HOOTENANNY: WEB ENABELED GEOSPATIAL VECTOR-DATA CONFLATION AND MAP GENERATION

Roberto Canavosio-Zuzelski*, Jason Surratt*, Drew Bower*, Joseph Governski*, Matthew Sorenson*

***Contractor for the National Geospatial-Intelligence Agency**

NATIONAL GEOSPATIAL-INTELLIGENCE AGENCY

7500 GEOINT Drive Springfield, VA 22150

Abstract

As geographic vector datasets continue to become more commonplace and available to the public, GIS users find they routinely work with maps that can vary in format, coverage, attribute schema, completeness, and accuracy. For example, datasets range from tightly controlled and regulated on the Government side (NSG Standards Registry, 2014; GWG, 2015; USGS National Map, 2015; Ordnance Survey, 2015) to community projects that are more loosely regulated and open for contributions by users with varying backgrounds and cartographic mapping expertise (OpenStreetMap, 2015). Often times the solution is to work with parts of each dataset in order to produce a superior "best-of-breed" conflated map that leverages the best features and information from each of the individual sources. To this end, automated conflation tools and services were developed in the Hootenanny project to provide an open source, standardized, way of conflating foundational features such as roads, buildings, and points-of-interest. This research paper discusses the various services, workflows, conflation and schema alignment strategies, and output formats being developed to facilitate automated conflation.

Keywords: Alignment, Attribute, Conflation, Feature Matching, Map Merging, Vector-to-Vector

INTRODUCTION

Mapping data is a commodity generated by commercial (Google Maps, 2015), Government (NSG Standards Registry, 2014; GWG, 2015; USGS National Map, 2015; Ordnance Survey, 2015) and crowd-sourced methods such as OpenStreetMap (OpenStreetMap, 2015) and Crowdmap (Ushahidi, 2015). The explosive increase in the demand for open data has been driven by the exponential growth of user friendly geographically enabled desktop and mobile applications. The challenge often encountered when leveraging this data is that no single source can provide all of the necessary contextual information in terms of feature geometry or relevant attribution. However, when the desirable qualities of each input source are merged together in a common schema and format, one is able to build a superior "best-of-breed" dataset that may serve as a gold standard.

Approved for Public Release, 15-333
Corresponding Author: Roberto Canavosio-Zuzelski, canavosr@nga.mil

This capability of combining multiple spatial data sources is commonly referred to as *conflation* and is generally limited to a small subset of available commercial desktop GIS packages. Conflation of maps refers to a combining of two digital map files to produce a third map file which is better than each of the component source maps (Saalfeld, 1985). A conflated dataset is critical to making a better map and serves as the foundation to providing more effective geospatial analytics, such as coverage analysis, change detection, geocoding, and routing.

The task of selecting which dataset(s) to use and merge becomes a daunting process often requiring a case-by-case evaluation where analysts are forced to resort to multiple ad-hoc conflation techniques (automated and manual) which quickly becomes labor intensive and stove-piped. As such, multiple conflation efforts occur in isolated projects by numerous users for specific purposes with varying workflows, rules, schemas, formats, and data quality. This type of approach creates inefficiencies and lacks standardization which is required for large scale organizational use, maintenance, and development activities.

To this end, *Hootenanny* was developed to provide an open source standards based approach to geospatial vector-data conflation. Hootenanny is designed to facilitate automated and semi-automated conflation of critical Foundation GEOINT features in the topographic domain, namely roads (polylines), buildings (polygons), and points-of-interest (POI's) (points). Conflation happens at the dataset level, where the user's workflow determines the best *reference* dataset and *source* content, geometry and attributes, to transfer to the output map. The input data must be normalized to allow processing and matching of features and attributes from different schemas. Hootenanny internal processing leverages the key value pair structure of OpenStreetMap (OSM) for improved utility and applicability to broader user groups, e.g. normalized attributes can be used to aid in feature matching and OSM's free tagging system allows the map to include an unlimited number of attributes describing each feature (OpenStreetMap/Map Features, 2015).

Historically, conflation applications have been specialized desktop tools which required a niche expertise with reoccurring licensing requirements and difficult to customize or add functionality. However, open source software provides an attractive business model where the user community can contribute, customize, share, and help maintain the software in an interactive environment. For this reason, Hootenanny is being developed under the open source General Public License (GPL) and will be hosted on the National Geospatial-Intelligence Agency's (NGA) GitHub website (NGA GitHub, 2015).

