New Jersey Broadband Mapping Project:

Community Anchor Institution Processing

September 30, 2011

Grantee:

New Jersey Office of Information Technology

200/300 Riverview Plaza

PO Box 212

Trenton, NJ 08625

Contact:

Shelley Bates

shelley.bates@oit.state.nj.us

609-633-9605

Contractor:

Telcordia Technologies, Inc.

1 Telcordia Drive

Piscataway, NJ 08854

Contact:

John R. Wullert, II

jwullert@telcordia.com

732-699-2687

Table of Contents

SUMMARY	**1**
LOCAL GOVERNMENT AND NON-GOVERNMENT ORGANIZATIONS	**2**
STATE GOVERNMENT	**2**
HOSPITALS	**3**
HIGHER EDUCATION	**4**
LIBRARIES	**5**
PRIVATE K-12 SCHOOLS	**6**
PUBLIC K-12 SCHOOLS	**7**
PUBLIC SAFETY ORGANIZATIONS	**9**
CAI LOAD PROCESSING	**10**

Summary

For each category of community anchor institution, we generally obtained data from two sources. One source was a reference source that provided a list of institutions with name, address and ID number where applicable. This reference source was expected to be nearly complete, representing all the institutions of the specified type in the state. The other source provided the broadband information. In most cases, the broadband information was supplied by the institutions via our Web site.

There were exceptions, however, to these guidelines. In the case of Higher Education, we obtained the broadband access information from NJEdge, an organization that collects data via its own survey. In the case of State Government, we obtained a list of broadband circuits provided to the state by Verizon; there was no reference list for comparison. We similarly had no reference list for local government and non-governmental organizations; we used only data from collected via our Web site for these classes of institution.

For each CAI category, t he following table provides the number of records we obtained from the reference source, the number of broadband access records we obtained, the total number of records we submitted to the NTIA and the number of complete records, with verified address information and broadband access information.

CAI Category	Reference Records	Broadband Records	Total Records Submitted	Complete Records Submitted
School K-12 (Public)	2603	796 (Web) 478 (eRate)	2598	175
School K-12 (Private)	1260 (NCES)		1267	169
Libraries	465 (IMLS)	89	472	50
Medical/Healthcare	1107 (NJ-HHS)	5	1108	5
Public Safety	343 (NJ 911 Comm.)	120	349	104
University	158 (NCES IPEDS)	39 (NJEdge)	158	39
Other – State Government		2007	1947	1947
Other – Local Government	0	54	54	54
Other – Non Government	0	8	8	8
Total			6964	2551

Local Government and Non-Government Organizations

1. Accepted data submitted by 54 local government and 8 non-governmental organizations via specially designed Web site. We merged data submitted to Web site for April 2011 delivery with that submitted between April and September. (Files lib_20110323.xml and lib_20110907.xml) Data collected included:
    i. Community Anchor Institution Category
    ii. Community Anchor Institution Name (System, Branch)
    iii. Address: Street, City, State, Zip, County
    iv. Contact info: Name, Phone, Email, Web address
    v. Wi-Fi access
    vi. Broadband info: Provider, Technology, Upstream and Downstream speeds
    vii. Comment
2. Generated Latitude and Longitude via geo-coding using Yahoo geocoder API.
    a. Ensured no errors were present, that at least on entry was returned and that quality metric was over 75. Also ensured that result was in New Jersey and that city and zip were not both blank.
    Output is in file Submitted_GovNGO_CAIs.xls.
3. Manually edited results to add street numbers in those places where it was missing via Google search for the institution name.

