METHODOLOGIES USED TO CREATE AND VALIDATE BROADBAND DATASETS FOR THE OCTOBER 8, 2010 SBDD SUBMISSION

Broadband data for Massachusetts was, integrated and verified by the Massachusetts Broadband Institute, a division of the Massachusetts Technology Collaborative. Data development was performed using ESRI ArcGIS 9.3 software.

Data Integration

Data were received from broadband service providers in varying formats and levels of detail. No two datasets were alike, which required a significant amount of manual review and editing to integrate the information into a common format. Although Excel and Shapefile templates were provided, very few datasets were received in the template formats and attributes were not always provided using the standardized coded values requested. In addition, attribute field names were inconsistent between datasets, contained spaces and special characters or were missing altogether. These differences prevented the use of automated data integration models to format and import data into standardized feature class templates.

All attributes were standardized so that the provider name, doing-business-as name and FCC registration numbers were consistent throughout the datasets and that attributes complied with valid value list (e.g., for technology of transmission, spectrums used, maximum advertised and typical speeds, etc.).

Geocoding

Unless otherwise specified, address data was geocoded using street addresses and zip codes from NAVTEQ 2008 Q4 streets data, which was developed though a partnership between NAVTEQ and the Massachusetts Office of Geographic Information (MassGIS) for increased geocoding accuracy and success rates for State 911 data.

Data transfer model loading

The final datasets for each provider were appended and loaded into the SBDD transfer schema. Geometry and topology checks were performed a final time and the data were checked for conformance with SBDD business rules.

DATA DEVELOPMENT – WIRELINE AVAILABILITY

This section describes the methods used to create the following datasets representing wireline broadband availability (e.g., cable, xDSL, other copper wireline, fiber optic and other unclassified wireline services) by census block and/or road segment:

- BB_Service_CensusBlock and
- BB_Service_RoadSegment

The various wireline broadband availability data formats received include:

1. Non-geographically referenced CAD files containing cable or fiber strands;
2. Geographically referenced Shapefiles containing census block polygons or road segments;
3. Excel spreadsheets or delimited text files containing census block IDs
4. Excel spreadsheets or delimited text files containing individual street addresses;
5. Excel spreadsheets or delimited text files containing street address ranges; and
6. Written or verbal narratives of service areas.

For areas where census blocks are less than or equal to 2 square miles in area, a template containing polygon geography for the appropriate census block vintage was used (2009 was the default census block vintage for data not tied to a census block ID). Otherwise, a template containing line geography from 2009 TIGER/Line roads that intersect 2000 vintage census blocks greater than 2 square miles in area. Associated attribute information included provider identification, technology of transmission and upload and download speeds.

The integration methods used, and described below, varied according to the source data format.

1. <u>Integrating CAD strands</u>: Cable strands submitted in CAD format were georeferenced to street centerlines and a 200 foot buffer was created from the strands. 2009 census blocks and 2009 TIGER/Line road segments (in census blocks greater than 2 square miles in area) that intersected the 200 foot buffer were classified as served and associated attribute information from tabular datasets or narratives were populated accordingly.

2. <u>Integrating census block and road segment polygons</u>: Data provided in Shapefile format required minor formatting of attribute field names and values to match the common schema.

   (a) The census block vintage was determined by reviewing ID values and attributes were imported into the census block template.

   (b) If vector data was provided from a source other than TIGER/Line roads, a spatial intersection with a 200 foot buffer was performed to transfer attributes to the corresponding TIGER/Line road segments.

3. <u>Integrating tabular data containing census block IDs</u>: Tabular information relating to census blocks referenced three different versions (or vintages) of census block data from 2000, 2008 and 2009 and were joined to the corresponding polygon geometry using the 15 or 16 character FIPS IDs. Prior to integrating the individual provider information into a single combined dataset, the 2008 and 2009 census block data were summarized and joined to the 2000 census block polygons using the first 15 characters of the FIPS ID for each unique transmission technology type while retaining the maximum advertised and typical speeds and other associated validation and data processing attributes.

4. <u>Integrating tabular data containing individual street addresses</u>: Tabular data containing individual street addresses, generally representing subscriber addresses, were geocoded using NAVTEQ 2008 Q4 streets data to generate point locations. 2009 census blocks and 2009

TIGER/Line road segments (in census blocks greater than 2 square miles in area) that intersect a 200 foot buffer of the points were classified as served. Associated attributes were also imported.

