OpenStreetMap Collaborative Prototype, Phase One

By Eric B. Wolf, Greg D. Matthews, Kevin McNinch, and Barbara S. Poore

Open-File Report 2011–1136

U.S. Department of the Interior
U.S. Geological Survey

U.S. Department of the Interior
KEN SALAZAR, Secretary

U.S. Geological Survey
Marcia K. McNutt, Director

U.S. Geological Survey, Reston, Virginia 2011

For product and ordering information:
World Wide Web: http://www.usgs.gov/pubprod
Telephone: 1-888-ASK-USGS

For more information on the USGS—the Federal source for science about the Earth,
its natural and living resources, natural hazards, and the environment:
World Wide Web: http://www.usgs.gov
Telephone: 1-888-ASK-USGS

Suggested citation:
Wolf, E.B., Matthews, G.D., McNinch, K., and Poore, B.S., 2011, OpenStreetMap Collaborative Prototype, Phase 1: U.S. Geological Survey Open-File Report 2011-1136, 23 p.

Contents

Figures

Tables

OpenStreetMap Collaborative Prototype, Phase One

By Eric B. Wolf, Greg D. Matthews, Kevin McNinch, and Barbara S. Poore

Abstract

Phase One of the OpenStreetMap Collaborative Prototype (OSMCP) attempts to determine if the open source software developed for the OpenStreetMap (OSM, *http://www.openstreetmap.org*) can be used for data contributions and improvements that meet or exceed the requirements for integration into *The National Map* (*http://www.nationalmap.gov*). OpenStreetMap Collaborative Prototype Phase One focused on road data aggregated at the state level by the Kansas Data Access and Support Center (DASC). Road data from the DASC were loaded into a system hosted by the U.S. Geological Survey (USGS) National Geospatial Technical Operations Center (NGTOC) in Rolla, Missouri. U.S. Geological Survey editing specifications were developed by NGTOC personnel (J. Walters and G. Matthews, USGS, unpub. report, 2010). Interstate and U.S. Highways in the dataset were edited to the specifications by NGTOC personnel while State roads were edited by DASC personnel. Resulting data were successfully improved to meet standards for *The National Map* once the system and specifications were in place. The OSM software proved effective in providing a usable platform for collaborative data editing.

Background

The National Map (*http://nationalmap.gov/*) is a collaborative effort among the USGS and other Federal, State, and local partners to deliver topographic information for the Nation. Geographic information content includes orthoimagery (aerial photographs), elevation, geographic names, hydrography, boundaries, transportation (including roads), structures, and land cover. *The National Map* is a significant contribution to the National Spatial Data Infrastructure (NSDI) and currently is being transformed to better serve the geospatial community by providing high quality, integrated geospatial data and improved products and services accessible through the Internet, including new generation digital topographic maps.

The USGS is responding to changes in data production and sharing on the Internet by exploring collaborations with non-traditional customers and data producers (Sugarbaker and others, 2009). The Geospatial Web (Scharl and Tochterman, 2007), or the merging of location with content information on the Internet, has democratized online mapping. Civilian access to Global Positioning Satellite (GPS) signals, the availability of application programming interfaces (API) that enable the mashup of geospatial data from disparate sources onto a map-based platform, the increasing maturity of open-source geospatial software, and the rapid spread of geo-enabled mobile devices have made it easier for the general public to access and use geospatial data online. These technical changes have fostered a culture of collaborative online mapping by users who are not Geographic Information Science (GIS) professionals. The trend of mapping data collection and use by non-professionals has been referred to as volunteered

geographic information (VGI) (Goodchild, 2007). Rather than being passive recipients of maps and data from official sources, these users are capable of producing their own data and innovative geo-technologies.

Historically, the USGS has had programs for volunteers to contribute data to USGS topographic maps. In the 1990s, volunteers from the Earth Science Corps "adopted a quad," annotating potential revisions to the paper maps and mailing them to the USGS. In the Internet era, the program was renamed *The National Map* Corps. The focus was on the collection and editing of structures information; including providing location information for buildings such as fire stations, schools, and post offices. Early on, citizens submitted the names and GPS coordinates of structures through a web portal. A later system based on ESRI ArcIMS used heads-up digitizing as the data capture methodology. This program was put on hiatus due to budgetary constraints.

