DATA DEVELOPMENT & VALIDATION METHODOLOGIES WHITE PAPER

Commonwealth of Pennsylvania
State Broadband Data and Development (SBDD)
Broadband Mapping Project

NTIA Data Submittal
March 31, 2011

Baker

Table of Contents

Introduction

The following sections of this document provide an overview of the process used for the SBDD Broadband Mapping data development for the Commonwealth of Pennsylvania. The following narrative is depicted in Appendix A, Commonwealth of Pennsylvania SBDD Process Workflow, and Appendix B, State Broadband Data Validation Workflow, included at the end of this document.

Broadband Provider Outreach

The following outreach procedure provides the framework for communicating with Broadband Service Providers (providers). The primary goals of the outreach approach documented herein are to:

- Promote provider understanding and acceptance of the Broadband Mapping process, results, and benefits
- Clarify NTIA Broadband Mapping requirements
- Facilitate data confidentiality agreements as required
- Minimize the submittal of invalid data
- Enhance provider understanding of the semi-annual update process
- Work with providers to evaluate submittal options to facilitate data submittals

Data Submission Guidelines

Guidelines for the providers' submission of Broadband Mapping Data are documented in the "Data Submission Guidelines". These Guidelines define technical requirements, submission specifications, and coordination and documentation activities.

Pennsylvania Broadband Providers Website

A URL was deployed (http://www.bakergis.com/PABroadbandProvider/) to communicate and distribute NTIA NOFA requirements to providers along with outreach and data submittal materials including:

- NTIA NOFA and subsequent clarification
- Outreach letters to providers
- Draft Non-Disclosure/Data Sharing Agreement
- Quick Start Guides
- Data Submission Guidelines
- Data Transmittal Letter
- Broadband Data Submittal Templates
- Census TIGER Data
- Data Submittal Assistance Contact Information

Outreach Delivery Vehicles

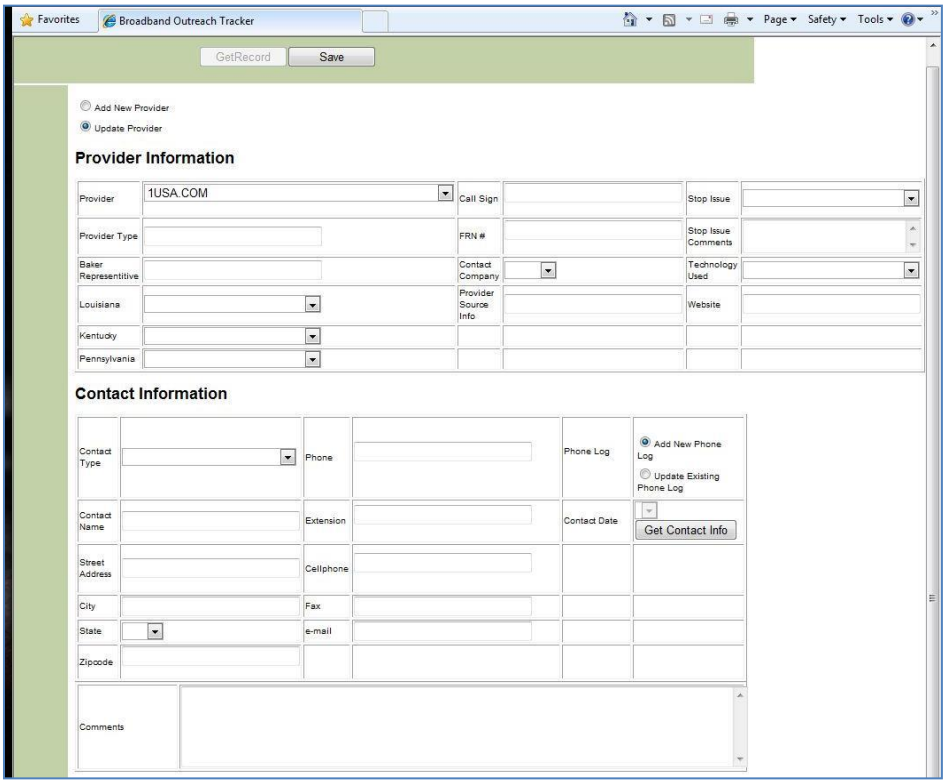
- A State Broadband Mapping Initiative Call for Data letter from the Commonwealth of Pennsylvania Department of Community and Economic Development (DCED) was emailed to all providers in the Commonwealth. This initial provider contact letter described the program and the role of Michael Baker Jr., Inc. (Baker) acting on behalf of the DCED for Broadband Data Collection and Mapping.

- Baker distributed a follow-up letter to all providers describing the data submittal requirements and material and help available to aid with the data submittals.
- Submittal assistance was provided to providers that needed help with data submittals.
- Presentations were conducted with various broadband provider associations to present the data submittal requirements and answer questions.
- Email communication and electronic transfer of data was encouraged to facilitate a faster delivery of data and information.
- A URL was deployed and promoted to distribute outreach material and information concerning the Broadband Mapping Project.
- A secure FTP URL was provided for submittal of broadband data by providers.