State Government

1. Obtained a listing of 2007 connections provided by the primary broadband service provider to the state. List of connections included the following data:
    a. Service address
        i. This field included an indication of the office or department being served and an extremely abbreviated version of the address
        ii. e.g.: "(SPNL)STATE OF NJ-TLS 19 LANDIS AV, UP DRFLD T"
    b. Speed (single value, 1.5 to 1000 Mbps)
    c. Technology (ATM, Ethernet, Frame Relay, PRI, Point-to-Point
2. Used an automated process to expand the town names in the Service Address field (flow for steps 2-6 is in file VerizonList_Geocode.arroyo; input file is Broadband Mapping Prod Sum 2500 Feb 11_Addressed_Ida_Murray3.xlsx)
    a. For example, replaced "PRSPY" with "Parsippany" and "FR LN" with "Fair Lawn"
3. Extracted address information from Service Address field by removing the following:
    a. Digits following and including a pound sign (e.g., NJ STATE PAROLE DIST **#6** 210 S BROAD)
    b. "P.O Box NNNN",
    c. Anything in parentheses (e.g., **(SPNL)**STATE OF NJ:OIT 90 STATE HWY NO 183)

    d. Any string consisting solely of letters, backslashes, colons, dashes, ampersands and spaces prior to the first number string in the address (e.g., **SONJ:DOE** 7 GLENWOOD AV, E O BLDG FLR 4;DES SUITE 401-402)
    e. Any string after the first comma (e.g., 7 GLENWOOD AV, **E O BLDG FLR 4;DES SUITE 401-402**
    f. Text prior to and including an ampersand (e.g., **NJ STATE DOT @** ROUTE 23)
    g. Replacing "AV," with "AVE,"
    h. Any text between commas (e.g., 3810 NEW JERSEY AV, **WILD DES DEPT LABOR,**)
4. Extracted CAI Name information using the following rules:
    a. Extract text between four characters in parentheses and the last number to appear
    b. Remove any number and text following it
    c. Remove any ampersand and the text following it
    d. Replace empty strings with SONJ
5. Merged city information and state information with extracted addresses.
6. Generated Latitude and Longitude via geo-coding using Yahoo geocoder API.
    a. Ensured no errors were present, that at least on entry was returned
    b. Ensured that state was New Jersey and that city and state values were populated.
7. For those that failed test with Yahoo geocoder API, attempted to match with Google geocoder API
    a. Ensured no errors were present, that at least on entry was returned
    b. Ensured that state was New Jersey and that city and state values were populated.
8. Results in successful geocoding of 1917 of the 2007 entries
    a. Results are in file NJ_State_Verizon_Geocoded.xls

Hospitals

1. Obtained a listing of 1107 hospitals from NJ Department of Health and Human Services . List of connections included the following data:
    a. Facility Name
    b. Address: Street, City, State, Zip
2. Generated Latitude and Longitude via geo-coding using Google geocoder API.
    a. Ensured that at least on entry was returned, that state was New Jersey and that city or zip were present in recognized address. (Used Arroyo flow HHS_Hospital_Process.arroyo)
    Output of this stage is in file Hospitals_Geocoded2.csv.
3. Merged NJ-HHS data with data collected from 5 hospitals via our hosted Web site to merge address and ID information with speed and Wi-Fi availability information. We merged data submitted to Web site for April 2011 delivery with that submitted between April and September. (Files lib_20110323.xml and lib_20110907.xml)
    a. Performed exact match between NJ-HHS and submitted data on institution name

i. Facilitated matching by Converting names to upper case, removing certain common words (THE, HOSPITAL, MEDICAL, CENTER, SYSTEM, HEALTHCARE), removing double spaces and trimming leading and trailing spaces.

This portion of the process occurs in SubmittedCAI_Hospital_Process.arroyo. Output is in file Hosp_Submitted_Matched.xls.

4. Manually matched inserted last hospital into list.

Higher Education

1. Obtained the following data from the named sources
    a. List of higher education institutions from National Center for Education Statistics IPEDS Data Center (http://nces.ed.gov/collegenavigator/?s=NJ). Table included information on 158 institutions with the following fields:
        i. Institution Name
        ii. Address: Street, City, County, State, ZIP
        iii. IPEDS ID

