

California Broadband Task Force

Appendix: Broadband Mapping by Speed Methodology

Introduction

The CBTF confronted several significant challenges in developing wireline broadband availability speed maps for California. These included:

- Accurately portraying broadband availability and unavailability
- Preserving provider competitive advantage and confidentiality
- Coordinating provider data through a neutral third party
- Agreeing on a mapping protocol (including the most appropriate scale for data collection, analysis and display – three different issues).
- Agreeing on a speed tier protocol

The Build-Out Working Group to the CBTF and staff described an initial mapping protocol that was presented to the CBTF on May 24. Michael Byrne, GIS Architect for the Office of Statewide Health Planning and Development and staff to the CBTF refined this protocol to meet the needs of wireline broadband providers and the CBTF. The California Emerging Technology Fund adopted the resulting mapping protocol for its project with the third party.

The intent of this mapping protocol was to provide the CBTF, Governor and Legislature with the most comprehensive and accurate assemblage of broadband availability in the State. This effort was accomplished by integrating provider data on speed and availability from the address level.

Supplied Data Specification

Wireline broadband providers submitted to a third party location-based reference(s) (e.g., discrete addresses or map-based service area delineations) for available broadband services. Each was coded by the highest available speed tier offered. Each speed tier represented a combined upstream and downstream speed. The tiers submitted were:

- 500 Kbps to 1 Mbps
- 1 Mbps to 5 Mbps
- 5 Mbps to 10 Mbps
- 10 Mbps to 100 Mbps
- 100 Mbps to 1 Gbps
- 1 Gbps to 10 Gbps

Providers delivered the above data to the third party, Michael Baker Corp. (Baker) in the following format alternatives:

Provider Alternative 1 – Preferred Format for Addresses:

A list of all addresses, in a parsed-address field format, with available broadband within the provider’s service area:

Provider Alternative 1 Š Address Record Format													
Parsed -address													
ID	House Number	Prefix Direction	Street Name	Street Type	Suffix Direction	City	County	Exchange * (Telco only)	ZIP Code	CBTF Speed	Latitude*	Longitude*	Census*
1	818	S	K	St	NW	Sacramento	Sacramento	Sacramento	95814	3	34.386543	-119.7653	60670100111001

Provider Alternative 2 – Secondary Format Preferences for Addresses:

A list of all addresses, in a concatenated-address field format, with available broadband within the provider’s service area:

Provider Alternative 2 Š Address Record Format									
Concatenated -address									
ID	Address	City	County	Exchange * (Telco only)	ZIP Code	CBTF Speed	Latitude*	Longitude*	Census*
1	818 SK St NW	Sacramento	Sacramento	Sacramento	95814	3	34.386543	-119.7653	60670100111001

- Notes Associated with Provider Alternative 1 or 2:
 - All non-* fields (ID, Address (parsed or concatenated), City, ZIP_Code, & CBTF_Speed) are required.
 - Fields with an * (Latitude, Longitude, Census) are optional if available.
 - If submitting Latitude and Longitude, Decimal Degrees (as shown above) is preferred over Degrees Minutes Seconds (e.g. - 119 47 13.83)
 - Census will be used (if populated) to assist with the mapping. If providing Census please describe what census codes you are submitting. The preferred census geography is the 15 digit block code (State + County + Tract + Block with leading zero’s).
 - The ID Field is a unique counting number (e.g. record number from 1 ... n)
 - For CBTF_Speed please submit speed tier (e.g. 1 – 5) or raw speed numbers and the vendor will translate for you.
 - A tab-delimited transferable digital table (e.g. tab separated values) with any associated notes and a contact person in a readme.txt file will be accepted.

Provider Alternative 3 – Map-based Service Area Delineations:

A GIS or CADD data file (an ESRI shapefile or personal geodatabase, or Autodesk AutoCAD DWG file, or Bentley Microstation DGN file) with available broadband within the provider’s service area only if such areas are delineated by CBTF_Speed as city blocks or smaller areas. The intent of this

alternative is to permit providers that maintain such broadband availability data as map-based representations an opportunity to provide such information as map-based representations in lieu of a list of all addresses.