Broadband Outreach Tracker Application

The Tracker application (Figure 1) was utilized to collect all correspondence with providers and feedback on the effectiveness of the outreach activities by tracking items such as:

- The number and content of incoming e-mails and letters submitted from the providers
- The number and source of comments, questions, and suggestions made by providers
- The number and source of comments, questions, and suggestions made by attendees at provider meetings and conference calls
- Provider contact information and data submittal status.



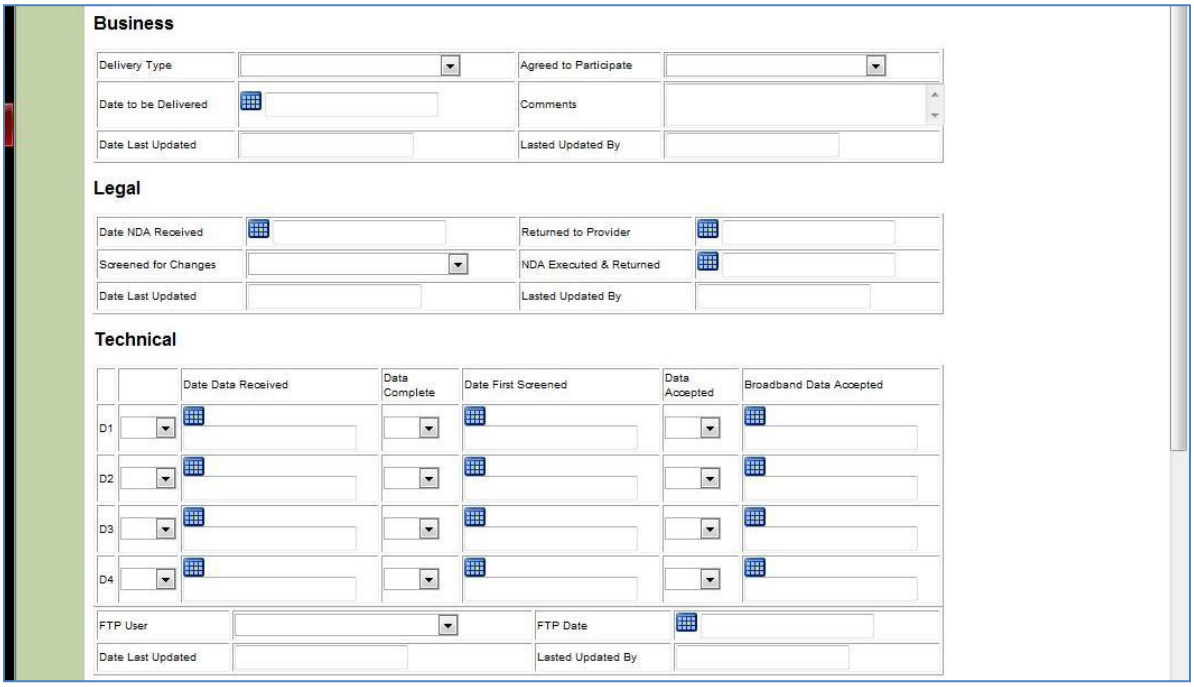


Figure 1 Broadband Outreach Tracker

Provider Submittal Validation

When a data submittal is received from a broadband service provider, it is updated in the Broadband Outreach Tracker and run through an initial validation process to assure that it meets the submittal guidelines.

Validation Checklist

The following items are part of this initial data validation process:

- Verify provider's transmittal letter requested in Data Submission Guideline with is complete and matches submitted data
- Verify the file naming conventions
- Verify each file is machine readable
- Verify data is in the correct GIS or Tabular format/file type
- Verify each field is populated and no empty or NULL values are present for mandatory fields
- Verify all ID (record number points) are unique within the submittal
- Verify all attribute data is formatted according to the submittal guidelines
- Verify topology for all geospatial submissions
- Verify Metadata for all submissions
- Verify the required contact information is included
- Verify adherence to Data Submittal Guidelines (see http://www.bakergis.com/PABroadbandProvider/ to access Data Submittal Guidelines)
  Broadband Service Availability (at least one)
  - Individual Street Addresses (Sec 3.1 & 4.1)

- Census Blocks < 2 sq mi (Sec 3.3 & 4.3)
- Street Segments for Census Blocks > 2 sq mi (Sec 3.2 & 4.2)
- Service Overview (Sec 3.4 & 4.4)
- Polygonal Boundary Area(s) (Sec 3.8 & 4.8)

Middle-mile Points (Sec 3.5 & 4.5)
Community Anchor Institutions (Sec 3.7 & 4.7)
Last Mile Connection Points (Sec 3.6 & 4.6)
WISP Antennas (Sec 4.9)

Data Usability Determination

The validation results are evaluated by the outreach and aggregation persons to determine the usability of the data. If the data meets the submission specifications, it is forwarded on for data aggregation. If it is determined to be unusable, it is returned to the provider for resolution. If the data can be manipulated to get it into a usable format, it is manipulated as required, and then forwarded on for data aggregation.

SBDD Data Development

Data from the providers may be submitted in various formats as defined in the Data Submittal Guidelines, or in some cases unspecified formats may be accepted to help facilitate provider participation. Depending on the format of the submitted data, it is processed through one of the following processes to upgrade it to the NTIA SDBB data standards.