To reevaluate the viability of a volunteer mapping program, the USGS held a workshop on VGI in January 2010 (USGS VGI Workshop) (*http://cegis.usgs.gov/vgi/index.html*). The workshop brought together representatives from organizations that had experience using data collected by volunteers. These ranged from citizen science efforts (for example, NOAA's Cooperative Observer Program) to online crowdsourced identification of historic photographs (for example, Library of Congress), commercial firms that use volunteers with sensors in their cars to update street networks (for example, TeleAtlas), and a non-profit building an online map of the world created entirely by volunteers (OpenStreetMap, *http://www.openstreetmap.org*).

Problem Statement

Even though the USGS VGI Workshop demonstrated that there are a growing number of programs in the public, private, and non-profit sectors that use information supplied by volunteers, there are numerous issues and questions that need to be explored in order to evaluate the usefulness and effectiveness of VGI for use by the USGS in the creation of *The National Map*. These issues are:

- How accurate are the data?

  One of the most urgent questions emerging from the workshop was related to data accuracy: how accurate are volunteered data? Some citizen science programs such as the Audubon Christmas Bird Count have been in operation for decades and their record of accuracy is well established (Cohn, 2008); however, these programs have highly structured protocols designed by professional scientists and provide intensive training for volunteers. The case for volunteered geographic information is more problematic. Volunteered Geographic Information is a fairly new phenomenon and has many variations (Goodchild, 2007).

- What types of tasks are well-suited to volunteer data collection?

  Not all spatial data are well suited for volunteer data collection due to different requirements for accuracy, completeness, and currentness. There are also complexity and security considerations, as well as differing relationships between data sets. Feature geometry that can be verified by visible identification in aerial imagery or ground truthing with GPS, such as buildings and roads, are better suited to VGI data collection

than more abstract features, such as political boundaries. Attribution also impacts how well certain feature types lend themselves to volunteer data collection. Features with complex attribution requiring significant quality control efforts are not as well suited to volunteer efforts as features with simple attribution.

Volunteered Geographic Information methods remain untested within the production systems of national mapping agencies that have particular requirements for data quality control and review. Non-spatial crowdsourced sites, such as Wikipedia, use volunteers specifically assigned to moderate and review volunteer contributions. Current crowdsourced map efforts, such as OSM, do not have a specific structure of moderation in place to review data submitted by volunteers. Instead, they rely on the entire volunteer community to correct erroneous data or identify cases of vandalism. This project will assay the possibilities of whether volunteers can be used not only to collect data, but also to review, test, and certify data for national mapping agencies.

- What motivates volunteers?

  Motivating volunteers and maintaining their interest was discussed at the USGS VGI Workshop as a main topic of consideration for creating a volunteer data collection program. It was noted that volunteers need to have a tangible reason for contributing their time, knowledge, and effort. Volunteer motivation could include being involved in a community that represents something larger than themselves and seeking to give back to that community, a hobby, an interest in a particular geography and a desire to improve its representation on an open-source map, seeking return of data for use in their own projects, rewards and recognition, or simply curiosity.

- What is the best way to structure such programs and provide incentives?

  The goal of establishing a successful volunteer program can be accomplished with many different project structures and organizations. At the USGS VGI Workshop many existing volunteer data collection efforts were discussed with very different project structures: Wikipedia, Audubon Christmas Bird Count, OSM, and so forth. The USGS is in the process of evaluating these different approaches to establish which would be suitable for a volunteer program to utilize when contributing to *The National Map*.

- How can volunteered data be integrated with data produced by professional organizations?

  The National Map has particular needs for authoritative data, whereas a project such as OpenSteetMap does not claim to contain authoritative data. How to certify volunteer data and integrate them into *The National Map* are significant problems that have not previously been researched in the geospatial community.

- What are the cost/benefit tradeoffs of using volunteered data?

  The usefulness, accuracy, and completeness of volunteer-collected data are investigated along with associated costs. The results of this cost/benefit analysis will be a factor in determining the future of the program.

- How sustainable are these volunteer programs?

