

Slow Small and Varied Data Management for Public Policy

5 September, 2019

The San Diego Regional Data Library

Eric Busboom, eric@sandiegodata.org

<http://sandiegodata.org>

Slow Small and Varied Data Management for Public Policy

5 September, 2019

The San Diego Regional Data Library

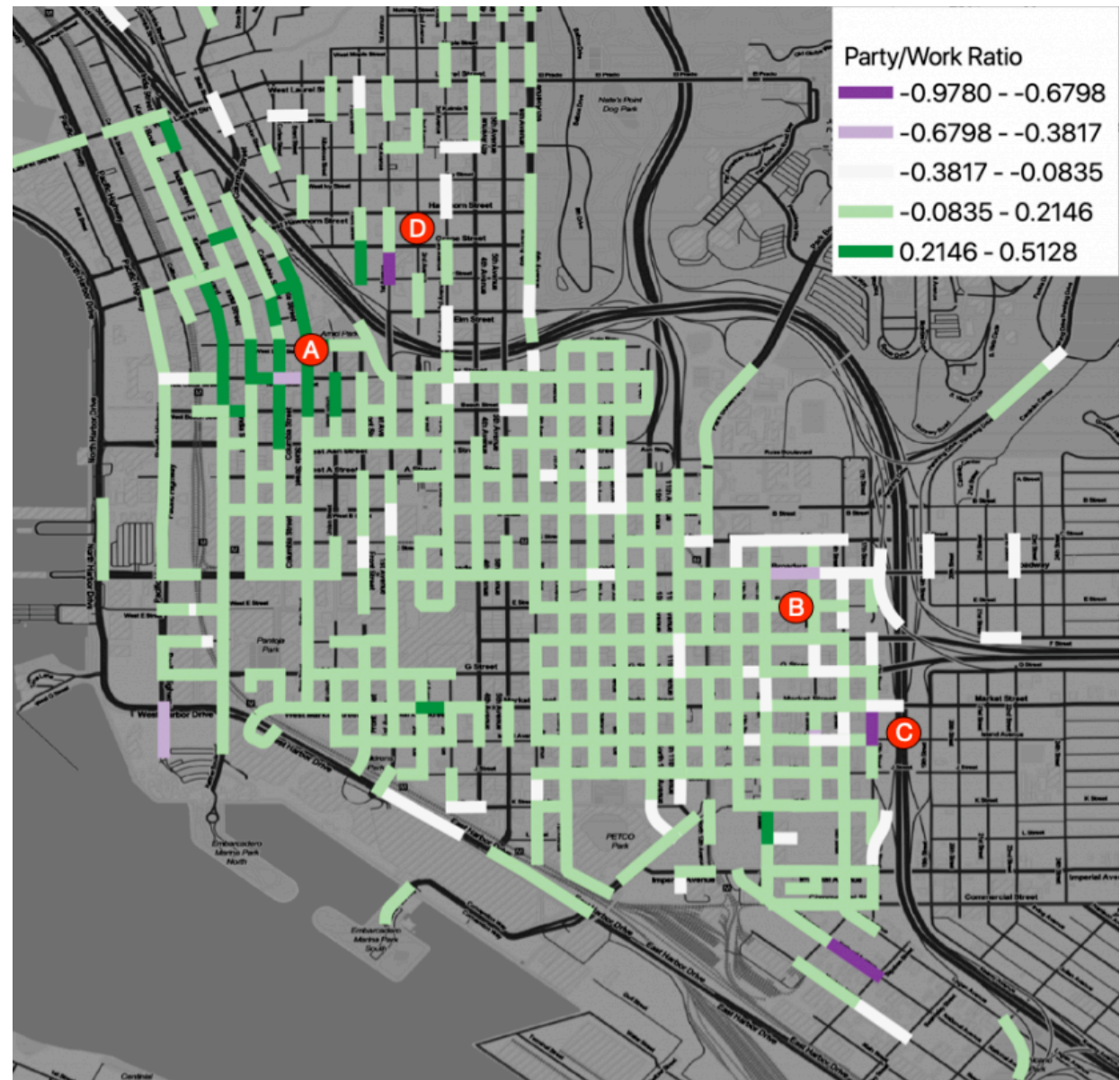
Eric Busboom, eric@sandiegodata.org

<http://sandiegodata.org>

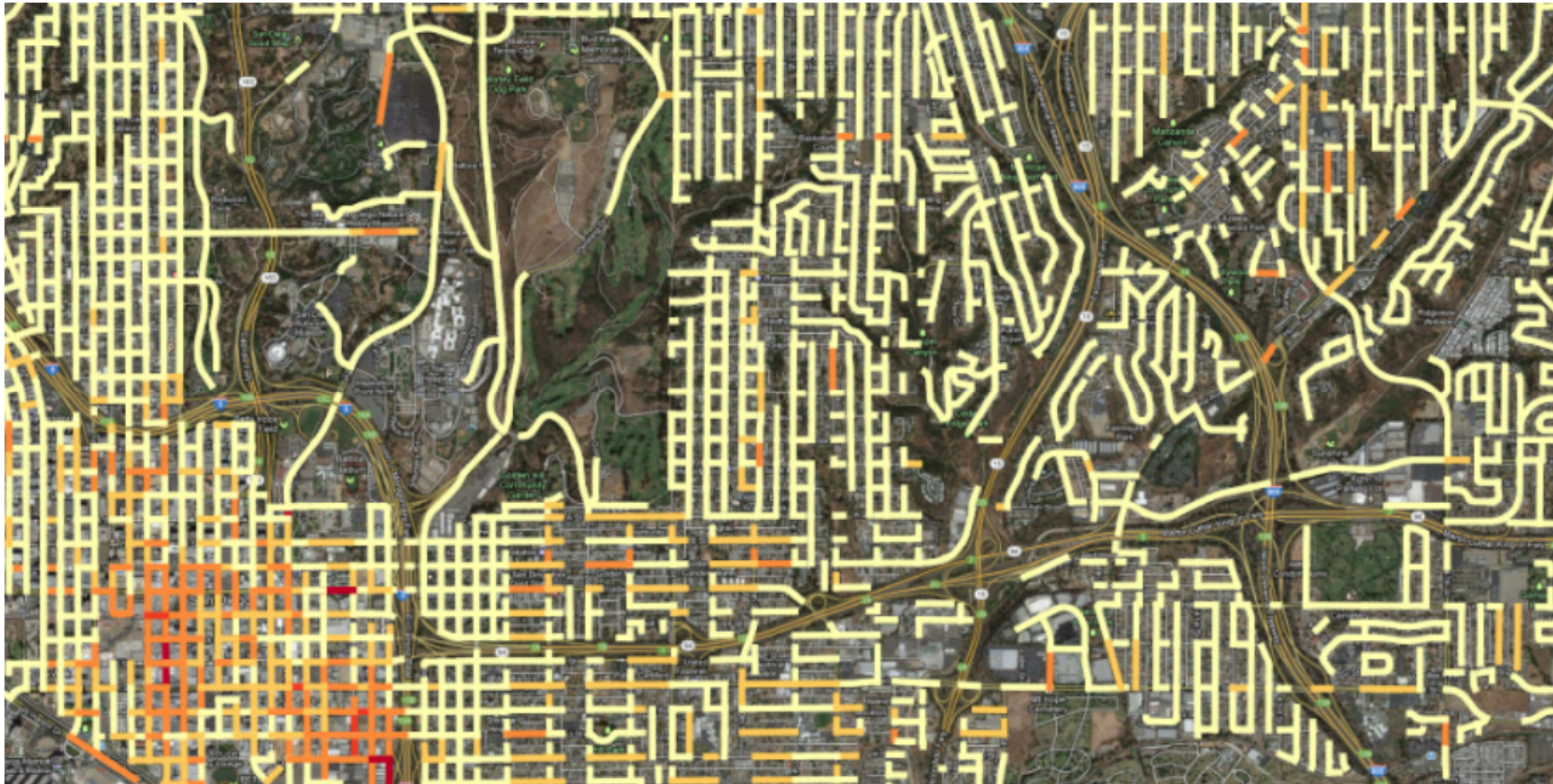
The Library's Mission

Provide volunteer data
services to San Diego

- Nonprofits
- Governments
- Journalists



Pedestrian Traffic by Street Segment,
Downtown San Diego



San Diego Downtown Crime Intensity Heat map, by street segment

Search

Collections

- [California Department of Education](#)

- [Urban Planning](#)

- [Homelessness](#)

The [San Diego Regional Data Library](#) ensures civic, social, governments and journalists in San Diego county have usable neighborhood level data they need to run their operations and tell their stories, so they can inform and influence grantors, decision makers and the public.

[Visit the Data Library's main website](#) for details of our work, projects and ideas.