Provider Alternative 3 S Map -based Service Area Delineations						Record Format
ID	City	County	Exchange* (Telco only)	ZIP Code	CBTF Speed	Census*
1	Sacramento	Sacramento	Sacramento	95814	3	60670100111001

- Notes Associated with Provider Alternative 3:
 - All areas must be represented as closed polygons with a single unique id.
 - Any variation of a record item, City, County, Exchange (if provided), Zip Code, CBTF_Speed or Census) must result in a separate formed and closed polygon.
 - All areas formed and contained within a closed polygon must have available broadband service within 50’ from the inside edge of the surface to which the polygon’s area is delineated.
 - All non-* fields (ID, City, County, ZIP_Code, & CBTF_Speed) are required within an associated database record accompanying the GIS or CADD data.
 - Fields with an * (Exchange, Census) are optional if available.
 - Decimal Degrees (as shown above) is preferred over Degrees Minutes Seconds (e.g. -119 47 13.83)
 - Albers Equal Area projection
 - Census will be used (if populated) to assist with the mapping. If providing Census please describe what census codes you are submitting. The preferred census geography is the 15 digit block code (State + County + Tract + Block with leading zero’s).
 - The ID Field is a unique counting number (e.g. record number from 1 ... n)
 - For CBTF_Speed please submit speed tier (e.g. 1 – 5) or raw speed numbers and the vendor will translate for you.
 - If a database table is provided for attribute information associated with the GIS or CADD file, please specify its format and schema or deliver the accompanying database records as a tab-delimited transferable digital table (e.g. tab separated values) with any associated notes and a contact person in a readme.txt file will be accepted.

Technical Workflow

The technical workflow used to process the California broadband availability data included seven steps:

- Data Submission Agreements
- Register / describe data,

- Pre-format data,
- Load event data,
- Load and geocode data,
- Determine data quality,
- Apply aggregate function,
- Provide final deliverables

Final Geographic Information Systems (GIS) data of availability by tier was delivered to the California Emerging Technology Fund and subsequently to the Office of Statewide Health Planning and Development (OSHPD). OSHPD then performed post-processing analysis to determine statewide and regional availability metrics and maps. This workflow included these steps:

- Metric Analysis
- Develop final maps for reporting

Data Submission Agreements

- a) Baker entered into individual Non-Disclosure Agreements (NDA) with each submitting wireline provider that wished to sign an NDA. These agreements typically stipulated full confidentiality of raw data.
- b) Once NDA's were signed, providers submitted data in the appropriate formats to Baker.
- c) After data submission, Baker analysts processed the data anonymously (i.e. they did not know which provider's data they were processing at any point).
- d) Once a final product was developed, all raw input data was destroyed (as/if required by NDA).

Register / Describe Data

- a. Receive a digital file from each provider (Provider Alternatives 1-3 above)
- b. Register metadata about the file (e.g. contact, file name, date received etc). Registered information was only used for project management and processing analysts were not provided with registration data..
- c. Ensure file format meets specifications outlined in the Task Force Request (check for complete required fields, populated optional fields, and determination of CBTF_Speed population – if the speed or speed scale is populated). If data is not in required format, determine if submitted information can still be used. If so, manipulate data into required format. If not, contact CBTF project staff and provider and obtain a solution
- d. Send receipt of data submitted to provider.

Pre-format

- a. Create three loading tables (using the Raw Data Submission Table as a format base); 1) for source data (tblBBTF_source), 2) for already geocoded data (e.g. tblBBTF_geocoded_event) and 3) for data not yet geocoded (e.g. tbl_BBTF_address).
- b. Determine if any data has come in with already geocoded values
- c. If so, split geocoded records out and append them into the event loading table (tblBBTF_geocoded_event). No tag regarding vendor, source file, or technology will be maintained at this point. Then go to step 3
- d. If not, load (append) non-geocoded data into the tbl_BBTF_address table. No tag regarding vendor, source file, or technology will be maintained. Proceed to step 4.

Load Event Data

- a. Generate point GIS file from already geocoded data based on the Latitude and Longitude delivered.
- b. Project resulting dataset to California standard Albers Equal Area projection.
- c. Evaluate census geography delivered for any already geocoded data. Match resulting points against framework census geography. Provide statistics of common match between reported census geography and matched census geography in the 50% deliverable.