5. Integrating tabular data containing street address ranges: (a) If tabular data was based on 2009 TIGER/Line roads and included a TIGER line ID, the TIGER line ID was concatenated with the address fields to create a unique ID and linked to corresponding TIGER/Line geography. Associated attributes were also imported.

   (b) If tabular data was not based on TIGER/Line roads or did not have a means for creating a unique ID to link to the TIGER/Line data, the minimum, mean and maximum left and right street addresses were geocoded using NAVTEQ 2008 Q4 streets data to generate point locations. As with the individual street address methodology above, 2009 census blocks and 2009 TIGER/Line road segments (in census blocks greater than 2 square miles in area) that intersect a 200 foot buffer of the points were classified as served. Associated attributes were also imported.

6. Integrating narrative data: (a) Location information provided in narrative form, such as the names of streets served or unserved, were incorporated by classifying the qualifying road segments as served. A spatial intersection was then performed to classify any census blocks with area less than 2 square miles as served.

   (b) Attribute information provided in narrative form generally applied to all records or an easily identifiable subset of records in a dataset and the standardized values were assigned to the appropriate field in batch.

Data standardization
For the many datasets based on census blocks from 2008 or 2009, the attributes were summarized by the 15 digit 2000 census block ID. Maximum advertised and typical speed information were retained for each unique provider and technology combination for each 2000 census block ID and imported into a template containing 2000 vintage census block geography. Records with download speeds below 768 kbps (i.e., that don't qualify as broadband service) were removed from the final dataset.

DATA DEVELOPMENT – WIRELESS AVAILABILITY

This section describes the methods used to create the following dataset representing wireless broadband availability (e.g., fixed and mobile wireless and satellite services) by service area:

- BB_Service_Wireless

The various wireless broadband availability data formats received include:

1. Geographically referenced Shapefiles or MapInfo files containing service area polygons;
2. Geographically referenced KML raster files depicting service areas;

3. Non-geographically referenced PDF files depicting service area polygons; and
4. Hard copy maps with hand-drawn service areas.

Associated attribute information included provider identification, technology of transmission, wireless spectrums used and upload and download speeds. In some cases, attributes were provided in a separate tabular or narrative form or had to be acquired from the provider's web site. If providers offered more than one spectrum, a separate polygon was created for each unique provider and spectrum combination.

Data integration methods used, and described below, varied according to the source data format.

1. <u>Integrating service area polygons</u>: Data provided in vector format required minor processing to fix geometry errors and create separate polygons for unique provider and spectrum combinations. Attribute field names and values were created, formatted and/or populated from tabular or narrative form to match the standardized template format.

2. <u>Integrating service area raster images</u>: Propagation model outputs provided as KML raster images were imported into the GIS system; however, the geographic reference information was not able to be preserved. The imported raster images were georeferenced in the GIS by matching the intersections of propagation area boundaries and roads in Google Earth. Once georeferenced, the raster images were converted to polygons, then tagged with and aggregated by the associated tower ID and spectrum information to create service areas polygons for each propagation model. Additional associated attribute values were populated from information provided in narrative form.

3. <u>Integrating static PDF maps</u>: The PDF maps containing wireless access points and service area buffers were georeferenced using known locations, such as road intersections. Point data were created from the georeferenced map and service area buffers were recreated. Individual service areas were tagged with spectrum information and aggregated into a single service area for the provider and spectrum combination. Additional associated attribute values were populated from information provided in narrative form and the resulting service area boundaries received confidence score of 1.

4. <u>Integrating hard copy maps</u>: Hard copy maps containing shaded service areas were reproduced by digitizing boundaries based on known map locations, such as road intersections. Associated attribute values were populated from information provided in narrative form and the resulting service area boundaries received confidence score of 1.

<u>Data standardization</u>

Service area datasets for each provider were clipped to the state boundary and self-intersecting lines were fixed prior to loading into the SBDD transfer schema.

DATA VERIFICATION – WIRELINE AND WIRELESS AVAILABILITY

This section describes the methods used to verify the following datasets representing wireline broadband availability (e.g., cable, xDSL, other copper wireline, fiber optic and other unclassified wireline services) by census block and/or road segment and wireless broadband availability (e.g., fixed and mobile wireless and satellite services) by service area:

- BB_Service_CensusBlock,
- BB_Service_RoadSegment and
- BB_Service_Wireless

Verification of availability data received from providers is essential to determining the accuracy and completeness of the resulting broadband availability maps and is an ongoing process. Methodologies continue to be developed and implemented for data verification and are incorporated into a confidence ranking process. The data verification and confidence ranking methods are described below.