        Final input data, including a few manual edits (see below) is in file CollegeNavigator_NJ_20110909_fixed2.xlsx

    b. Generated Latitude and Longitude via geo-coding using Yahoo geocoder API (flow IPEDS_HigherEd_Geocode.arroyo).
        i. Ensured no errors were present, that at least on entry was returned
        ii. Ensured that state was New Jersey and that city and state values were populated.
    c. For those that failed test with Yahoo geocoder API, attempted to match with Google geocoder API (Flow IPEDS_HigherEd_Geocode.arroyo)
        i. Ensured no errors were present, that at least on entry was returned
        ii. Ensured that state was New Jersey and that city and state values were populated.
    d. Manually updated a few addresses that failed to produce maps. Result was that 156 of 158 institutions were properly geocoded.
2. List of members of NJEdge (Format-edited version is in file Mapping Bandwidth_Mb_07112011_edit.xlsx). Table included information on 50 institutions, most of which (39) were unique state, community or private institutions of higher learning. Information from NJEdge included:
        i. Institution Name
        ii. Address
        iii. Technology Type
        iv. Upstream and downstream speeds
3. Merged IPEDS and NJEdge data to match institution data with broadband access information
    a. Performed exact match on institution name

i. Facilitated matching by Converting names to upper case and trimming excess spaces

b. Of those NJEdge data entries that did not match, used approximate matching based on institution name
   i. Preprocess prior to approximate match involved
      1. Removing strings COLLEGE, UNIVERSITY, NEW JERSEY
      2. Removing any punctuation
   ii. Matched using Levenshtein Distance metric with threshold of 4.

c. Reviewed unmatched NJEdge data manually and identified three additional matches.

4. Successfully merged data from all 39 NJEdge institutions into IPEDS data for total of 158 institutions
   a. Note that remaining NJEDGE institution (Fairleigh Dickenson) has different address than either of the campuses in the IPEDS data.

   Final output is in file HigherEd_Geocoded_RateMatched.xls
5. Manually edited results to add street numbers in those places where it was missing via Google search for the institution name.

Libraries

1. Obtained the following data from the named sources
   a. Obtained the file "Public Libraries Survey Fiscal Year 2009" from http://harvester.census.gov/imls/data/pls/index.asp. Used file puout09.txt
      i. Manually extracted 465 records for the state of New Jersey
      ii. Used the following data items:
         1. FSCSKEY
         2. FSCS_SEQ
         3. LIBNAME
         4. ADDRESS
         5. CITY
         6. ZIP
         7. LATITUDE
         8. LONGITUDE
   b. Data submitted by 89 library organizations via specially designed Web site. Data collected included same fields listed above for Local Governmental organizations
2. Merged library survey data with data collected from libraries via our hosted Web site to merge address and ID information with speed and Wi-Fi availability information.
   a. Performed exact match between survey and submitted data on library name
      i. Facilitated matching by Converting library names to upper case, cutting submitted names to fixed-field length of survey data (60 characters) and trimming excess spaces
   b. For those submitted data entries that did not match, performed an approximate match based on library name

i. Preprocess prior to approximate match involved
    1. Removing strings “P.L.”, “FREE”, “PUBLIC”, “LIBRARY”, TOWNSHIP, TSWP, PUB, LIB, THE, SYSTEM
    2. Removing any punctuation
    3. Converting “NO”/”SO” at start of line to NORTH and SOUTH respectively
ii. Matched using Levenshtein Distance metric with threshold of 3.

c. Successfully matched all but twelve submitted entries to Library Survey Data
    i. Manual comparison showed that two of those libraries were not present in the survey data.
    ii. Remaining ten were branches of Newark Public Library, but all were submitted with the same address, so they could not be successfully geocoded.

Results (LibraryPlusSubmitted.xls) include 472 Library entries. This is larger than the 465 from the survey because some libraries submitted more than one broadband provider.

3. Manually edited results to add street numbers in those places where it was missing via Google search for the institution name.
    a. During manual edit, also removed entries that were labeled “bookmobile” as not indicating a valid anchor institution able to receive broadband.