Spatial Data

After validation and any required manipulation of any spatial data submitted by the providers, it is georeferenced and simply loaded into the appropriate NTIA geodatabase feature class.

Address Data Geocoding

If not already in the standard address point template, the provider tabular address data is first loaded into that template. The data is then exported to a geodatabase table using the ArcGIS Conversion Tools. ArcGIS geocoding tools are then utilized geospatially locate the address points for the tabular records. Interactive address rematching is performed against two additional street centerline datasets as needed to increase geocoding matching results. The NTIA deliverable is the geocoded address point geodatabase table. The geocoded address points are also subsequently aggregated to the census block or road segment feature class for public web map display.

Census Block Aggregation

If not already in the standard census block template, the provider tabular census block data is first loaded into that template. The data is then exported to a geodatabase table using the ArcGIS Conversion Tools. The provider tabular census block records are then joined to the geodatabase 2000 U.S. Census Block. This join is performed as many times as necessary for multiple Trans Tech values for each Provider/Census Block combination. The NTIA deliverable is the census block geodatabase table.

If the list of census blocks contains blocks > 2 sq. miles then these blocks are used to select all the 2000 U.S. Census TIGER centerlines that intersect those blocks. The Census Block record data is aggregated to each Road Segment within the Census Block. This process is performed as many times as necessary for multiple Trans Tech values for each Provider/Census Block combination.

Road Segment Aggregation

If not already in the standard road segment template, the provider road segment data is first loaded into that template. The data is then exported to a geodatabase table using the ArcGIS Conversion Tools. If the provider submittal included graphic centerline segments, these are migrated into the delivery geodatabase along with the linked attribute records. If the provider submittal was tabular road segment records only, they are then joined to the geodatabase 2000 U.S. Census TIGER centerline feature class. This join is performed as many times as necessary for multiple Trans Tech values for each Provider/Road Segment combination. The NTIA deliverable is the road segment geodatabase table.

If the provider road segment data lie within census blocks <= 2 sq. miles then the road segment data is aggregated to the census block. This process is performed as many times as necessary for multiple Trans Tech values for each Provider/Road Segment combination. The NTIA deliverable is the road segment geodatabase table.

Overview Data Aggregation

Provider Service Availability Areas submitted for entire county areas are loaded into the NTIA geodatabase Overview table. If not already in the standard template, the provider data is first loaded into that template. The data is then exported to a geodatabase table using the ArcGIS Conversion Tools. The provider overview records are then joined to the geodatabase 2000 U.S. Census County feature class. This join is performed as many times as necessary for multiple Trans Tech values for each Provider/County Area combination.

Polygonal Boundary Aggregation/Integration

Providers submitting polygonal service area data are handled in two ways. Wireline Provider data is aggregated to the census block feature class for areas where census blocks <= 2 sq. mi., or road segment feature class for areas where census blocks > 2 sq. mi. Wireless Provider Service Availability Areas submitted by polygonal area are simply loaded into the NTIA geodatabase Poly_Bndry feature class.

Wireline Provider

The polygonal data is georeferenced and loaded into the Poly_Bndry feature class. The polygon is then attributed, manually if necessary. Depending on the area, census blocks < or => 2 sq. mi., a selection set of either census blocks or road segments that intersect the polygon boundary is created. The attributed polygon boundary is then joined with census blocks or road segments table to attribute accordingly. This join is performed as many times as necessary for multiple Trans Tech values for each Provider/County Area combination. The NTIA deliverable is the census block or road segment geodatabase table.

Wireless Provider

The polygonal data is georeferenced and loaded into the Poly_Bndry feature class. The polygon is then attributed, manually if necessary. Multiple Poly_Bndry records are created for multiple Trans Tech values for each provider. The NTIA deliverable is the polygon boundary geodatabase table.

Middle/Last Mile Data Integration

If not already in the standard template, the data is first loaded into that template. The data is then exported to a geodatabase table using the ArcGIS Conversion Tools. The point features are geo-located utilizing the lat/long information provided. The NTIA deliverable is the middle or last mile geodatabase table.

Community Anchor Institution Integration

Providers supplied some Community Anchor Institution (CAI) data with the data submittals. But the majority of the data was collected from existing GIS Layers maintained by the Commonwealth of Pennsylvania, outreaching to CAIs through state agencies and their contacts, and having CAIs complete an online survey at http://www.bakerbb.com/pa_institution_survey/.

Provider CAIs

If not already in the standard template, the data is first loaded into that template. The data is then exported to a geodatabase table using the ArcGIS Conversion Tools. The point features are geo-located utilizing the lat/long information provided. Address data is used to geocode locations only when lat/long data is not provided.

Commonwealth CAIs

CAI shapefiles were provided through the Commonwealth's other geospatial efforts. The shapefiles were then exported to the NTIA geodatabase CAI feature class. Various sources for obtaining broadband information for the CAIs were utilized. Various state agencies provided some of the information, i.e. Pennsylvania Department of Education (PDE) provided tabular broadband information for schools, PDE provided tabular broadband information for libraries, Pennsylvania State Police provided tabular broadband information for their facilities. A CAI data survey website was also deployed and the URL distributed by various state agencies to the CAI contacts. Data from all of these sources were then aggregated into the CAI geodatabase table for the NTIA deliverable.