The data verification process employs, or will employ (designated by an asterisk *), the following methods, which supply input for the confidence ranking methodology.

1. <u>Cable service area modeling</u>: Cable strand data for incumbent cable providers were acquired as georeferenced MapInfo files from the MA Department of Telecommunications and Cable (DTC) in 93% of the 305 cable-served towns. The strands were imported and a 200 foot buffer was created to approximate the distance from the cable that a structure can receive service without excessive cost or delay. The 200 foot distance was selected based on observed distances between poles and the acceptable distances of structures from cable as defined in cable license agreements. Census blocks and road segments acquired from providers that intersected the resulting service area buffers for that provider were given an increased confidence score.

2. <u>DSL service area modeling</u>: DSL service areas were modeled from known DSL-equipped central office locations, which were geocoded using NAVTEQ 2008 Q4 streets data and refined using aerial photography, street views and bird’s-eye views from Google Maps and Bing Maps. A linear network was developed, using a comprehensive roads dataset maintained by the MA Department of Transportation (MassDOT), that encompassed all roadways within 17,800 linear feet of the central office location. A 200 foot buffer of the network was created to define a maximum service distance of 18,000 feet from the central office to the service location, based on input from industry experts, with the same 200 foot distance from pole to structure that was used in the cable model. The resulting service area buffers were cropped at town boundaries except where central offices were known to serve neighboring towns. Census blocks and road segments acquired from providers that intersected the estimated service areas for that provider were given an increased confidence score.

3. Infrastructure field surveys: Targeted field work has been performed to located broadband infrastructure, such as DSL-equipped remote terminals (RTs). As with the central offices, locations were mapped using address and landmark information acquired in the field by geocoding with NAVTEQ 2008 Q4 streets data and refining with aerial photography, street views and bird's-eye views from Google Maps and Bing Maps. Although many DSL-equipped RTs have been located in the field, they have not yet been incorporated into the DSL service area model yet due to the difficulty of predicting the directional nature of services provided from those locations. However, the locations are valuable for visual review areas of DSL coverage claimed by providers that fall outside of modeled service areas to evaluate the likelihood of service from a given RT location. These visual reviews are performed by team consisting of a GIS expert and a DSL technology expert and confidence scores modified accordingly.

4. Public surveys: Broadband subscription information is collected through web-based broadband surveys from the public and from community anchor institutions (see www.massbroadband.org/mapping/survey.html). The surveys are publicized through targeted events and publications and MBI email notifications. Information collected includes location, provider name, transmission technology, price, and speed for homes, businesses, and institutions throughout the state. At this time, the survey data is only used to verify availability by provider name and transmission technology. Census blocks and road segments acquired from providers that are within 200 feet of survey locations are given an increased confidence score. As with the service area models, the 200 foot distance represents the distance at which service can be provided without excessive cost or delay. In the future, speed test results will be summarized by census block to verify typical speed information received from providers as well.

   Responses to the public survey are geocoded through Google Maps and visually refined by the user if desired. Responses to the community anchor institution surveys are linked to existing point locations maintained by the Massachusetts Office of Geographic Information (MassGIS) or affiliated agency. Community anchor institutions that have changed addresses or are not already in the MassGIS datasets are geocoded using NAVTEQ 2008 Q4 streets data and refined using a combination of institution web sites and aerial photography, street views and bird's-eye views from Google Maps and Bing Maps.

   At this time, responses from the FCC's consumer broadband test are not used for data verification, but will be evaluated for inclusion in future data verification phases.

5. Provider web site information: If information acquired by providers – including availability and speed – appeared to be questionable, a search was performed on the provider's web site to confirm it. This was type of verification was only performed when uncertainties arose during visual review of the data. In the future, this type of review may be incorporated into a more structured approach to validate locations that are geographically dispersed throughout a provider's service area.

6. * Community cable and DSL feedback: In collaboration with Regional Planning Agencies (RPAs), availability maps will be generated and distributed to carefully selected community representatives, such as local broadband committee members or town officials, with local knowledge of cable and/or DSL services in their town. The community representatives will review and mark-up hard copy maps to identify services areas that extend too far or not far enough and to provide the location or address of the last known service location along a road. This will initially be implemented through a pilot project for the member communities of two Regional Planning Agencies. The pilot project will allow evaluation and refinement of the process before being rolled out statewide, with a focus on low confidence areas. Confidence scores will be modified based on feedback from the community representatives.