  Keeping volunteer contributors engaged for the long-term was an issue discussed in detail at the USGS VGI Workshop. The consensus was that it is easier to get volunteers to create new data than to update, maintain, and validate data. How successful volunteer projects such as Wikipedia and OSM have maintained volunteer contributions remains to be studied.

- How well suited are existing web-based collection systems for potential USGS VGI efforts?

  An evaluation of existing systems for user-contributed data revealed that the OSM software stack has actually delivered real world success and looks promising.

OSMCP Project Approach

USGS set up a research project, the OSMCP, to address some of the questions raised during the USGS VGI Workshop. Phase One was initiated in Fiscal Year 2010 (FY2010) and was limited to evaluating the suitability of existing web-based collection systems for USGS VGI efforts. Phase One was targeted at creating a system for collaborative editing with USGS partners and did not incorporate volunteer contributions. Later phases will build on this initial phase and expand the scope to amateur volunteer contributors.

A prototype was built using open source software that supports the OSM system. This software was replicated, deployed, and customized on a USGS system. Data developed by a USGS partner organization were uploaded into the system and edited by both the USGS and the partner in a web-based environment. The following components comprised the Phase One project:

- Selecting and deploying hardware and software.
- Developing documentation, including specifications for collaborative editing of roads data for *The National Map*.
- Selecting a partner (and partner data) for collaborative editing.
- Loading and editing partner data in the OSM system environment.
- Evaluating the data resulting from collaborative editing.

The OSMCP was to start with a prototype system that could be used to gain experience in collaborative editing and set up an infrastructure to support future project phases. Subsequent phases of the OSMCP will build upon Phase One and continue to explore the use of VGI at the USGS. There are plans for future phases to include wider user groups, including contributions from volunteer groups. Phase One is meant to be an early milestone in the process of increasing system capability in order to address additional questions raised at the USGS VGI Workshop related to accuracy, volunteer motivation, and cost/benefit.

Materials and Methods

Phase One of the project included exploring the capability of a web-based collaborative editing tool to facilitate cross-agency co-editing and data integration of the edited data into The National Map.

Selecting and Deploying Software and Hardware

At the USGS VGI Workshop various software systems for collecting user contributed geographic information were discussed. Factors considered in the selection of software to support the project included:

- User-friendly interface that can be used effectively with a minimum of training or user documentation;
- Low cost and compatibility with existing USGS hardware;
- Does not require a user license by the data contributor;
- Can be supported by existing systems support staff;
- Supports the basic functions for large numbers of simultaneous users of:
    1. Web-based editing via a browser including Internet Explorer, Mozilla Firefox, and Apple Safari.
    2. User creation and editing of spatial features including the position and attributes through heads-up digitizing.
    3. Provides an orthorectified image as a backdrop and source for heads-up digitizing.
    4. Customization of feature types and attributes without programming.
    5. Storage or transfer of data to a database easily exported to ESRI ArcGIS.

The software architecture options considered were:

1. USGS direct contributors to OSM (*http://www.openstreetmap.org*)
2. USGS download and use OSM data as part of *The National Map*.
3. Modify the USGS National Hydrography Dataset (NHD) Stewardship Editing Tools. (*http://webhosts.cr.usgs.gov/steward/scripts/st2_software.pl*)
4. Modify Alabama NHD Web Edit Tool (WET). (*http://nhd-wet.alabama.gov/*)
5. Develop a new custom web application in ESRI ArcGIS Server or the open source GeoServer web map server software. (*http://www.geoserver.org*)
6. Use the OSM Software Architecture with a public domain database.

One option considered in the USGS VGI Workshop was to use USGS staff to encourage the collection of data useful for *The National Map* within the OSM platform. This option was not pursued due to data licensing issues. Data in OSM are licensed under a Creative Commons Share-Alike license (*http://www.openstreetmap.org/copyright*), which requires anyone using the data to provide improvements to the community under the same license. Most USGS geospatial data are in the public domain. There are no restrictions on use or redistribution of the data. This difference prevents the USGS from harvesting data collected through the OSM system for use as part of public domain data sources, including *The National Map*.