A REST API is in place to connect the web browser based User Interface (UI) with the core conflation algorithms and database. The translation and conflation operations are also exposed through an Open Geospatial Consortium (OGC) Web Processing Service (WPS) and the resulting vector data is accessible via a Web Feature Service (WFS) for additional open interoperability (OGC, 2015). This allows for multiple users to perform conflation in a standardized manner on a system that can easily be upgraded, deployed, and maintained at the enterprise level.

HOOTENANNY APPROACH AND ASSUMPTIONS

Hootenanny is designed to perform fully or semi-automated conflation of two vector data sets using either a simple web user interface or an advanced command line utility. The primary function of Hootenanny is to take two input files and produce a single conflated output file. This conflation

process can be repeated in pairwise fashion with multiple iterations, ultimately providing users with the ability to conflate as many datasets as needed to produce a superior dataset with the most complete geometry and attribution.

In order to provide the user with a friendly intuitive conflation experience, Hootenanny was built upon the open source Mapbox iD Editor (Mapbox, 2015). This offers several benefits including its open license that allows users to customize and add functionality, along with an in-depth editing capability that was originally designed for interactive editing of OSM features. Traditionally, map conflation was achieved in either a semi-automated or manual method. Hootenanny's focus is on providing both fully and semi automated functionalities, along with providing iD for editing and conflict resolution of the automated solution.

Hootenanny offers several types of conflation, including *reference* (vertical), *cookie-cutter* (horizontal), and *average*. Reference conflation is where the user specifies one of the input datasets as the "reference" and the other as the "source". Traditionally, the reference has the best geometry and attributes that will be carried forward to the conflated map. The source usually has updated geometry and/or attributes that would add value to the reference and make it more complete. Hootenanny allows the user to select different geometry and attribute references because one dataset may not serve both purposes. Cookie-cutter is an implementation of horizontal conflation where one of the input sources is superior, in every way, to the other. In this case, the less detailed source is cut-out and replaced by the superior one and conflated around the edges to ensure a seamless transition. Applications of cookie-cutter have been applied in urban areas where a dense local level dataset is conflated with a sparser regional level map. Average refers to computing a weighted average location based on the geospatial accuracy of the two matching features.

Matched and unmatched features refer to the geometry that are the same and different, respectively, between the two inputs. In general, the way in which these features are handled and transferred to the conflated map is dependent upon feature and what type of conflation is being performed. For example, in reference conflation matched road geometry from the reference will persist, but unmatched or new roads from the source dataset will also be carried forward. However, for building conflation matched polygons will always select the most detailed representation of the shape to carry forward, independent of what dataset is specified as reference or source. The idea here is that the most detailed building is considered the best geometry.

POI conflation refers to merging features that exist as point geometries, e.g. restaurants, taverns, café, shops, offices, buildings, tourist attractions, etc. Hootenanny uses both attributes (name, type) and proximity (distance) to help determine whether two points are a match. With POI's the location from the user defined reference dataset will persist in the conflated map.

Similarly, any existing feature attributes are transferred to the conflated map based on whether it is matched or unmatched. For matched features, the matched attributes are transferred based on a user decision of what *attribute reference* dataset to use, while any unmatched or new attributes from a matched source feature will be transferred to the corresponding reference feature in the conflated map. Unmatched or new features that exist in the source and not in the reference will be transferred in whole to the conflated map.

HOOTENANNY CONFLATION WORKFLOW

The general case of the Hootenanny conflation workflow is shown in Figure 1 and depicts the high-level steps necessary to conflate data and generate an output map in Hootenanny. It is important for the user to understand these functions as each have implications on the conflated results. The squares represent a specific conflation task, while the oval canisters represent a database function. The workflow is described as follows:

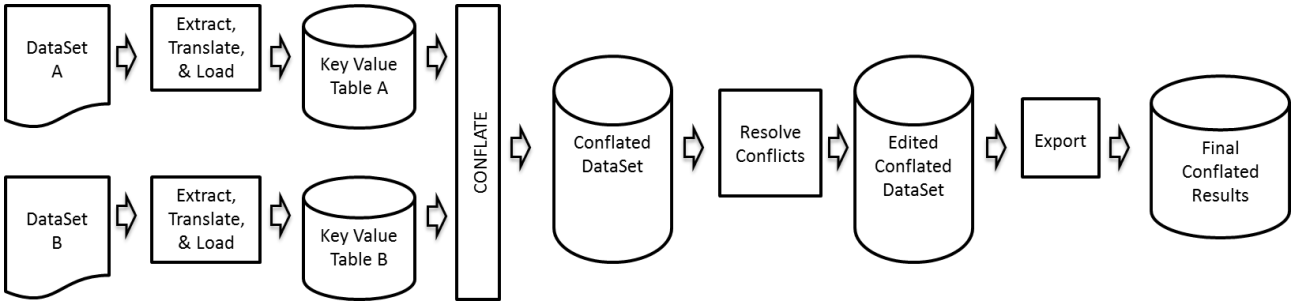

Figure 1. Hootenanny Conflation Workflow

Datasets - The datasets refer to the reference and source inputs that will be imported into Hootenanny and used for conflation. Several formats are supported as input sources including ESRI Shapefile, File Geodatabase, and native OSM. A zip file of multiple Shapefiles is also supported. The user should be familiar with the inputs and be ready to identify the reference and source geometry, attribute reference, input schemas, and have a general understanding of spatial misalignment for parameter tuning.

Extract, Translate, and Load (ETL) - Hootenanny is built upon the OSM key value pair tag concept and PostgreSQL database structure. Additional work was done to extend the base OSM tag structure to accommodate customer attributes that did not directly map to an OSM tag. Therefore, data going into Hootenanny has to be translated from the input schema to OSM. Currently, there are two options for translation, an automated process that supports some pre-defined standard customer schemas like: Topographic Data Store (TDS), Multi-National Geospatial Co-Production Program (MGCP), Urban Tactical Planner (UTP), Urban Feature Data (UFD), and NAVTEQ. Additionally, for datasets that do not fit into one of these pre-defined formats, a semi-automated *Custom Translation* capability is provided where the user maps attributes from the input schema to the corresponding tag in OSM.

Conflation – Conflation algorithms do the heavy lifting to provide the automated conflation solution. Hootenanny currently provides algorithms to support roads (polylines), buildings (polygons), and POI's (points) conflation. Roads include most linear features found in transportation including highways, cart tracks, trails, and bridges and tunnels in some cases. Automated conflation only

accounts for a portion of the overall conflation process, albeit we strive for a large portion on the order of 80-90%, there is a semi-automated and manual process (*Conflict Resolution*) left to bring the conflated dataset to 100% complete. As such, the automated algorithms in Hootenanny are based on training datasets that describe the common problems a user would encounter when performing manual conflation, and therefore keep track of matching performance and deviation from a given matching threshold. The advantage here is that Hootenanny can flag suspect or ambiguous matching performance for the user to fix in the Conflict Resolution phase. Generic Line Conflation is under current development and provides command line support for other linear features like rivers, streams, walls, and shorelines through the Java Script (JS) Application Programmers Interface (API).

Conflict Resolution – As mentioned above, automated conflation only provides part of the overall solution. Therefore, the user must play a role in resolving ambiguous feature matching (conflicts) and editing any geometry artifacts left during the automated conflation. Hootenanny leverages the Mapbox iD Editor software to provide the user a streamlined intuitive method to resolve these conflicts. Hootenanny cues up conflicts and drives the user to each conflict to make the appropriate geometry or attribute fix. Additionally, the user has the option to edit other features they may come across that need correcting while working through the conflict cue. For users that do not want to perform this manual editing or want to accept the automated solution in full, there is the option to *accept all* or *discard all* conflicts. This becomes a user choice and ultimately a risk assessment between accepting geometry and attribute errors resulting from automated algorithms versus the use purpose of the end product.

Export – Export refers to getting a conflated dataset out of Hootenanny (OSM) and into a different format. The Export functionality support outputs as ESRI Shapefile, File Geodatabase, WFS, and native OSM formats. Similar to ETL, Export has to translate the Hootenanny internal OSM schema to an approved standard schema. Currently, the user has the option to output conflated datasets in TDS, MGCP, or OSM standard schemas. At this point, mapping specifications (NSG Standards Registry, 2014; GWG, 2015) can be applied to the standardized output file to produce a stylized mapping product. It's worth mentioning that the potential for attribute loss may exist when input attributes do not directly map to a corresponding spot in a non OSM export specifications.