Provider Validation

After data development, service availability maps are generated and submitted to the providers to validate their mapping results. This provides a "sign off" on the interpretation of the submitted data and extends the outreach efforts by providing a visual representation of the data to be delivered to the State and the NTIA.

Types of Provider Maps

Provider maps generally consist of the following types.

Outreach Maps

Often, providers will send data which does not contain all the information needed for a NTIA compliant dataset. In such cases, as an aid to the outreach communication, it may be necessary to produce a map to help the

provider locate their service area or verify data they have provided. These maps may take many forms, but generally are of two types:

- General Location Maps – these maps are often produced when the provider does not have a list of address or other standard submittal data and needs help defining their service area. A typical map will show counties, major roads, and towns of the general area the provider has stated as their service area. The intent of the map is to give the provider a way to markup or delineate their service area. If a provider has not provided required attribute information such as Technology of Transmission, Speed Data, etc. then it may be necessary to add a visual clue to this data like an information stamp on the map that they can easily fill out. If the provider sends the map back with a service area boundary, this can then be digitized and sent back to the provider for verification.
- Verification of Provider Supplied Boundaries – these maps are produced when the provider has sent service area boundary information which is confusing or otherwise unclear. Often these are produced when providers send CAD maps, hand drawn maps that need digitization, or lists of zip codes or counties served. A typical map will place the interpreted boundary over a location map so the provider can verify the service area. As with the General Location Map, information stamps or other visual clues may be placed on the map.

Initial Verification Maps

Once the provider data has been processed and the census block and road segment feature classes created, an Initial Verification Map (Figure 2) is produced to give the provider a visual representation of their service area by census block. These maps enable the provider to verify their service area and make changes if necessary. Initial Verification Maps are produced using a set of standards and produced at the highest resolution necessary to convey the map information to the provider. Initial Verification Maps are also produced for Wireless Polygon areas.

Detailed Verification Maps

Providers who have questions about their service areas may request additional information to help clarify issues. In these cases, it may be necessary to create a Detailed Verification Map to highlight the areas in question. Detailed Verification Maps provide the same information as Initial Verification Maps only at a higher resolution. Several maps may be needed to accurately portray an area in question.

Revised Maps

Revised maps take two forms:

- Initial or Detailed Verification Maps which have been annotated or marked-up by the provider
- Outreach produced Initial or Detailed Verification Maps incorporating provider changes