Another alternative considered for an USGS VGI platform was to leverage the software developed for the NHD Stewardship. The NHD is the hydrographic layer of *The National Map* and is maintained through a network of steward-contributors. These stewards receive training and support and are encouraged to sign stewardship agreements with USGS. The software developed for processing steward changes to the NHD requires an ESRI ArcGIS software license. The NHD update software is designed specifically around the data types in NHD. The NHD uses a topological network representation. The software is designed specifically to provide

editing of that kind of representation. This option was not pursued because of the requirement of having an ESRI ArcGIS software license and the complexity of the ESRI ArcGIS suite compared to other available web editing tools.

In addition to the tools created by the USGS for user contributions to the NHD, the State of Alabama is developing a web-based mark-up and editing tool for NHD. The Alabama Web Edit Tool (WET) is built on ESRI ArcGIS Server and was under development at the time of the USGS VGI Workshop. The WET tool requires a software login, and the edits are not incorporated directly back into the dataset until they are reviewed by an NHD steward. Because of this workflow limitation and the fact that the tool was still in development when the USGS VGI investigation began, this option was not pursued.

A fourth option included building a custom web-application to support a USGS VGI effort using one of many commercially available or open-source geoprocessing and mapping servers, such as ESRI ArcGIS Server or the open source GeoServer. The high cost of developing new editing software prohibited that option.

The final option considered (the one that was ultimately adopted) involved using the OSM software replicated on USGS systems. OSM is one of the most ambitious efforts at producing a basemap of the world through volunteer contributions. The OSM community, which was well represented at the USGS VGI Workshop, proposed that the USGS utilize the open-source software infrastructure that makes up the system. The OSM software architecture is well documented, easy to configure, requires no additional software licenses for users of the system, and has a large community of contributors, open-source developers, and forum and wiki contributors that would be available for technical support or questions. The OSM editing user interface is easy to use, and was evaluated by USGS staff to verify that it supported all required editing functions needed to allow data to be incorporated into The National Map. The OSM software is easy to modify, configure, and customize to suit the project. The decision was made to use the OSM software in Phase One of this project due to its ease of use, open source licensing, potential support, and the leadership the OSM community has shown using volunteer mappers on the web.

The OSMCP Software System

The OSM community has developed substantial software architecture (OpenStreetMap, 2010) to collect user-generated spatial information. The software is free; open source-licensed; and utilizes the Linux operating system, Apache web server, Ruby on Rails web architecture, and PostgreSQL relational database (see Figure 1). A variety of different applications have been developed to facilitate user contributions including the web-browser based Potlatch and Potlatch 2 interfaces, the desktop applications JOSM and Merkaartor, as well as several iPhone applications. The OSM software does not impose any data structures other than nodes (points), ways (lines, polygons), and relations. Further, feature attribution in OSM is handled via ad hoc tags. No feature ontology is imposed on data collected using OSM.