Private K-12 Schools

1. Obtained the following data from the named sources:
    a. List of private K-12 education institutions from National Center for Education Statistics Private School Universe Survey (http://nces.ed.gov/surveys/pss/pssdata.asp). Table included information on 1260 institutions with the following fields:
        i. Name
        ii. Address: Street, City, State, ZIP
        iii. NCES_ID
        iv. Latitude/Longitude
    b. Data submitted by schools via specially designed Web site. Data collected included same fields listed above for Local Governmental organizations. Total number of Public and Private schools submitting information was 796.
    c. Data from the USAC eRate program, listing schools that have obtained subsidized Internet access, including following relevant fields
        i. Name
        ii. Address: Street, City, State, ZIP
        iii. Provider

    There were 478 records that corresponded to schools and Internet access.
2. Merged NCES private school with data collected from private schools via our hosted Web site to merge address and ID information with speed information.
    a. Performed exact match between NCES and submitted data on institution name and zip code
        i. Facilitated matching by:

1. Converting school names to upper case
2. Removing string ", NJ"
3. Converting string SAINT to ST

b. For those submitted data entries that did not match NCES data, performed an approximate match based on institution name
   i. Preprocess prior to approximate match involved
      1. Replacing string SCHOO or SCHO with SCHOOL
      2. Replacing string "HIGH SCHOOL" with HS and string "ELEMENTARY" with ELEM
      3. Removing strings SCHOOL, THE, REGIONAL, HIGH and ACADEMY
      4. Trimming excess spaces
   ii. Matched using Levenshtein Distance metric with threshold of 3.

c. Successfully merged data from 71 submitted private school into 1260 NCES institutions
   i. Manual comparison resulted in matching of two additional institutions
   ii. Remaining institutions were ambiguous or not present in the NCES data.

3. Combined results of step 2 with eRate data to merge address and ID information with access and provider data. (Flow in file K-12_eRateProcess.arroyo, handles both public and private schools)
   a. Performed exact match between step-2 results and eRate data on institution name and zip code
   b. Verified uniqueness of results based on institution name, zip code and provider
   c. When a match was detected, set the Availability flag to "y" and filled in provider name from eRate data. (Unless provider name was already present from Web-submitted data)
   d. Filled in an 128 additional records
4. Generated 1267 records to submit. Note that some schools had more than one service provider and thus include multiple records.
   a. Output file is PrivateSchool_GeoMatched.xls
5. Manually edited results to add street numbers in those places where it was missing via Google search for the institution name. Note that not all schools have a street number – some list an intersection and others simply list the street as their address.

Public K-12 Schools

1. Obtained the following data from the named sources:
   a. List of public K-12 education institutions from National Center for Education Statistics Public School Universe Survey. (Went to http://nces.ed.gov/ccd/schoolsearch/ , searched for schools in New Jersey, then selected option at bottom of results page to download an Excel file: ncesdata_DE2476A3.xls.) Table included information on 2603 institutions with the following fields:
      i. Name
      ii. Address: Street, City, State, ZIP

iii. NCES_ID

b. Data submitted by schools via specially designed Web site. This was entries in the school category that did not match any of the NCES private schools. Total number of Public and Private schools submitting information was 796. Of those, 673 did not match private schools.

c. Data from the USAC eRate program, listing schools that have obtained subsidized Internet access, including following relevant fields
   i. Name
   ii. Address: Street, City, State, ZIP
   iii. Provider

   There were 478 records that corresponded to schools and Internet access.

2. Merged NCES private school with data collected from private schools via our hosted Web site to merge address and ID information with speed information. (Flow in file PublicK-12Process.arroyo)
   a. Performed exact match between NCES and submitted data on institution name and zip code
      i. Facilitated matching by:
         1. Removing SCHOOL and all truncated versions of the word from the ends of any string
         2. Performing the following conversions
            a. "SENIOR HIGH" and HIGH to HS
            b. "MIDDLE", "M S", "MID" and "MIDD" to MS
            c. "ELEMENTARY" to ELEM
            d. CHARTER to CS
            e. BOROUGH to BORO
            f. AVENUE to AVE
            g. TOWNSHIP to TWP
            h. STREET to ST
         3. Removing the strings REGIONAL, " REG" and ACADEMY
         4. Removing punctuation and double spaces
         5. Trimming any leading or trailing spaces
   b. For those submitted data entries that did not match NCES data, performed an approximate match based on concatenation of institution name and zip code
      i. Preprocess prior to approximate match involved
         1. Removing the following phrases
            a. "BOARD OF EDUCATION" and all truncated versions
            b. BOE
            c. DISTRICT and all truncated versions
            d. PRIMARY, INTERMEDIATE, ELEM, MS, HS, SR, JR
            e. # or any digits
            f. PUBLIC
         2. Trimming excess spaces