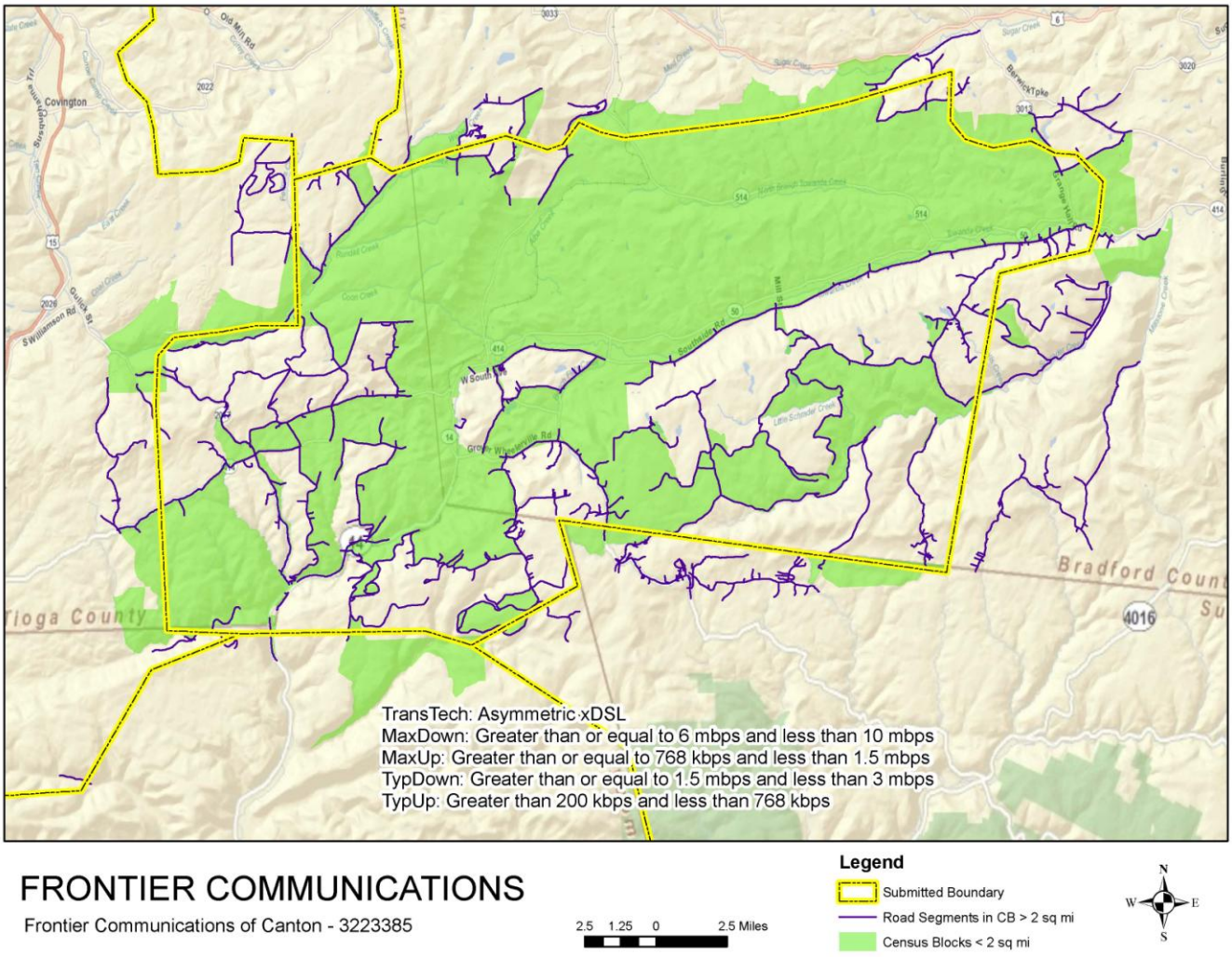


Figure 2 Provider Map

Data Validation

A critical component of the project is the validation of the data submitted by the broadband service providers. Data from various sources, as described in more detail in the following sections, is utilized to develop a level of confidence in the data received from the broadband providers.

Validation Data Set Collection and Development

This validation process employs data sets developed or acquired from different sources as described in the following sections.

Provider Feedback Loop: Maps of completed provider service areas and data are furnished back to the providers for confirmation of the processed/aggregated information. Feedback is integrated into the each provider's dataset.

Telogical Systems Wireline Market Intelligence Data: This commercially available dataset was developed using a methodology that incorporates deep web crawling and additional means, including direct mail harvesting and advertising collaterals (including door to door) to gather cable and telecommunication provider information. This dataset is used as a validation source for wireline provider service area coverage, Technology of Transmission, and Speed.

American Roamer Wireless Market Intelligence Data: This commercially available dataset is used as an independent source to verify information submitted by providers of wireless broadband service. This dataset is used as a validation source for wireless provider service area coverage.

Prior Commonwealth Broadband Mapping Dataset: Under the requirements of the Commonwealth's Act 183 of 2004 legislation, broadband coverage data was previously collected by the Commonwealth. These datasets are used as a validation source for provider service area coverage and Technology of Transmission.

FCC Speed Test: The FCC speed test data includes the IP addresses for each specific speed test conducted. This IP address is queried against a web search engine to determine the provider assigned to that address and is used as a validation source for the provider service coverage and typical speeds.

Fixed Wireless Line of Sight Analysis: Utilizing the existing PAMAP LiDAR for topography generation and determining tower/antennae heights, line of sight analysis is performed to determine areas of reported fixed wireless broadband coverage that is questionable.

Field Data Acquisition: Broadband technicians visited a sampling of census block locations to gather broadband data to be used for validation. The following criteria were taken into account when developing the census block sampling dataset:

- urban vs. rural census block characteristic
- censes block grouping
- land vs. water census block characteristic

The overarching mission of the Federal broadband stimulus program is to expand Broadband service to areas that are currently unserved and underserved. Also, the market intelligence validation sources typically represent some rural, but more urban areas. Thus, our field data collection efforts were targeted more towards the rural areas; split 90% rural, 10% urban.

Additionally, a study by Penn State University (Glasmeier 2002) notes that a large number of census block groups typically fit within any given cable or telephone company service areas. Therefore, our field sample was also based on selection of one census block per block group and a land mass greater than 50% to avoid field visiting areas covered mostly by water. There are a total of 10,387 block groups in PA. Using a statistical sample size calculator based upon the number of block groups in the state and +/- 4% margin of error at a 95% confidence level, the sample size is 568 census block locations statewide. The procedure for selecting the calculated field verification census blocks is provided below.

1. Select one census block per census block group
    a. Convert the census block groups polygon to label points.

b. Select the census block polygon by doing a spatial selection using census block groups label points.

2. Select from the current selection where the census block land mass is 50% or greater and the block is rural.
3. Export the selected blocks to a new shapefile. This reset the FID for the next step.
4. Select every 2nd, 3rd, 4th, or so on to get the desired number of blocks. Query used to select: MOD("FID",2) = 0. This will select every other record.

The planned census block field locations are shown in Figure 3.