Data stewardship refers to the business rules, processes and tradecrafts used by Geographic Information System (GIS) professionals to accomplish a certain task, provide a service, or generate a product. As such, data stewardship has an impact on the conflation process and resulting product. For example, the data steward should perform a pre-analysis of the datasets to determine if conflation would in fact be beneficial, dataset restrictions or license considerations, what source is best suited for the reference, conflation logic, or output format/schema. In some cases, conflation might not be the answer or provide only marginal benefit depending on issues identified in pre-analysis or expected output.

SCREENSHOTS

The following figures provide a visual overview of Hootenanny as implemented in the Mapbox iD software. Datasets are courtesy of Washington D.C. GIS Roads (D.C. GIS, 2015) and TIGER/Line Shapefiles (US Census Bureau, 2015). Note, background imagery services are available in Tile Map Service (TMS) format (TMS Specification, 2012), but were not used here for convenience.

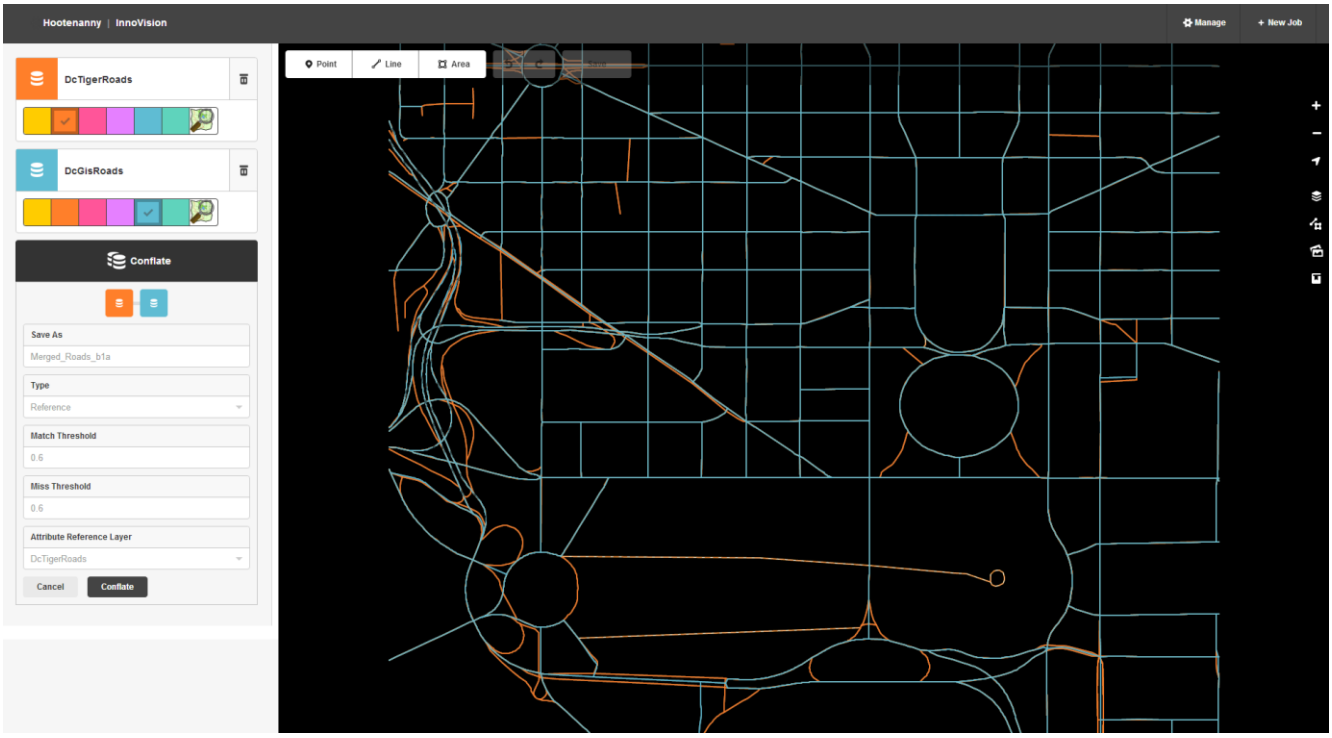

Figure 2. Sample Input Datasets: Reference (Orange) and Source (Blue)

Figure 2 represents two input datasets prior to conflation. For matched roads the geometry from the reference dataset will be transferred to the conflated map, along with any new attribution contained on the source road. Any unmatched/new roads in the either source will also be pushed forward.