3. Submitted entries that were blank after these operations were removed.

ii. Matched using Levenshtein Distance metric with threshold of 2.

c. Successfully merged data from 169 submitted entries into 2595 NCES institutions

i. Dropped 8 NCES institutions as incomplete

ii. Recurring issue was information submitted for districts that did not correspond to a specific school

3. Combined results of step 2 with eRate data to merge address and ID information with access and provider data. (Flow in file K-12_eRateProcess.arroyo, handles both public and private schools)
    a. Performed exact match between step-2 results and eRate data on institution name and zip code
    b. Verified uniqueness of results based on institution name, zip code and provider
    c. When a match was detected, set the Availability flag to "y" and filled in provider name from eRate data. (Unless provider name was already present from Web-submitted data)
    d. Filled in nine additional records
4. Generated Latitude and Longitude via geo-coding using Yahoo geocoder API.
    a. Ensured no errors were present, that at least on entry was returned and that quality metric was over 75.
    b. Ensured that state was New Jersey and that city and/or zip value was populated.
5. Generated 2598 records to submit. Note that some schools had more than one service provider and thus include multiple records.
    a. Output file is PublicSchool_GeoMatched.xls
6. Manually edited results to add street numbers in those places where it was missing via Google search for the institution name. This was only attempted for a portion of those with missing street numbers. Of those attempted, only 25% were producing a valid street number, so the manual step was not deemed worthwhile.

Public Safety Organizations

1. Obtained the following data from the named sources:
    a. List of local and state public safety organizations obtained from NJ State 911 Commission. (Reused data from April 2011 - PSAP's & PSDP's_Geocoded.xls) Table included information on 343 institutions with the following fields:
        i. Name
        ii. Address: Street, City, State, ZIP, County
        iii. NCES_ID
    b. Data submitted by 120 public safety organizations via specially designed Web site. Data collected included same fields listed above for Local Governmental organizations
2. Generated on 911 Commission Data Latitude and Longitude via geo-coding using Yahoo geocoder API.

    a. Ensured no errors were present, that at least on entry was returned and that quality metric was over 75.
3. Merged 911 Commission data with PSAP data collected from via our hosted Web site (99 entries) to merge address and ID information with speed information.
    a. Performed exact match between 911 and submitted data on institution name
        i. Facilitated matching by:
            1. Converting names to upper case
            2. Removing the Strings DEPARTMENT, DEPT, TOWNSHIP, TWP
            3. Removing punctuation and double-spaces
            4. Replacing string PD with POLICE and string BOROUGH with BORO
    b. Performed manual merging to integrate additional submitted records that were not matched.
        i. Successfully merged 104 submitted PSAP entries with 911 Commission data. Output in file PSAP_911_Matched.xls
4. Manually edited results to add street numbers in those places where it was missing via Google search for the institution name.

CAI Load Processing

Submission date: October 2011

This report presents details on processing data about Community Anchor Institutions for delivery to the National Telecommunications and Information Administration.

For each location submitted to us we report a street address, details about the data service at the Institution (if known), a point shape corresponding to the street address, and the ID of the enclosing Year 2010 Census Block.

Overview of Transfer Model Table

The following table lists the columns in the NTIA data-transfer table.