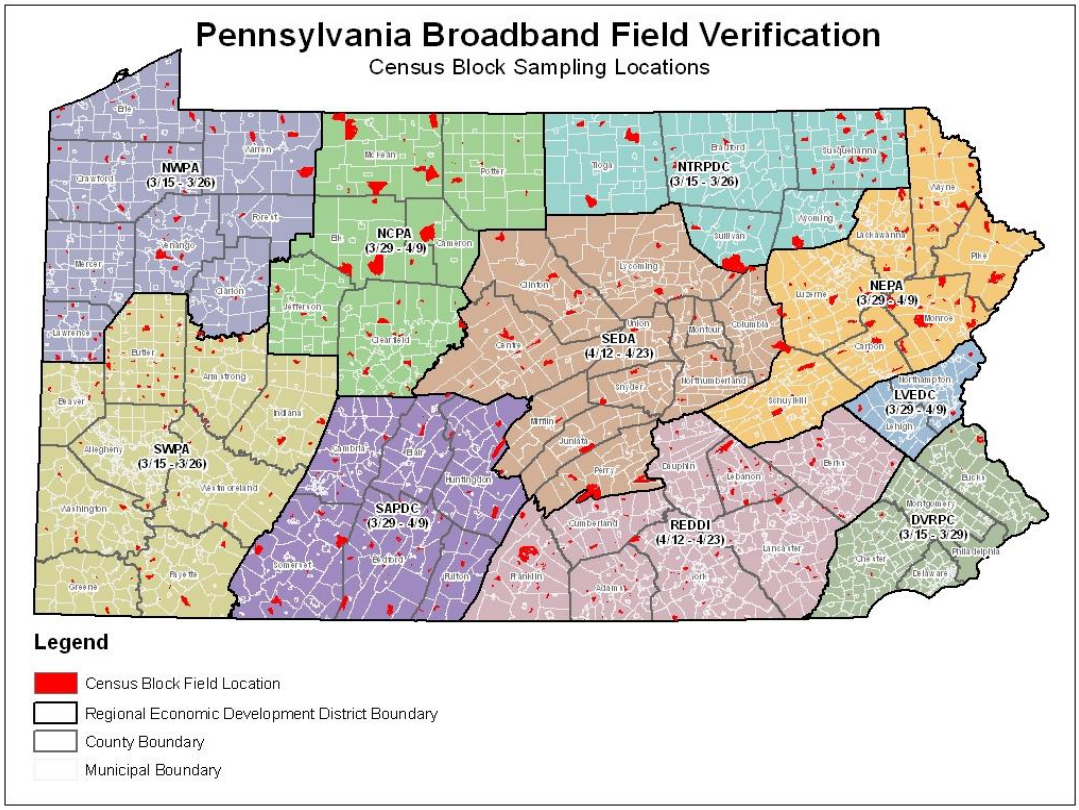


Figure 3 Planned Field Verification Census Block Locations

For each census block in the sample set, broadband technicians collected data using Panasonic Toughbook computers, loaded with MapPoint mapping software, and a customized Microsoft Access data collection form with the ability to automatically import GPS coordinates. The sample census blocks were pre-loaded and directly accessible from MapPoint. Two types of data collection were conducted (infrastructure observation and wireless speed testing) and the results were recorded and linked to the corresponding field location coordinates within the designated sample census block. The information collected by the field broadband technicians includes:

Wireline:

- GPS coordinates
- circuit infrastructure feeding the area (copper, fiber, cable)

- local distribution hut equipment inspection, where allowed/possible
- witness access circuit speed tests, where allowed/possible
- facility elevation (measurement relative to grade), where allowed/possible
- distance from DSLAM measurement where applicable and determine access speed capability with an accuracy within 500ft using mapping software
- collect site pictures

Wireless:

- GPS coordinates
- internet speed test

The map in Figure 4 shows the locations (blue points) of the census block field surveys that were performed.