Figure 3 shows a sample of the Hootenanny Conflict Resolution capability. The ambiguous match is shown in pink along with some options to accept or discard the conflict. Alternatively, the user has the option to edit the feature with the standard iD editing tool pallet, which appears when the feature is selected. Some basic navigational abilities exist to scroll through the conflict cue.

Figure 4 shows a sample conflated dataset. Note, the green roads signify their presence in the conflated dataset.

Figure 3. Sample Conflict Resolution and Editing Process

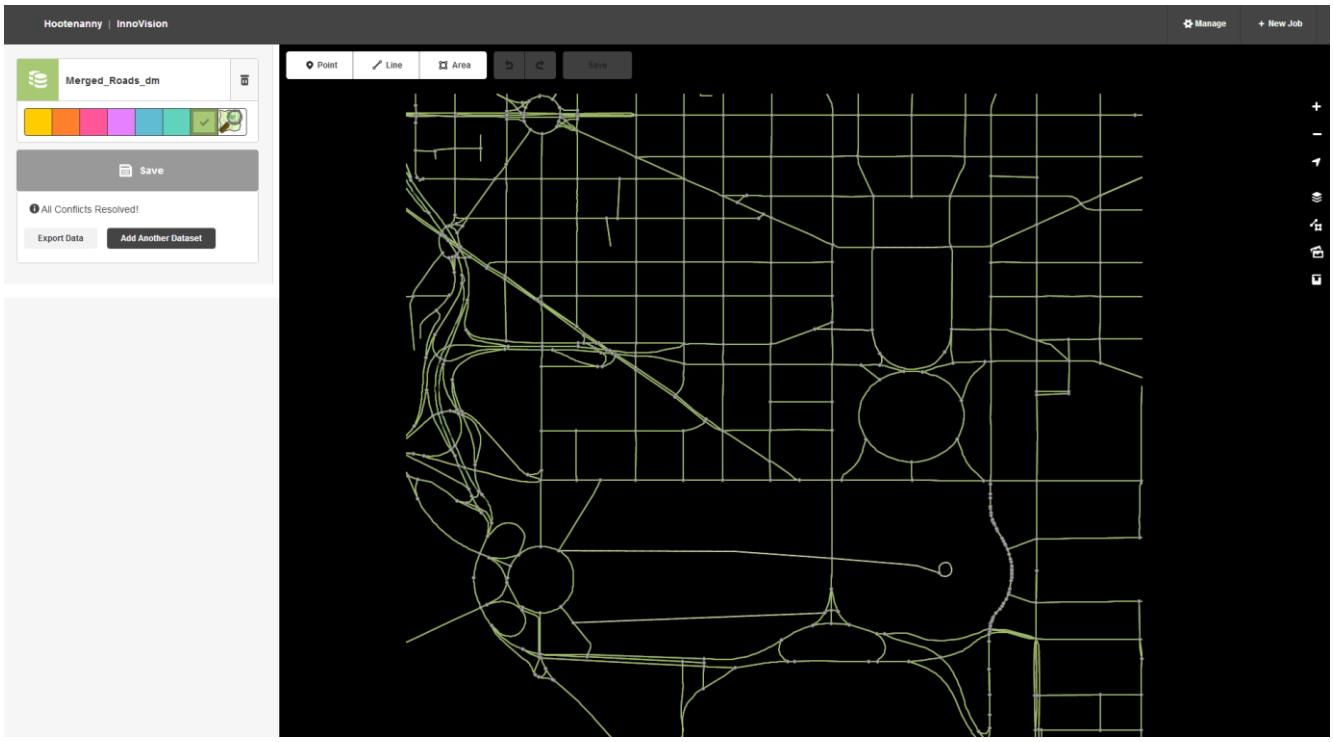

Figure 4. Sample Conflated Dataset

CONCLUSIONS

The challenge often faced when working between multiple new and existing sources of geospatial data is that no single source provides all of the necessary contextual information, e.g. updated geometry, coverage, or relevant attribution. However, when the desirable qualities of multiple sources are conflated together in a common schema and format, one is able to build a superior "best-of-breed" dataset that may serve as a gold standard.