Homelessness

Datasets related to homelessness, for San Diego and Los Angeles.

National Datasets

- [HIC Facility Geocoded Addresses](#) - Geocoded addresses for housing facilities in the San Diego CoC.
- [HUD Housing Inventory Count](#) - Project-level housing inventory for 2012 to 2018.
- [HUD Point in Time Counts by CoC](#) - Combined years of sheltered and unsheltered homeless counts, by CoC.

Los Angeles Datasets

- [San Diego Point In Time Count 2018, Tracts](#) - Unsheltered counts per tract and sheltered counts per facility, for 2018 San
- [Homelessness Risk Tutorial](#) - Data related to homelessness from an ArcGIS tutorial
- [LA Homelessness Demographic Survey](#) - Standardized Records from the Demographic Survey of Unsheltered Individuals
- [LAHSA Homeless Count](#) - Counts of homeless per tract, from the Los Angeles Point In Time Count, for 2018

San Diego Datasets

- [San Diego Point In Time Count 2010, Geo Points](#) - Point data for locations of homeless for the 2010 PiTC count in San Diego

Recent Posts

[San Diego Unified School Enrollment Boundaries](#)
[Enrollment by school](#)
[Demographics for California School Districts](#)
[California School District Links to US Census](#)
[San Diego Point In Time Count 2010, Geo Points](#)

Categories

[Analysis](#)
[build_environment](#)

3Vs of Big Data

- Small
- Slow
- Varied



Industrial Process For Data

- 80% of effort in data preparation
- Wrangling, Analysis, Reporting are different roles
- We've solved problems like this before!



Non Requirements

- Hadoop, Spark, Data Lake
- Databases
- Clusters

Requirements

- Easily separate roles
- Version control, repeatability
- Provenance; Where did it come from?
- Documentation, Metadata
- Easy to use data in Jupyter, Excel, Tableau

Metatab

- Structured metadata in a CSV file
- Use spreadsheet for formatting data
- Easy single-file data packages in Excel



<http://metatab.org>

Metatab Features

- Easy structured metadata; uses CSV
- Package dependencies, provenance
- Custom URLs handle complexity of tabular data
- Load data from URLs, programs or Python functions
- Publish to Wordpress, CKAN, S3, Data.world

Metatab

	A	B	C	D	
1	Declare	metatab-latest			
2	Title	California Residential Elder Care Facilities			
3	Description	A list of licensed Residential Elder Care Facilities (RCFEs) from the California Department of Social Services.			
4	Identifier	18a22824-1f07-425d-8651-4d4d1975271f			
5	Name	cdss.ca.gov-residential_care_facilities-2017-ca-4			
6	Time	2017			
7	Version	6			
8					
9	Section	Resources	Name	Description	Source
10	Reference	home/getstatedata/ResidentialElderCareFacility	raw_facilities	Unprocessed list of RCFE	
11	Datafile	program:scripts/process_facilities.py	facilities	Processed list of RCFE	
12	Datafile	program:scripts/geocode.py	geocodes	Facilities geocoded to localtions, tracts and blocks with the Census geocoder.	
13					
14	Section	Contacts	email	Organization	Url
15	Origin	dss.ca.gov		California Department of Social Services	bySea
16	Wrangler	Eric Busboom	eric@sandiegodata.org	San Diego Regional Data Library	ownl a.org
17					
18	Section	Schema	DataType	AltName	Desc
19	Table	facilities			
20	Table.Column	Facility Type	text	facility_type	
21	Table.Column	Facility Number	integer	facility_number	
22	Table.Column	Facility Name	text	facility_name	

Create a New Package

- **mp new -o example.com -d datasetname**
- Add Urls for resources and references
- Add metadata, contact, documentation links
- Schemas / Data Dictionaries automatically generated

Metatab Urls

- Urls specify resource, target in resource
- Can indicate encoding, start line, format, etc.
- http://.../simple-example.foo#&target_format=csv
- http://.../test_data.zip#renter_cost_excel07.xlsx;2
- <gs://1VGEkgXXmpWya7KLkrAPHp3BLGbXibxHqZvfn9zA800w>
- metatab+http://.../example.com-simple_example-2017-us-1#random-names
- `python:pylib#implicit_dataframe`