7. * Wireless drive studies: In coordination with local colleges, teams of student volunteers will be trained by an experienced field engineer to perform wireless drive studies. The students will drive pre-defined routes with intermittent stops to collect wireless signal location and quality information using Android phones operating QoS Solutions' QMapper and QPerf software (see www.qos-solutions.com). An initial pilot study will be performed, in the same two RPA regions as the community cable and DSL feedback projects, to test and refine the survey methodologies before they are rolled out statewide. Confidence scores will be modified based on results of the wireless drive studies.

Confidence Ranking

As availability data is verified, the verification status is documented in each individual census block or road segment record or subdivision of a wireless service area. The records are also assigned numeric values from 1 to 5 that represent the level of confidence in the likelihood that service is available at that location. When service availability for a given provider and technology is verified by an alternate source, the confidence value for that location is increased by one, up to a maximum score of 5. A value of 1 represents the lowest confidence in provider data and no corroborating information from alternate sources. A value of 5 represents 3 or more corroborating sources or confirmation through field work. Data of all confidence levels are included in the availability datasets; however, locations that are deemed to be inaccurate as a result of the data verification process may have their confidence value reduced and may be tagged as not part of the service area.

General guidelines of the confidence ranking process are as follows:

- Initial rankings: Data records submitted by providers are given an initial confidence ranking of "1" or "2" depending on the level of ambiguity in the submission method. For example, availability information provided by census block ID, street address or spatial object is given a confidence ranking of 2. Whereas, availability information provided as hand-drawn or narrative estimates may be given a confidence ranking of 1.

- Verification from alternate datasets: If availability at a given location is corroborated by an alternate dataset (such as the cable or DSL models, broadband survey responses, cable or DSL service area feedback from community representatives, or wireless drive study data interpolation), the verified location receives a 1 point increase in the confidence score for each corroborating dataset, with a minimum score of 3 and a maximum score of 5.

- Field confirmation: If availability at a given location is confirmed by known service locations identified through field work, it is given a confidence score of 5. Confirmed field locations include known infrastructure, such as DSL-equipped remote terminals, or known service availability acquired in wireless drive studies.

Provider Feedback Loop

All providers that submitted data received a written data submission report that described the format and completeness of the datasets they provided. This report included requests for additional information or alternate formats in the next submission and other data clarifications or corrections needed. Additional feedback was provided by phone or email conversations as needed. In certain cases, hard copy or PDF maps of estimated services were provided for verification and/or modification. In the future, all providers will receive maps and/or vector data for review and verification of census blocks, street segments or wireless service areas. Information on conflicting alternate data sources may also be provided for comment or challenge. This process will be standardized and formalized through the development of a web-based provider data portal.

DATA DEVELOPMENT – BROADBAND SPEED OVERVIEWS

This section describes the methods used to create the following dataset representing maximum advertised speeds for each provider and technology by region (e.g., county):

- BB_Service_Overview

Overview datasets received from providers corresponded to several different regional geographies – including county, cellular market area (CMA), US Census Bureau core based statistical area (CBSA) and rural statistical area (RSA) – or were provided by street address. Data were formatted and imported into a standardized polygon feature class containing geometry for the appropriate regional geography by ID (e.g., 5 digit state/county FIPS code, 3 digit CMA ID, 5 digit CBSA ID or 4 digit RSA ID). Attribute information included provider identification, transmission technology, maximum advertised speeds and subscriber weighted nominal speeds.

The MBI did not have alternate data sources for the verification of these datasets.

Data standardization

Due to common boundaries in the county and CMA geographies and uniformity of the attribute values, all CMA based datasets were able to be converted to county geography with no loss of data. However, in order to maintain data integrity, the CBSA and RSA-based data were left in the original geographies and the IDs were stored in the state/county FIPS code field. Only data for providers that did not supply speed information at the census block, road segment and/or street address level were included in the SBDD data transfer dataset.

DATA DEVELOPMENT – MIDDLE MILE INTERCONNECTION FACILITIES

This section describes the methods used to create the following dataset representing the location, technology and capacity of facilities that connect a service provider's network to another provider's network or the Internet:

- BB_ConnectionPoint_MiddleMile

Tabular data – including provider identification and facility ownership, capacity and type – were received from providers by street address or latitude and longitude. Latitude and longitude values were used to create point geometry when possible. Otherwise, street address data was geocoded using NAVTEQ 2008 Q4 streets data.

