replace the Standard Industrial Classification (SIC) system. The use of these codes in categorizing businesses enables analysis with this large wealth of data.

Publicly Available Datasets with Business Location by Type

Collectively there are rosters of businesses published by regulators of specific industries and entity types. This data can be used in combination to create a snapshot of selected industries across Colorado, many of which are maintained on a regular basis to accommodate changes over time. For datasets like Restaurant Inspections that are only available at a local level and for selected locations, competitors are encouraged to develop with the data as made available with the plan to utilize data from other locations in future iterations of the product.

Childcare Facilities and Healthcare Facilities

The Colorado Department of Human Services (CDHS) updates the [Licensed Child Care Facilities](#) data set monthly, and Colorado Department of Public Health and Environment updates [Health Facilities](#) annually. The business names of the child care facility programs and the physical location addresses. For both datasets, each location, the total licensed capacity and

the types of services and programs are provided. Health Facilities has related datasets for Community Behavioral Health Centers, Drug Treatment Programs and Resources Facilities.

Restaurant Inspections in Tri-County Colorado

The Tri-County Health Department provides public health services to Adams, Arapahoe, and Douglas counties, including public inspections of restaurants, grocery stores, and wholesalers. [Restaurant Inspections in Tri-County Colorado](#) includes the point of sale location (address and lat/long) for all entities regulated in the three counties.

Assessor Data for Counties in Colorado

The [Statewide Aggregate Parcels in Colorado 2017](#) dataset was created to facilitate a statewide view of parcels. The fields in this dataset have been intentionally selected to facilitate mapping attribute data across municipal boundaries and published by the Governor's Office of Information Technology. In many cases, but not always, the assessor data is housed and managed with the municipalities' GIS shop. For the management of the parcel data, the official record is not the GIS file, instead it is the actual Legal Land Survey, and most parcel data should be viewed as an approximate location unless the data specifies otherwise. Parcel data includes the address of the location and the name and address of the owner. Quite often this does not provide context for what business is operating at the location, but sometimes it does.

countyFips	parcel_id	situsAdd	sitAddCty /	sitAddZip	owner	owner2	ownerAdd	ownAddCty	ownAddStt	ownAddZip
014	157535115006	370 AGATE ST	BROOMFIELD	NULL	RODR...	NULL	NULL	BROOMFIELD	CO	80020-1924
014	157528207098	13456 VIA VARRA	BROOMFIELD	NULL	PAND...	NULL	NULL	BROOMFIELD	CO	80020

This data includes some categorical information that can be extracted from the 'zoningCode' field, designating the location as residential, commercial, industrial, agricultural, and a variety of other codes, which can be used to validate business types in given locations.

legalDesc	landSqft	block	lot	bCoc	subName	zoningCode
Subd: BR...	7501.2...	18	38	NULL	BRM HEIGHTS #1 & AMD	R-1

The assessor data provided about each site includes information like the value of the land as compared to the appraised value and the most recent sale date and sale price.

saleDate	salePrice	apprValTot	asedValTot
2014-10-29	253000	268000	21330

Colorado Anchor Institutions. Locations of schools, libraries, medical and healthcare providers, public safety entities, community colleges and other institutions of higher education, and other community support organizations and entities can be found in the [Community Anchor Institutions dataset](#). These locations are gathered in the National Telecommunications Information Administration (NTIA) and State Broadband Data Development Program (SBDD) in an effort to identify existing and potential broadband availability to the public.

The Homeland Infrastructure Foundation-Level Data.

The HIFLD Subcommittee was established in 2002 to address improvements in collection, processing, sharing and protection of National geospatial information, and the [Homeland Infrastructure](#). 'location by type' data includes DHL Facilities, Landfills, Federally Insured Bank Locations, Major Sport Venues, Overhead Cables, Critical Habitats, Watershed Boundaries, Dam Areas, Bridge Areas, Landslide Areas, and from Geographic Names Information System (GNIS): Structures, Historical Features, Cultural Features, Landform Features, Community Features.