Table Column	Null?	Data Source / Transformation
ANCHORNAME	NO	
ADDRESS	NO	
BLDGNBR	NO	
PREDIR	YES	Usually set to null

STREETNAME	NO	
STREETTYPE	YES	Usually set to null
SUFFDIR	YES	Usually set to null
CITY	NO	
STATECODE	NO	Set to "NJ"
ZIP5	NO	
ZIP4	YES	Set to null
LATITUDE	NO	Submitted or geocoded from street address
LONGITUDE	NO	Submitted or geocoded from street address
CAICAT	NO	Set to appropriate category
BBSERVICE	NO	
PUBLICWIFI	NO	
URL	YES	
TRANSTECH	YES	Note: set to “0” if unknown.
FULLFIPSID	NO	ID of enclosing Year 2010 Census Block
CAIID	YES	
SUBSCRBDOWN	YES	
SUBSRBUP	YES	Note missing “C” in name
SHAPE	NO	Created for the latitude, longitude value pair

Overview of Data Load Process

In general all data went thru the following processing steps:

1. Geocoded the addresses using an Arroyo flow and either the Google or the Yahoo geocoder, leaving the result with address and (lat, long) pairs in an Excel spreadsheet.
2. Imported the spreadsheet to the geodatabase. During this step it was essential to

import latitude and longitude columns as type Double (ESRI will convert if they are stored as Text and the target type is specified). The result is a simple table (not a feature class) with a name "catN_name"; e.g., "cat3_hosp".

3. Created a feature class with point shapes corresponding to each (lat, long) pair from the table using ArcCatalog's "Create Feature Class from XY Table" option. Remember, X is Longitude, Y is Latitude, and use WGS 1984 as the coordinate system. The result is a feature class with a name like "catN_name_point".
4. Created a feature class with the same content as the previous step but with the geographic coordinate tolerance value the same as the transfer model. The result is a feature class with a name like "catN_name_point_tol".
5. Created a new feature class with a column containing the ID of the containing year 2010 Census Block ID using ArcCatalog's spatial join feature. The result is a feature class with a name like "catN_name_point_tol_cb".
6. Copied the data to the bb_service_cainstitutions table

Category 1: Schools

Private school source file is "PrivateSchool_GeoMatched.xls", tab "Matched" with 1,267 rows.

Public school source file is "PublicSchool_GeoMatched.xls", tab "Matched" with 2,598 rows.

Internal processing notes:

1. The private-school spreadsheet has a column "OBJECTID" that should be imported, or if it is, it must be renamed first.
2. The school names often are sharply abbreviated; e.g., "Woodside", with no indication of whether it is an elementary, secondary, high school, etc.

Category 2: Public Libraries

Source file is "LibraryPlusSubmitted.xls", tab "SurveyPlusSubmitted", with 472 rows.

Internal processing notes: None.

Category 3: Hospitals

Source file is "Hosp_Submitted_Matched.xls", tab "NJHA plus Survey Matched", with 1108 rows.

Internal processing notes: None.

Category 4: Public Safety

Source file is "PSAP_911_Matched.xls", tab "911 with Matched", with 351 rows.

Internal processing notes: File had some rows shifted off by one position, which resulted in discarding “public wifi” field.

Category 5: Higher Education

Source file is “HigherEd_Geocoded_RateMatched.xls”, tab NCES+NJEDGE”, with 157 rows.

Internal processing notes: None.

Category 6: Local Government

Source file is “NJ_State_Verizon_Geocoded.xls”, tab “GoodGeocoding”, with 1,947 rows.

Internal processing notes: Excel sheet yielded several empty rows when loaded to a simple table. Ignored these, they will fail in the spatial join and will be excluded during the final load operation.

Categories 6 and 7: Other Community Support

Source file is “Submitted_GovNGO_CAIs.xls”, tab “Sheet0”, with 62 rows.

Internal processing notes: None.

Validation

Discarded records that did not meet transfer model requirements:

No zip code: 4

No building number: 328

No street name: 1

No city: 19

State outside NJ: 3

Successfully Loaded: 7221

Our investigation revealed that many institutions, particularly public schools, do not post or use street/building numbers. Therefore, the requirement for a street number is causing us to discard many otherwise valid records.