

American Community Survey
PostgreSQL Database Integration



PROJECT PROPOSAL

US Census American Community Survey 2012
PostgreSQL Database Integration

April 2014 rev 1405

contact: Brian M Hamlin maplabs@light42.com

Overview:

The **America Community Survey** (ACS) is a collection of data products distributed periodically by the U.S. Federal Government [1] – the ACS 2012 5 year study was published in December 2013. There exist a very limited set of tools to directly access the vast statistical content therein.

Proposal:

Ingest and index the contents of the ACS 2012 5 Year product using widely available database technology. Publish the scripts and procedures used as GPL free software. (value-add front-ends or other uses are left to interested parties)

Technical Details and Proof-of-Concept:

State-by-state ingestion has been performed on a trial basis, for California, New Hampshire, Maine and Colorado, using a combination of PostgreSQL and python language scripts. [2] The result is ~1480 database tables per state, with summary data for thousands of standardized geographical entities, including county, census tract and metropolitan statistical areas (MSA).

Some Questions and Answers on the Proposed Project:

What benefit does this provide to a user that isn't already provided by the Census API?

Local storage and retrieval creates a responsive research environment not currently available with online search; SQL access to all tables; enables graphical styling of data using a wide variety of toolchains; enables data exploration, visualization, and re-aggregation of data.

Can you provide an example of an anticipated use case?

Build visualizations of sets of data using available software toolchains

A researcher knows the name of a table of interest (in census nomenclature) – query for all available geographies for that summary; query for all available geographies under a primary geography, available for that summary (e.g. all census blocks in Los Angeles county)

A researcher does not know the specific field names in tables of interest – explore the tables using SQL wildcards to find available data.

A researcher wants to get a non-standard summary of tables, built with standardized geographical entities – write a SQL query to aggregate and export the custom set.

[1] https://www.census.gov/acs/www/about_the_survey/american_community_survey/

[2] http://ct.light42.com/www_shared/Census_Misc/census20125_desc0.txt.bz2

Signatures are being gathered of interested parties to substantiate the effort and enable a resource gathering inquiry.

Parties with Interest in this Project:

Dr. Jon Stiles
Executive Director, Census Research Data Center (RDC)
UC Berkeley representative
ICPSR; The Roper Center; Census State Data Center
University of California, Berkeley

-

Nathaniel Roth Programmer IV-Supervisor
Information Center for the Environment (ICE)
Department of Environmental Science & Policy
University of California Davis

-

Raymond Yee
Lecturer, School of Information, UC Berkeley

-

Dr. Saleh Elmohamed
Senior Researcher, Mathematics and Engineering
Cornell University; UC Berkeley EECS, College of Engineering

-

Garlynn Woodsong
Project Manager, Urban Design / Transportation Planning
Calthorpe and Associates

(add your name and title here)

Technical Appendix A: Q&A on database details

== General

-- where is the base data ?

http://www2.census.gov/acs2012_5yr/summaryfile/2008-2012_ACSSF_By_State_All_Tables/

== Tables

-- what tables are available ? (disregard imputations tables)

```
SELECT
  distinct(table_subject_area)
FROM
  acs20125_tables_list
ORDER BY
  table_subject_area;
#--
SELECT
  table_indb_name,
  table_subject_area,
  table_universe,
  table_title
FROM
  acs20125_tables_list
WHERE
  table_subject_area <> 'Imputations'
ORDER BY
  table_subject_area,
  table_universe,
  table_title;
```

-- what geographies are available ?

```
SELECT   geoid, geo_name
FROM     acs20125_geo_defs_ca
ORDER BY length(geoid), geoid;
```

```
select length(geoid) as geoid_len, count(*) from acs20125_geo_defs_ca
group by length(geoid)
order by length(geoid);
```

geoid_len	count	
9	12	-- statewide summary
11	53	-- congressional district
12	186	-- counties; Combined Statistical Area (CSA); Senate dist; Assm Dist
13	216	-- Native Am reservations
14	3189	-- Census-designated place (CDP); School districts
16	114	-- more Native Am reservations
17	2284	-- Census county division (CCD); Urban Area (ua); Urban Cluster (uc) http://www.census.gov/geo/reference/ua/urban-rural-2010.html
18	8203	-- tracts; more Native Am resrv
19	23438	-- block groups; metro areas
22	2245	-- CCDs
28	12842	-- census tracts

Note there are close to 900 tables populated with 21513 rows each
one row per geoid. So not all of these entities are represented.