Resources and References

References

- Metatab Urls
- Not included in output packages
- Inputs to resource programs

Resources

- Metatab Urls
- Included in output package

Build a Package

- **mp build**
- **Creates new Metatab package from Metatab source**
- **Can build Filesystem, Zip, CSV, Excel packages**

```
cde.ca.gov-frpm — mp build — 60x24
(metatab-dev) crispin:cde.ca.gov-frpm eric$ mp build
Name is: cde.ca.gov-frpm-2
Building fs package (Package doesn't exist)
Loading documentation for 'README', 'README.md' to '/Users/eric/proj/data-projects/cde.ca.gov/cde.ca.gov-frpm/_packages/cde.ca.gov-frpm-2/./README.md'
Reading resource frpm04 from https://www.cde.ca.gov/ds/sd/sd/documents/frpm0405.xls#1
Loading data for 'frpm04'
Processed 9368 rows in 1.0 sec, rate = 14431.51 rows/sec
Reading resource frpm05 from https://www.cde.ca.gov/ds/sd/sd/documents/frpm0506.xls#1
Loading data for 'frpm05'
Processed 9497 rows in 1.0 sec, rate = 15553.27 rows/sec
Reading resource frpm06 from https://www.cde.ca.gov/ds/sd/sd/documents/frpm0607.xls#1
Loading data for 'frpm06'
Processed 9626 rows in 1.0 sec, rate = 14677.83 rows/sec
Reading resource frpm07 from https://www.cde.ca.gov/ds/sd/sd/documents/frpm0708.xls#1
Loading data for 'frpm07'
Processed 9782 rows in 1.0 sec, rate = 14787.63 rows/sec
Reading resource frpm08 from https://www.cde.ca.gov/ds/sd/sd/documents/frpm0809.xls#1
```

Publish to S3

- Publish to packages to S3
- Can access ZIP of all files, or files individually

```
$ mp s3 -s library.metatab.org
Packaged saved to: metapack+http://library.metatab.org/cde.ca.gov-frpm-2/

Wrote these files:
path                                     url
-----
_packages/cde.ca.gov-frpm-2.csv         http://library.metatab.org/cde.ca.gov-frpm-2/
cde.ca.gov-frpm.csv                     http://library.metatab.org/cde.ca.gov-frpm.csv

Skipped these files:
path                                     url
-----
index.html                             http://library.metatab.org/cde.ca.gov-frpm-2/index.html
datapackage.json                       http://library.metatab.org/cde.ca.gov-frpm-2/datapackage.json
README.md                              http://library.metatab.org/cde.ca.gov-frpm-2/README.md
metadata.csv                           http://library.metatab.org/cde.ca.gov-frpm-2/metadata.csv
data/frpm15.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm15.csv
data/frpm14.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm14.csv
data/frpm16.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm16.csv
data/frpm17.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm17.csv
data/frpm13.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm13.csv
data/frpm07.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm07.csv
data/frpm06.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm06.csv
data/frpm12.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm12.csv
data/frpm04.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm04.csv
data/frpm10.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm10.csv
data/frpm11.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm11.csv
data/frpm05.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm05.csv
data/frpm08.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm08.csv
data/frpm09.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm09.csv
data/free_rp_meals.csv                 http://library.metatab.org/cde.ca.gov-frpm-2/data/free_rp_meals.csv
data/frpm18.csv                        http://library.metatab.org/cde.ca.gov-frpm-2/data/frpm18.csv
notebooks/Explore.ipynb                http://library.metatab.org/cde.ca.gov-frpm-2/notebooks/Explore.ipynb
index.html                             http://library.metatab.org/cde.ca.gov-frpm-2/index.html

Synchronized these Package Urls
-----
http://library.metatab.org/cde.ca.gov-frpm-2/metadata.csv
s3://library.metatab.org/cde.ca.gov-frpm-2.csv
http://library.metatab.org/cde.ca.gov-frpm-2.csv
http://library.metatab.org/cde.ca.gov-frpm.csv
-----
$
```

Publish to Wordpress, CKAN, Data.World

- S3 packages can be linked to data repositories
- Wordpress, CKAN and Data.World are supported

San Diego Unified School Enrollment Boundaries

August 14, 2019 by metapack

Geographic boundaries for schools in the San Diego Unified school district

sandiegounified.org-enrollment_zones-2014-1

[Resources](#) | [Packages](#) | [Documentation](#) | [Contacts](#) | [References](#) | [Data Dictionary](#)

Resources

- [sdusd_boundaries](#). Select boundaries of school enrollments areas in San Diego Unified school district.
- [school_tract_xwalk](#). Porportions of overlap between tracts and school enrollment areas.