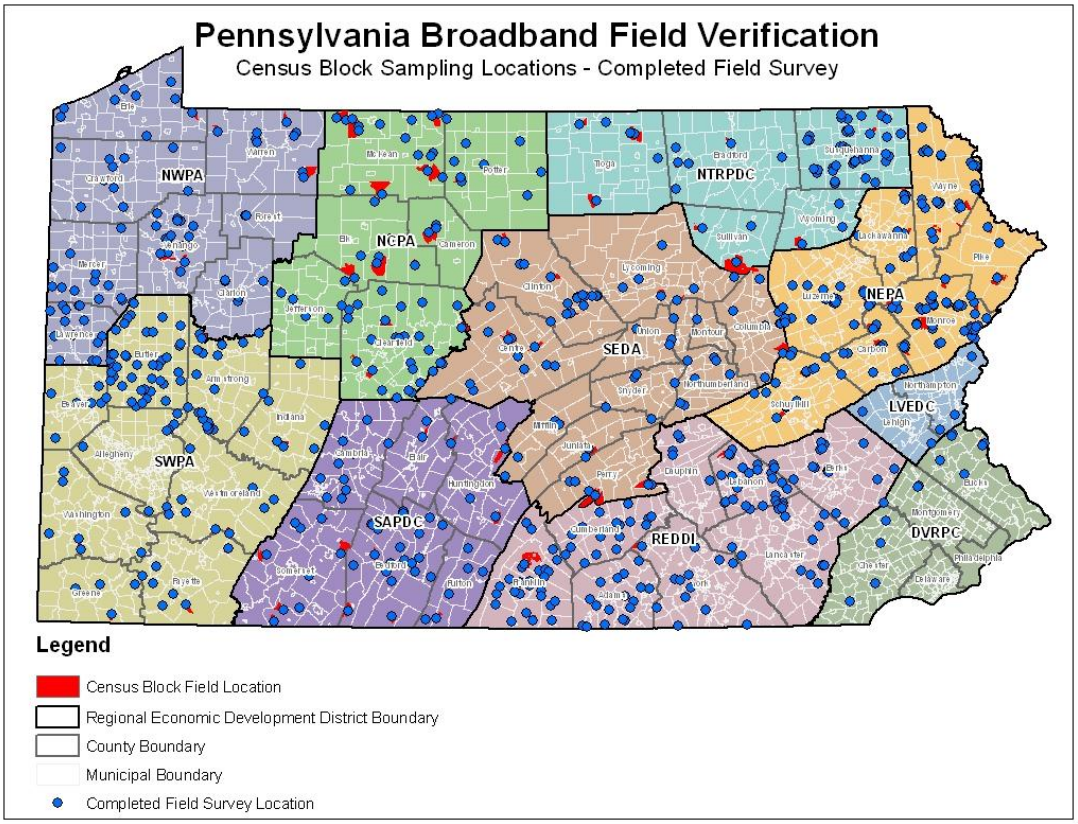


Figure 4 Completed Field Verification Locations

For the 568 census blocks that were visited, 2821 individual wired/wireless data elements were recorded and 3666 pictures were taken at those locations. This field collected dataset is used as a validation source primarily for wireline and wireless technology of transmission, middle mile, and wireless speed.

Provider Data Validation Process

Provider Feedback Loop: Feedback received from the providers is visually inspected and integrated directly in the mapping GIS database.

Service Area Validation Data: The Telogical wireline service area data is tabular and contains a separate record for each provider/technology of transmission combination with an associated census block or TIGER road segment, depending on the whether the size of the census block area (=/< or > 2 sq. mi.). This data is exported into an ArcGIS data format. The American Roamer wireless service area data is already in and ArcGIS data format. The validation data is then joined to the provider service area data by census block or TIGER road segment ID. Any database records in the provider or validation tables that cannot be joined are output to a separate layer that indicates the areas of discrepancy between the two datasets. The joined tables are then queried to detect any speed discrepancies which are also output to a separate discrepancy layer.

Field Validation Data: The field data are also collected in tabular database format, and represent a specific lat/long spatial location for each record. This data is also exported into an ArcGIS data format, joined to the provider data, queried to validate pertinent attribution. Again, records not joined and/or with detected attribution discrepancies are output to separate GIS layers.

Topology: The ArcGIS Validate Topology Tool is used to flag any topology issues in the broadband data. Flagged issues are reviewed to identify false positives and update true errors as required.

SBDD Check Submission: The NTIA-provided SBDD Check Submission tool is utilized to validate that the deliverable broadband data is consistent with the business logic rules set forth by the NTIA and a passing receipt is provided with the data submittal to NTIA.

Stakeholder Feedback: The state broadband mapping website includes a feedback function. Comments received from stakeholders such as the regional Economic Development Districts and the public are reviewed and used to validate the provider data submissions.

Validation and Confidence Level Reporting

To facilitate validation and confidence level reporting, Baker deployed a validation application called Statistical Evaluation and Assessment System (SEAS), shown in Figure 5, which automatically compares the multiple independent validation datasets against the broadband service provider supplied information. The SEAS application uses statistical methodologies to report the confidence level in the spatial and attribute accuracy of the information. Appendix B shows the validation workflow.

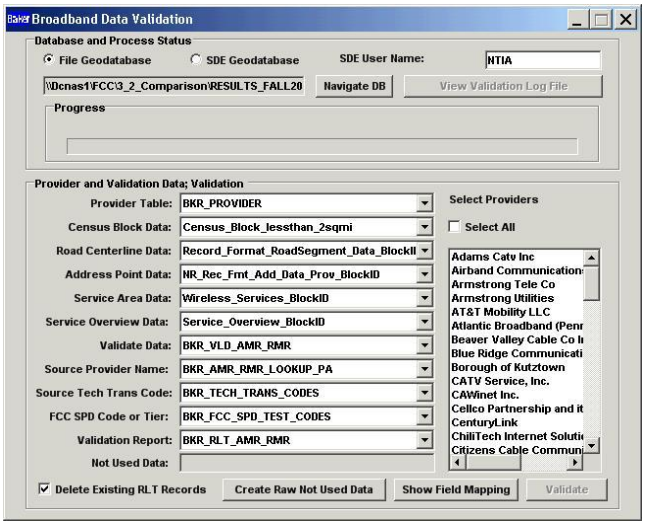


Figure 5 SEAS

The SEAS comparison is a three-part validation process:

1. Comparison of the collected validation source against the aggregated broadband provider data.
2. Match percentage calculation for each provider reported in the DataPackage.xls, "Provider Table" tab, "Comments" column.
3. Confidence score calculation displayed on the state broadband website.