In this system, the geoid is found by JOINing to acs20125_seq_metadata
getting the e_logrecno, and JOINing to acs20125_geo_defs_ca

== Table 60

(pick an example table and describe it in detail)

```
SELECT
  table_indb_name,
  table_id, table_seq_num,
  table_title, table_subject_area,
  table_universe
FROM acs20125_tables_list
WHERE
  table_indb_name = 'acs_mtable_60_raw';
```

```

-[ RECORD 1 ]-----+-----
table_indb_name   | acs_mtable_60_raw
table_id          | B05006
table_seq_num     | 0010
table_title       | PLACE OF BIRTH FOR THE FOREIGN-BORN POPULATION IN THE UNITED STATES
table_subject_area | Foreign Birth
table_universe    | Universe: Foreign-born population excluding population born at sea
#--

```

```

-- What does the table_id for table 60 mean ?
https://ask.census.gov/faq.php?id=5000&faqId=1687

```

```

B05006 =>
  B -> base table
  05 -> Foreign Born; Citizenship; Year or Entry; Nativity
  006 -> sixth table in this category

```

```

-- display table 60
  emit .csv header from acs20125_table_cols_list; emit .csv from acs_mtable_60_raw

```

```

-- what are the columns of table 60 ?

```

```

SELECT table_col, table_col_orig
FROM acs20125_table_cols_list
where
table_indb_name = 'acs_mtable_60_raw'
order by pkey;

```

```

-- what are the locations for each row of acs_mtable_60_raw

```

```

SELECT
  a.pkey as mtable_60_pkey,
  g.geoid,
  g.geo_name
FROM
  public.acs20125_seq_metadata,
  public.acs20125_geo_defs_ca g,
  public.acs_mtable_60_raw_ca a
WHERE
  g.logrecno = acs20125_seq_metadata.e_logrecno AND
  a.seq_metadata_key = acs20125_seq_metadata.pkey
ORDER BY
  a.pkey;

```

```

#--
mtable_60_pkey; geoid; geo_name
1355322;"04000US06";"California"
1355332;"04001US06";"California -- Urban"
1355342;"04043US06";"California -- Rural"
1355352;"040A0US06";"California -- In metropolitan or micropolitan statistical area"
1355362;"040C0US06";"California -- In metropolitan statistical area"
1355392;"040E0US06";"California -- In micropolitan statistical area"
1355422;"040G0US06";"California -- Not in metropolitan or micropolitan statistical area"
1355432;"040H0US06";"California -- Not in metropolitan statistical area"
1355442;"05000US06001";"Alameda County, California"
1355452;"05000US06003";"Alpine County, California"
1355762;"05000US06065";"Riverside County, California"
1355772;"05000US06067";"Sacramento County, California"
1355782;"05000US06069";"San Benito County, California"
1356002;"05000US06113";"Yolo County, California"
1356012;"05000US06115";"Yuba County, California"
1356022;"06000US0600190020";"Alameda CCD, Alameda County, California"
1356032;"06000US0600190200";"Berkeley CCD, Alameda County, California"
1356042;"06000US0600191070";"Fremont CCD, Alameda County, California"
1356052;"06000US0600191260";"Hayward CCD, Alameda County, California"
1356062;"06000US0600191660";"Livermore-Pleasanton CCD, Alameda County, California"
1356072;"06000US0600192230";"Oakland CCD, Alameda County, California"
1356082;"06000US0600391890";"Markleeville CCD, Alpine County, California"
...
1358782;"06000US0607992860";"San Luis Obispo CCD, San Luis Obispo County, California"
1358792;"06000US0608191215";"Half Moon Bay CCD, San Mateo County, California"
1358802;"06000US0608192870";"San Mateo CCD, San Mateo County, California"
1358812;"06000US0608193170";"South San Francisco CCD, San Mateo County, California"
1358822;"06000US0608390378";"Carpinteria CCD, Santa Barbara County, California"