Documentation

Building

The data/sdusd.zip file is built manually from the three data/*.zip files, using the notebooks/CombineFiles.ipynb notebook

Use in Jupyter

Accessing Packages in Metapack

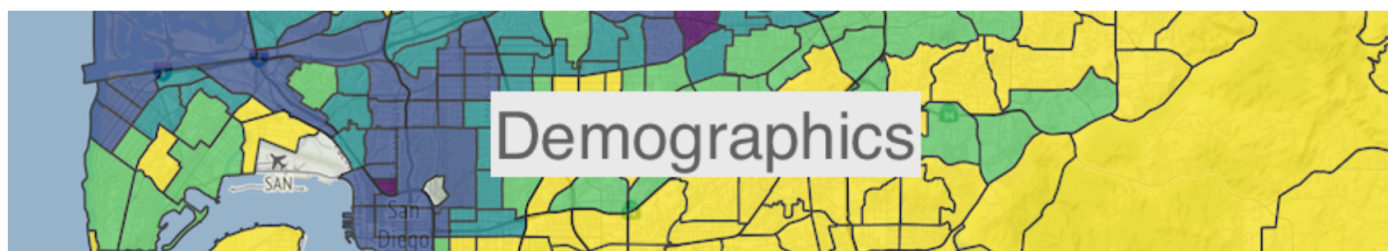
```
import metapack as mp
pkg = mp.open_package('http://library.metatab.org/sandiegounified.org-enrollmen

# Create Dataframes
sdusd_boundaries_gdf = pkg.resource('sdusd_boundaries').geoframe()

school_tract_xwalk_df = pkg.resource('school_tract_xwalk').dataframe()
```

- Easy access to packages in Python/Jupyter
- Method to create Pandas Dataframes
- Example access code included in data repository pages

School Versus Enrollment Area Demographics



eric Education 0 Comments

Eric Busboom.

A descriptive analysis of the differences between the demographics of schools versus the schools' enrollment areas, for schools in the San Diego Unified district.

 Full notebook on github

Since at least the early 1980's, San Diego unified has pursued a voluntary busing policy for desegregating neighborhood schools. The core of that policy is Magnet Schools, high-quality schools oriented around specific topics or educational goals, which were preferentially situated in lower-income and minority neighborhoods. Along with a generous city-wide busing program, the magnet schools would encourage higher-income families to voluntarily send their children to schools in lower income areas. The District also made it easier for lower-income parents to send their children to schools in higher-income neighborhoods.



Recent Posts

[School Versus Enrollment Area](#)

[Demographics](#)

[Pedestrians Rhythms By Neighborhood](#)

[Los Angeles Homelessness By Land Use](#)

[Los Angeles Homelessness Clusters](#)

[Crime Rhythm Maps](#)

Recent Comments

Archives

[August 2019](#)

[April 2019](#)