Proprietary Datasets with Business Location and Type

Datasets in this category require adherence to a restrictive use licence. Use of the data constitutes agreement to the license terms and conditions. Some companies make data available at a cost. These resources are available for free within the constraints of the use license.

Better Business Bureau

One of the Go Code Colorado In-Kind partners this year includes the Colorado Better Business Bureau, who provides their [BBB API](#) based on their work to collect and provide free business reviews, and to serve as an intermediary between consumers and businesses. This does not provide a list of all Businesses for Colorado, but does provide contextual information in the form of consumer reviews, as well as a corroboration point for active business location validation and some business types. GoCode Colorado participants are allowed to download and use the data offline in any way they see fit, as long as they refresh the data from the BBB according to their terms and conditions.

Google Places API

Google Developers' have 'Place Types' in their [Google Places API](#), however, use restrictions permit the creation of an offline copy of the data and limit their 'calls per API key' to 25,000 a day.

Yelp

Yelp has an extensive API library to access their business location information and business type information. However, users must abide by their terms and conditions, such as putting an attribution on their site ("powered by Yelp"), and not creating an 'offline' copy of the database by collectively saving daily pulls.

Factual

This data source, [factual](#), is a leader in providing reliable business point of sale data, and is available via a freemium model.

Open Datasets with Business Location and Type

[Open Street Map](#) has millions of points of interest which will serve well for this purpose. These are "nodes" (points) each tagged with a set of keywords that indicate what each point is. For example, a node with latitude 51.398, longitude -1.082, and tags amenity=place_of_worship, name=St Mary's Church. This indicates that there's a church called St Mary's Church at that lat and long. To get this data out of OSM, the first step is to identify which tags of interest. For example, everything tagged as a church: in other words, everything with amenity=place_of_worship. (The first part of the tag is the 'key', the second the 'value'. Find the long list of common OSM tags on the [OSM wiki](#)). The next step of exporting the data can actually be done in two ways. One is by using XAPI, the Extended API. XAPI allows for querying the server for data matching certain tags. It can be very slow, so it certainly isn't suitable for live use (don't use OSM servers for a live app anyway), but it's an easy way to download a lot of data matching certain criteria. The ideal way is to download the OpenStreetMap data dump and filter it according to needs. The data dump is called planet.osm. Either download the whole world (really huge), or just an extract covering one country or state <http://download.geofabrik.de/north-america/us/colorado.html>. Once downloaded, a Java program called Osmosis will do all the hard work of extracting the data, and either loading the result into a database or outputting it as a file. Although it is an open database, OSM data does have a use license. Be sure to read and understand if planning to productionize off this data.

DIY Business Location and Type Dataset

For selected industries in Colorado, rosters of registered business locations can be used to create a 'slice' of the Point of Sale (POS) locations. The goal of Go Code Colorado is to add value from combining data to create new valuable data for use in analysis. [Business Entities](#) is one of the best examples of how value can be created from Public Data.

- The problem with the Google API is that it doesn't have accurate data on whether a business is open/closed (there is out of date and incorrect data).
- The problem with the "Business Entities in Colorado" dataset is that it doesn't have reliable information on the point of sale location for all business (it contains mailing address, which in many cases is different than point of sale address)
- Mash em up and together you've got something!

Depending on the use case for the tool being built, there are a great many different crosswalks that can be designed to create the "Business Location and Type" dataset. Common studies that are done require the creation of a category column in the main location database, wherein each record is validated and analyzed against the NAICS codes to populate the category column.

Staging areas for incoming data allow for productionizing on incoming Public Data. Next add in Google Places API and BBB API as validators as they both will give the location for the point of sale for many businesses, along with the category of the industry, but cannot be retained locally in bulk. Then, depending on the 'flavor' of the need of the tool being built, incorporate any one or more of the additional data sources to create rulesets that improve the confidence in the 'business location' and 'business type' match, as well as the validity of associated variables like 'hours of operation', etc.

Questions Answered by this Document

- Any data sets that relate business name/entity ID to type of business?
- What datasets allow for comparison of businesses by delinquent status and location?
- Any public data sets that would help identify the industry in which a business provides product or services?