Load and Geocode data

- a. Create a composite geocoding solution with the following hierarchical services
 - i. Returns a Service Match Name (e.g. which part in the geocode service did it match – Address Points etc), Status (e.g. Match, Tied, Unmatched), Match Score, Side, X, Y.
 - ii. Returns point data with industry standard match score (to be determined between vendor and CBTF Staff) in the California standard Albers Equal Area Projection
- b. Geocode this table against composite geocoding service

Load and Process Map-based Service Area Data

- a. Project map-based data if necessary to California standard Albers Equal Area Projection
- b. Intersect map-based polygon data with other acceptable data sources if attribution is incomplete (City, County, Exchange, Zip Code, Census)
- c. Aggregate provider data by common properties to form homogenous representations of specific spatial extents.
- d. Process data to ESRI Grid (Raster) for analysis with geocoded point data and aggregation as described below.

Determine Data Quality

- a. The contractor regularly analyzed the following statistics;
 - iii. Number and percent provided as already geocoded
 - iv. Number and percent geocoded with a Match Score higher than determined industry standard (from 3.a.v above).
 - v. Number and percent tied geocodes.
 - vi. Number and percent unmatched
 - vii. Breakdown of City / Zip (number and percent) of unmatched geocodes

Aggregation and Analysis:

The vendor incorporated all data from providers and aggregated this data into a single availability dataset. This aggregation included the merging of geocoded and polygonal data from providers.

Apply Aggregate Function

- a. Append points from resulting event and geocode match into one point dataset
- b. Perform a data conversion to move availability points to a raster grid format. This processing was accomplished with a simple 1000 meter point to grid function; each point / speed combination resulted in a 1000 meter cell at the highest available speed level only. The final grid also was filtered with a 3x3 grid filter to smooth the edges and to reduce potential breaches of confidentiality.

Metric Analysis

A final dataset was delivered to the California Emerging Technology Fund and the Office of Statewide Health Planning and Development (OSHPD) through the CBTF. OSHPD then developed a metric to determine ‘percent’ availability and the number of communities unserved. This metric is based on the number of housing units inside and outside the final availability map. Housing units were used as a base metric for several reasons including 1) this map is a wire-line availability map (e.g. to households and businesses) 2) broadband as an infrastructure is essentially a personal or business utility, 3) the output availability map is a self forming geography (e.g. not based on any other geographic bounds) and therefore cannot be analyzed against other demographic boundary data without significantly introducing error 4) simply summing the area of availability does not provide a useable metric and 5) other large states with significant rural populations have provided a similar analysis approach (see http://gis.esri.com/library/userconf/proc06/papers/papers/pap_1436.pdf).

To develop the structures data set, OSHPD used two primary sources: 1) Tele Atlas address point data and 2) US Census data of number of Housing Units.

Tele Atlas address points are a third party solution integrating information from hundreds of private and government sources. These points represent real parcel addresses (real address locations). Currently this dataset has over 6 million real building address points in California.

The US Census reports over 12 million housing units in California. Using the number of housing units per US Census Block Group (provided by the California Department of Forestry and Fire Protection – see <http://frap.cdf.ca.gov/data/frapgisdata/select.asp>) and augmented with Claritas produced number of housing units per Census Tract, OSHPD subtracted the number of existing Tele Atlas points and distributed the remaining number of housing units along roads within the Census Block Group. The housing unit distribution process works on the principle that the distribution of units is a direct function of the distribution of roads; the denser the roads, the denser the houses. Given this principle, OSHPD randomly selected road segments and randomly distributed houses at locations within the real address ranges of the road segments. This method resulted in a distribution of over 6 million additional housing units in the State. For a copy of the code base to create these structures, please contact Michael Byrne at mbyrne@oshpd.ca.gov.

To assemble the metric OSHPD performed the following:

- Assembled both Tele Atlas and OSHPD generated points into one dataset. This dataset is now the surrogate for housing units within the state.
- Coded all houses as Rural, or Urban based on OSHPD Medical Service Study Areas (see <http://gis.ca.gov/catalog/BrowseCatalog.epl?id=1044>)
- Coded all houses for which County it was in
- Coded all houses for which Broadband Region it was in (defined by the CBTF)
- Coded all houses for broadband availability (or not) based on the delivered availability mapping protocol described above
- Coded broadband available houses for the broadband speed available

		Rural	Urban	Total
Statewide Broadband Availability				

Availability per Region				
1	Northern Sierra			
2	North Coast			
3	Sacramento Valley			
4	Sacramento Metropolitan			
5	Bay Area			
6	Mother Lode			
7	San Joaquin Valley			
8	East Side			
9	Central Coast			
10	Los Angeles / Orange			
11	Inland Empire			

12	Southern Border			
----	-----------------	--	--	--