The MBI did not have alternate data sources for the verification of these datasets.

Data standardization
Facility ownership, capacity and type values were standardized to comply with valid value lists. Due to the field type of double used to store latitude and longitude, values with trailing 0's did not meet the 6 digit business rule. However, to preserve the accuracy of the data, these values were not modified to contain 6 digits. Latitude and longitude values received from providers with less than 6 digits were also not modified to prevent misrepresenting the data as more accurate than it really was.

DATA DEVELOPMENT – COMMUNITY ANCHOR INSTITUTION SERVICE SUBSCRIPTIONS

This section describes the methods used to create the following dataset representing the location and broadband service subscription of community anchor institutions throughout the state:

- BB_Service_CAInstitutions

The community anchor institution datasets deemed most relevant to broadband issues in Massachusetts were K-12 schools, colleges and universities, public libraries, hospitals, community health centers, police and sheriffs, career centers and town halls. Existing spatial datasets containing community anchor institution names and locations were acquired from state and regional agencies. The attributes were standardized and imported into a template dataset. Missing attributes (e.g., zip codes) were acquired through web searches (e.g., on institution web sites or from the US Postal Service).

In addition to the SBDD anchor type categories, an MBI subcategory field was maintained that further divides the SBDD categories, as shown below. However, the SBDD data transfer model was not modified to include this field. Only those subcategories followed by an asterisk (*) were pursued for broadband service subscription information.

SBDD Anchor Category	*MBI Anchor Subcategory*
1: School (K-12)	100: Public school * 101: Charter school * 102: Private school
2: Library	200: Public library *
3: Medical/ Healthcare	300: Acute care hospital * 301: Non-acute care hospital * 302: Community health center * 303: Board of health 304: Clinic 305: Home health agency 306: Nursing facility 307: Outpatient surgery center 308: Laboratory
4: Public safety	400: State police * 401: County sheriff * 402: Local police * 403: Fire station 404: Ambulance service 405: Public safety answering point
5: University, college or other post secondary	500: Community college * 501: Public college or university * 502: Private college or university *
6: Other community support - government	600: Career center * 601: Town hall * 602: Prison or correction facility 603: Department of public works
7: Other community support - nongovernmental	700: Church 701: Senior center

Initial data requests were made to state and regional agencies and/or associations to acquire any existing compilations of information on broadband service information at affiliated anchor institutions. Complete or almost complete datasets for career centers, state police and county sheriffs were acquired from the MA Executive Office of Labor and Workforce Development (EOLWD) and MA Executive Office of Public Safety and Security (EOPSS).

For the remainder of the anchor institutions, a campaign was implemented to acquire information through phone, email and web-based surveys from individuals associated with individual anchor institutions that were knowledgeable about its broadband services. Requests were also made through targeted outreach at events and in publications targeted at anchor institutions to increase awareness of broadband issues and participation the broadband survey. Agencies and organizations that assisted in this effort included the MA Department of Secondary and Elementary Education (ESE), MA Board of Library Commissioners (MBLC), MA Chiefs of Police Association (MCOPA), Massachusetts Municipal Association (MMA) and MA

Department of Revenue (DOR), Mass League of Community Health Centers (MLCHC) and a CIO group for public and community colleges.

Data standardization

Survey questions were developed to request information that easily understood and acquired by anchor institution staff. As a result, survey results required additional formatting to standardize the information in accordance with SBDD valid values. This information included broadband subscription status, transmission technology and maximum advertised speeds were collected and standardized to comply with valid value lists. In addition, street addresses for new anchor institutions that were not in the original GIS datasets were geocoded using NAVTEQ 2008 Q4 streets data and refined using visual references such as Google satellite photography and street view imagery.

In some cases, standardized transmission technology attribute values were used by the MBI to track uncertain technology categories. These were converted in the final datasets, as shown below, to comply with SBDD valid values.

MBI Technology Values	SBDD Technology Values
1: Unknown	0: Other
42: Cable - DOCSIS Unknown	41: Cable - DOCSIS Other
72: Fixed Wireless - Unknown	70: Fixed Wireless - Unlicensed

Due to the requirement of the technology subtype in the SBDD data transfer model, only those community anchor institutions that had broadband service information (BBSERVICE IS NOT NULL) were included. For anchor institutions that did not have broadband service (BBSERVICE = N), the NULL value in the technology type field was set to 0 (Other) in order to get the records to load.