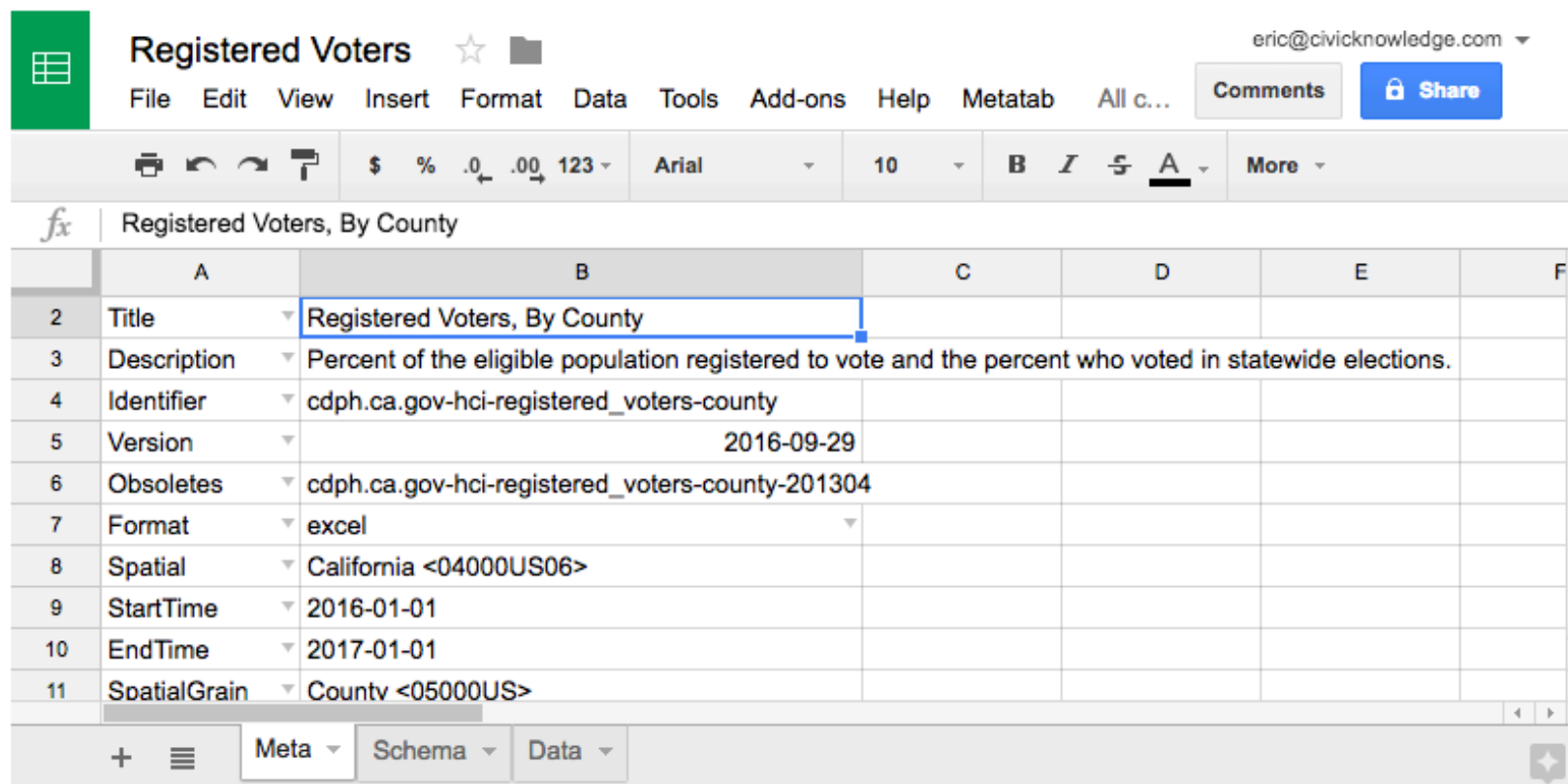
[January 2019](#)

<http://metatab.org>

Metatab

Metadata for Mortals

Metatab stores metadata in a spreadsheet, alongside data, ensuring that the metadata is easy to create, easy to read, and cannot be separated from the data.



The screenshot shows the Metatab interface. At the top, there's a green icon and the title 'Registered Voters'. Below it is a menu bar with options: File, Edit, View, Insert, Format, Data, Tools, Add-ons, Help, Metatab, and All c... There are also buttons for 'Comments' and 'Share'. Below the menu bar is a toolbar with various icons for formatting and editing. The main area is a spreadsheet titled 'Registered Voters, By County'. The spreadsheet has columns A through F. The first column (A) contains metadata fields, and the second column (B) contains the corresponding values. The metadata fields include Title, Description, Identifier, Version, Obsoletes, Format, Spatial, StartTime, EndTime, and SpatialGrain. The values are: Registered Voters, By County; Percent of the eligible population registered to vote and the percent who voted in statewide elections.; cdph.ca.gov-hci-registered_voters-county; 2016-09-29; cdph.ca.gov-hci-registered_voters-county-201304; excel; California <04000US06>; 2016-01-01; 2017-01-01; County <05000US>.

	A	B	C	D	E	F
2	Title	Registered Voters, By County				
3	Description	Percent of the eligible population registered to vote and the percent who voted in statewide elections.				
4	Identifier	cdph.ca.gov-hci-registered_voters-county				
5	Version	2016-09-29				
6	Obsoletes	cdph.ca.gov-hci-registered_voters-county-201304				
7	Format	excel				
8	Spatial	California <04000US06>				
9	StartTime	2016-01-01				
10	EndTime	2017-01-01				
11	SpatialGrain	County <05000US>				

[CLI Tools](#)

[Install Add-On](#)

[Metatab Spec](#)

[Data Pacakage
Spec](#)

Slow Small and Varied Data Management for Public Policy

The San Diego Regional Data Library

Eric Busboom, eric@sandiegodata.org

<http://sandiegodata.org>