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
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.

San Diego Downtown Crime Intensity Heat map, by street segment

Search

Search for datasets

Collections

- California Department of Education
- Urban Planning
- Homelessness
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
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
- 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
<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Declare</td>
<td>metatab-latest</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Title</td>
<td>California Residential Elder Care</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Facilities</td>
<td>Facilities (RCFEs) from the California</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Description</td>
<td>Department of Social Services.</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Identifier</td>
<td>18a22824-1f07-425d-8651-4d4d1975271f</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Name</td>
<td>cdss.ca.gov-residential_care_facilities-2017-ca-4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Time</td>
<td>2017</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Version</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Section</td>
<td>Resources</td>
<td>Name</td>
</tr>
<tr>
<td>10</td>
<td>Reference</td>
<td>home/getstatedata/ResidentialElderCareFacility</td>
<td>raw_facilities</td>
</tr>
<tr>
<td>11</td>
<td>Datafile</td>
<td>program/scripts/process_facilities.py</td>
<td>facilities</td>
</tr>
<tr>
<td>12</td>
<td>Datafile</td>
<td>program/scripts/geocode.py</td>
<td>geocodes</td>
</tr>
<tr>
<td>13</td>
<td></td>
<td>Facilities geocoded to locations, tracts</td>
<td>blocks with the Census geocoder.</td>
</tr>
<tr>
<td>14</td>
<td>Section</td>
<td>Contacts</td>
<td>email</td>
</tr>
<tr>
<td>15</td>
<td>Origin</td>
<td>dss.ca.gov</td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Wrangler</td>
<td>Eric Busboom</td>
<td><a href="mailto:eric@sandiegodata.org">eric@sandiegodata.org</a></td>
</tr>
<tr>
<td>18</td>
<td>Section</td>
<td>Schema</td>
<td>DataField</td>
</tr>
<tr>
<td>19</td>
<td>Table</td>
<td>facilities</td>
<td></td>
</tr>
<tr>
<td>20</td>
<td>Table.Column</td>
<td>Facility Type</td>
<td>text</td>
</tr>
<tr>
<td>21</td>
<td>Table.Column</td>
<td>Facility Number</td>
<td>integer</td>
</tr>
<tr>
<td>22</td>
<td>Table.Column</td>
<td>Facility Name</td>
<td>text</td>
</tr>
</tbody>
</table>
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
Publish to S3

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

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

# 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 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.

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.
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.
Slow Small and Varied Data Management for Public Policy

The San Diego Regional Data Library

Eric Busboom, eric@sandiegodata.org
http://sandiegodata.org