```

-- What are the population values, table 60, for any county, labelled ?

```
SELECT
  a.pkey as mtable_60_pkey,
  g.geoid,
  g.geo_name,
a.mtbl_60_col_1 as "Europe",
a.mtbl_60_col_2 as " Northern Europe",
a.mtbl_60_col_12 as " Western Europe",
a.mtbl_60_col_20 as " Southern Europe",
a.mtbl_60_col_26 as " Eastern Europe",

a.mtbl_60_col_45 as "Asia",
a.mtbl_60_col_46 as " Eastern Asia",
a.mtbl_60_col_54 as " South Central Asia",
a.mtbl_60_col_65 as " South Eastern Asia",
a.mtbl_60_col_76 as " Western Asia",

a.mtbl_60_col_89 as "Africa",
a.mtbl_60_col_90 as " Eastern Africa",
a.mtbl_60_col_95 as " Middle Africa",
a.mtbl_60_col_98 as " Northern Africa",
a.mtbl_60_col_104 as " South Africa",

a.mtbl_60_col_114 as "Oceania",

a.mtbl_60_col_120 as "Americas",
a.mtbl_60_col_145 as " South America",
a.mtbl_60_col_121 as " Latin America",
a.mtbl_60_col_122 as " Caribbean",

a.mtbl_60_col_158 as "Canada"

FROM
  public.acs20125_seq_metadata,
  public.acs20125_geo_defs_ca g,
  public.acs_mtable_60_raw_ca a
WHERE
  g.logrecno = acs20125_seq_metadata.e_logrecno AND
  a.seq_metadata_key = acs20125_seq_metadata.pkey AND
  g.geoid ~ '^05000US06'

ORDER BY
  g.geo_name;
```

```
-[ RECORD 1 ]-----+-----
mtable_60_pkey | 7013303
geoid          | 05000US06001
geo_name       | Alameda County, California
Europe        | 29836
 Northern Europe | 7220
 Western Europe  | 8037
 Southern Europe | 5450
 Eastern Europe  | 9047
Asia          | 271146
 Eastern Asia   | 103636
 South Central Asia | 65314
 South Eastern Asia | 96637
 Western Asia   | 5195
Africa        | 10276
 Eastern Africa | 4843
 Middle Africa  | 478
 Northern Africa | 1873
 South Africa   | 413
Oceania       | 7286
Americas      | 145345
 South America  | 10378
 Latin America  | 140289
 Caribbean     | 2386
Canada        | 5026
...
```

-- What are the population values, table 60, for Alameda County ?

```

SELECT
  a.pkey as mtable_60_pkey,
  g.geoid,
  g.geo_name,
  a.*
FROM
  public.acs20125_seq_metadata,
  public.acs20125_geo_defs_ca g,
  public.acs_mtable_60_raw_ca a
WHERE
  g.logrecno = acs20125_seq_metadata.e_logrecno AND
  a.seq_metadata_key = acs20125_seq_metadata.pkey AND
  g.geoid = '05000US06001'

ORDER BY
  a.pkey;

```

7013303		05000US06001		Alameda County, California		7013303		709942		463889		
29836				5349		2453		2569		327		921
226		244		372		108		8037		442		156
2234		3776				352		104		5450		621
1390		2671				81		9047		83		404
135		200		39		402		249		142		34
125		1028		1264		2329		969		632		80
576		82		271146		103636		85550		61181		11285
13084		4916		12440		30		65314		5751		629
49713		3497		101				3399		623		212
464		96637		3028				1863		1026		3073
53480		686		2749		288		31		5195		308
1106		329		254		750				254		224
1873		970		334		96				473		445
413		32		1960		0		180		246		
1169		72		293		677		7286		971		
555		416		4705		1610		145345		140289		
2386		48		47		505		22		242		
26		84		815		0		439		78		
80		127525		103681		364		250		12030		
5922		1579		3325		328		46		10378		
859		381		2121		637		1281		604		
327		3180		73		649		266		5056		
5026		30		+ ca								