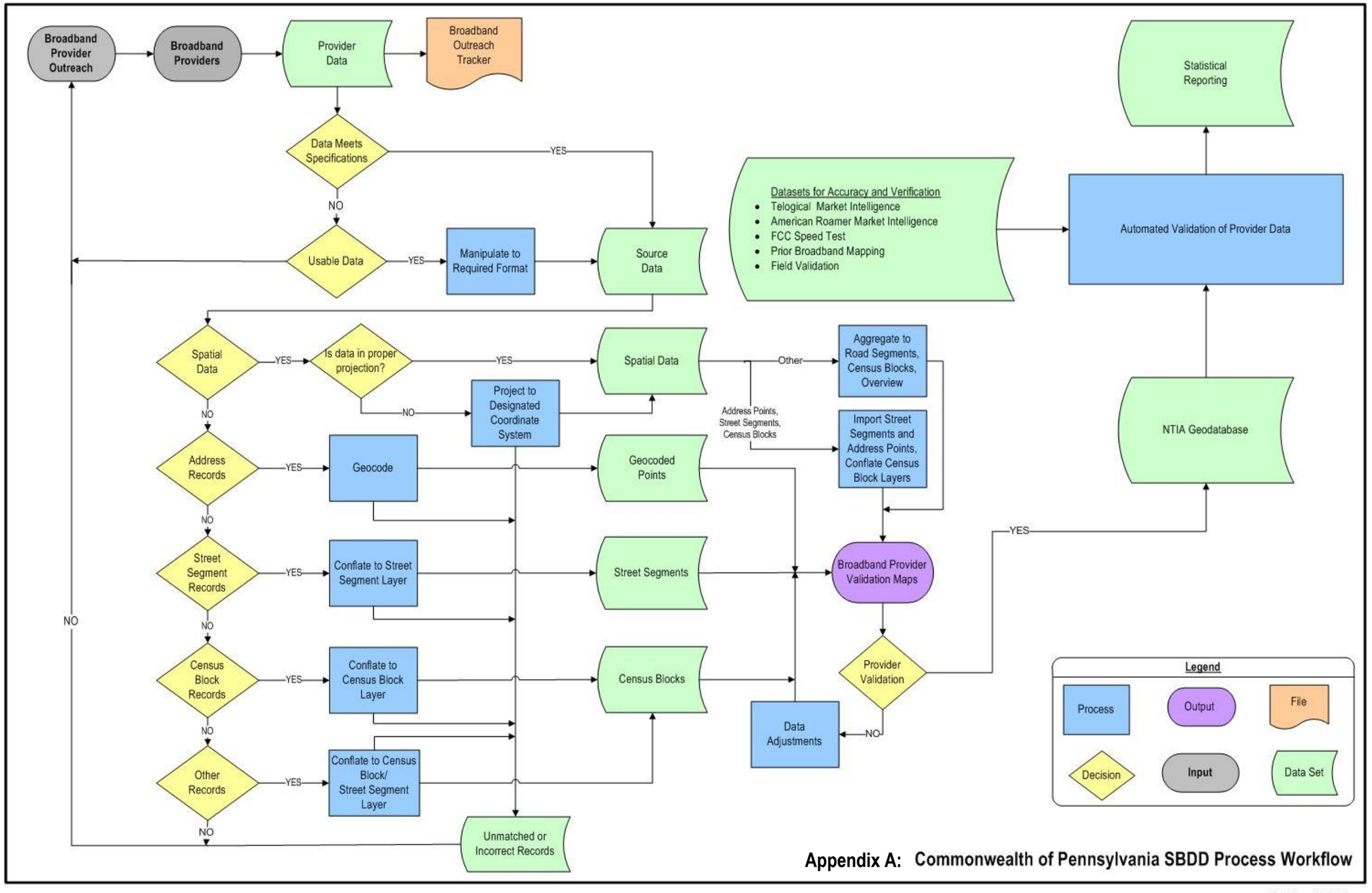
After completing all validation data source collections, SEAS is used to automatically compare the multiple validation datasets against the aggregated broadband data which came from the providers. Through the SEAS accumulation table, it produces a match percentage per broadband service record based upon the number of matches that record has against each validation source. The matched percentage for each record is the result of the total count of the matched validations for the record divided by the total validation source being compared against the record. Validation confidence rating/score is assigned on a scale of 1 to 5 based upon the percentage of validation source matches as per the following score results:

- 1 Star = 0% - 19% Match
- 2 Stars = 20% - 39% Match
- 3 Stars = 40% - 59% Match
- 4 Stars = 60% = 79% Match
- 5 Stars = 80% - 100% Match
- "No Analytics" = No validation source available for that provider

The Commonwealth's public broadband mapping website (www.broadbandinpa.com) is updated with the confidence level results at the record level based upon the queried geographic location and the following shows an example of this representation.

Provider Name	Transmission Technology	Max Download Speed	Max Upload Speed	Confidence Score
AT&T Mobility	Mobile Wireless	Greater than or e…	Greater than or e…	★★★☆☆
Verizon	Asymmetric xDSL	Greater than or e…	Greater than or e…	NO ANALYTICS
Comcast	Cable Modem – Other	Greater than or e…	Greater than or e…	★★★★☆

The matched percentage for the records for each provider are summarized and then divided by the total count of the records to create the final matched percentage for the specific provider. These percentages are included in DataPackage.xls on the Provider Table tab in the Comments column.



Appendix A: Commonwealth of Pennsylvania SBDD Process Workflow

October 1, 2010

Appendix B: State Broadband Data Validation Workflow