Hootenanny was developed to provide an open source standards based solution to geospatial vector-data conflation. The open source model is attractive because it gives interested users the ability to contribute to further development of Hootenanny, while providing flexibility and cost savings associated with site licensing requirements. Similarly, Hootenanny was developed around open OGC compliant web services using the Mapbox iD Editor software to provide a standardized approach to conflation. The advantage offers control of both the conflation business rules and output product, where it can be integrated into official business processes and holdings.

Hootenanny is designed to facilitate automated and semi-automated conflation of critical Foundation GEOINT features in the topographic domain, namely roads (polylines), buildings (polygons), and POI's (points), while providing a Conflict Resolution and editing capability to address any geometry and attribute problems that arise during the conflation process.

Hootenanny was built upon the OSM key value pair attribute concept and PostgreSQL database. Therefore, input sources must be mapped to Hootenanny's internal OSM schema. The advantage here is that once attributes are mapped they can be used to aid in feature matching and output into standardized specifications. Translation for supported schemas like TDS, MGCP, UTP, UFD, NAVTEQ, and OSM is automated, while a Custom Translation tool is provided that gives the user the flexibility to map additional schemas into the OSM format. Hootenanny supports ESRI Shapefile, File Geodatabase, and OSM formats as input and output formats, while adding OGC compliant WFS as an export option. The Hootenanny software will be licensed as open source GPL and hosted on the NGA GitHub website.

FUTURE WORK

Hootenanny is under current development and the open source model offers interested users the ability to contribute to Hootenanny's future development. The most logical starting point would be to extend input and output translation capabilities. Currently, several standard NGA formats make up the lion share of our work; however the Custom Translation tool offers users the ability to create their own translation by interactively mapping an input source to OSM. These translations could be provided back to the user community. Secondly, support for additional features such as railways, rivers, streams, utilities, etc. are constantly being considered and functionality added over time. Users with a stake in a particular feature could feasibly contribute to the development of a specific conflation

algorithm. Similarly, the work on generic conflation routines is ongoing and hopes to develop customizable business rules that users can use to conflate multiple feature sets.

Point-to-polygon conflation capabilities have utility. Some datasets may represent a particular feature as a point in one dataset and a polygon in another, e.g. buildings in different scaled maps, therefore it is desirable to further development in this area. Finally, Hootenanny is specifically focused on vector-to-vector conflation and addresses a significant part of a much broader geospatial data management requirement. Additional exploration into related areas to include vector-to-image conflation, data quality and topology compliance, and data provenance are required.

REFERENCES

Crowdmap website: http://www.ushahidi.com/product/crowdmap/ accessed March 24, 2015.

District of Columbia (D.C.) GIS Data Clearing House Catalog website: http://dcatlas.dcgis.dc.gov/catalog/ accessed March 30, 2015.

Geospatial-Intelligence Standards Working Group (GWG) website: http://www.gwg.nga.mil/disr_standards.php accessed March 26, 2015.

Google Maps website: https://www.google.com/maps accessed March 24, 2015.

Lynch, M. and A. Saalfeld, Bureau of Census, Statistical Research Division, 1985. "Conflation: Automated Map Compilation, A Video Game Approach".

Mapbox iD Editor website: https://www.mapbox.com/blog/new-map-editor-launches-openstreetmap/ accessed March 24, 2015.

National Geospatial-Intelligence Agency (NGA) GitHub website: https://github.com/ngageoint accessed March 26, 2015.

National System for Geospatial-Intelligence (NSG), 2014, website: https://nsgreg.nga.mil/JESC-approved.jsp accessed March 26, 2015.

Open Geospatial Consortium (OGC), Implementation Standards website: http://www.opengeospatial.org/standards/is accessed March 25, 2015.

OpenStreetMap website: http://www.openstreetmap.org accessed March 6, 2015.

OpenStreetMap/Map Features website: http://wiki.openstreetmap.org/wiki/Map_Features accessed March 26, 2015.

Ordnance Survey website*:* https://www.ordnancesurvey.co.uk/ accessed March 6, 2015.

Tile Map Service (TMS) Specification, 2012, website:
http://wiki.osgeo.org/wiki/Tile_Map_Service_Specification accessed April 7, 2015.

United States Census Bureau, Washington D.C. TIGER/Line Shapefiles website:
https://www.census.gov/geo/maps-data/data/tiger-line.html accessed March 30, 2015.

United State Department of Interior, United States Geological Survey (USGS), The National Map website: http://nationalmap.gov/index.html accessed March 6, 2015.