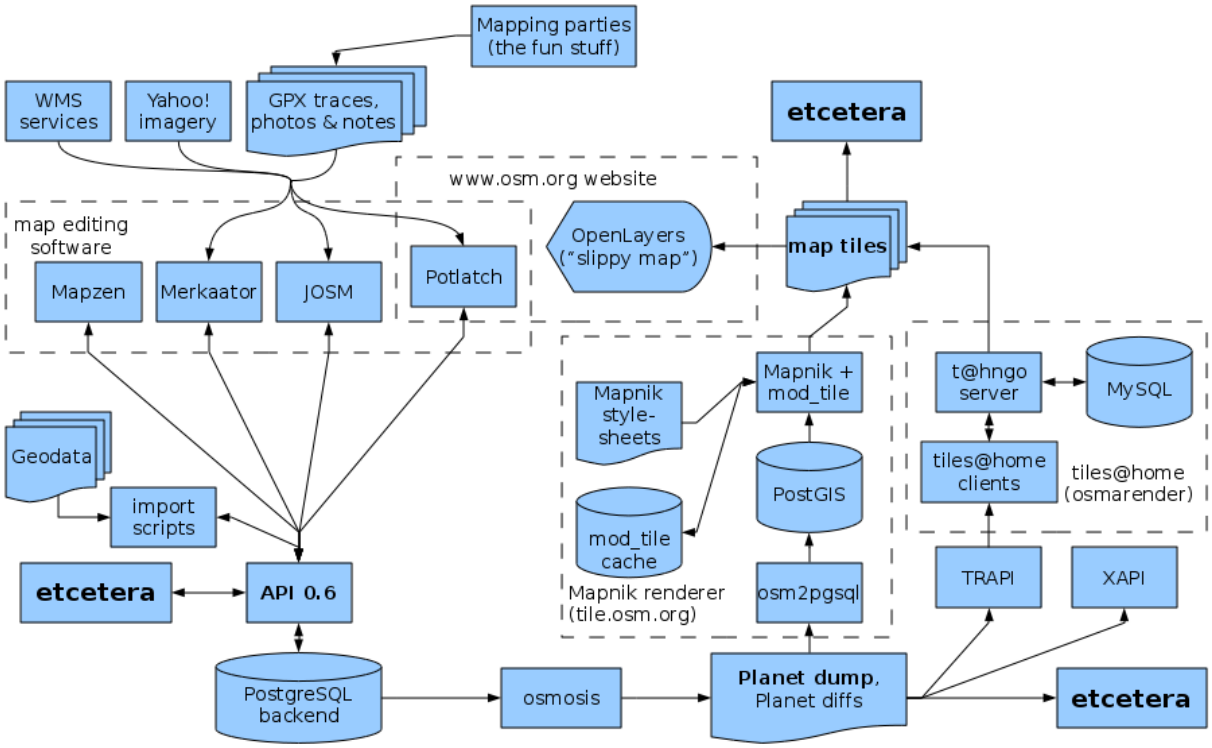

Figure 1. Components of the OSM systems architecture. (OpenStreetMap, 2010)

The OSM system architecture is based on a distributed model consisting of a PostgreSQL database backend with a middle tier, called API 0.6, developed in Ruby on Rails (Figure 1). Large data imports and exports can be accomplished via a direct load with a tool called osmosis. Editing and smaller data inputs are provided via the API 0.6 interface with tools like JOSM, Merkaartor and the browser-based Potlatch editor.

In Phase One we used a server physically located at the USGS NGTOC facility in Rolla, Missouri connected to the internet via a publicly available address: *navigator.er.usgs.gov*. This server was placed in the demilitarized zone (DMZ) behind the USGS Enterprise Secure Application Service (eSAS). Computer servers in the DMZ are accessible from the open internet. The eSAS protects servers in the DMZ by filtering web requests. The server had four Intel Xeon 3.6Ghz CPU cores with 6GB of RAM and 650GB of disk space. Ubuntu Server 9.10 "Karmic Koala" 64-bit was the operating system, with Apache 2.2 providing HTTP services.

The OSM web interface was reconfigured for the desired The National Map data schema (Table 1) (*http://nationalmap.gov/transport.html*). The OSM software does not impose any particular data schema or ontology. In the case of the OSM community, the data schema continues to evolve through a collaborative process. One of the main challenges in the OSMCP was to demonstrate that the OSM software could support a pre-determined schema. The Potlatch interface was configured for this schema by editing several text files (see Appendix B).

Table 1. OSM and Best Practices Schema differences.

OSM	Best Practices
highway:primary	highway:USHighway
name:West 6th Street	Full_Street_Name: W 6th ST
ref: US 40-59	Road_Class: 10002
	State:KS
	Surface_Type:99
	US_Route1:US-40
	US_Route2:US-59
	Geodb_oid:162151

The OSM schema was less complex than the Best Practices (BP) schema. For example, the Potlatch editor recognizes the "highway" tag key as a particular type of line representing a part of an automobile road network but does not specifically mean a limited access, high speed road. In the OSM schema, common uses of the highway tag are highway="residential road" and highway="service road".

The more technical modifications focused on the Potlatch editor in Figure 2. Potlatch is an Adobe Flash-based client written in Action Script 1 designed to be run inside the web browser. Vector symbology in Potlatch is controlled via a series of text files (listed in Table 2) stored on the server. The Potlatch files were modified to provide better symbology based on the BP schema. Figure 3 shows the Potlatch interface with vector symbology coded per the schema.

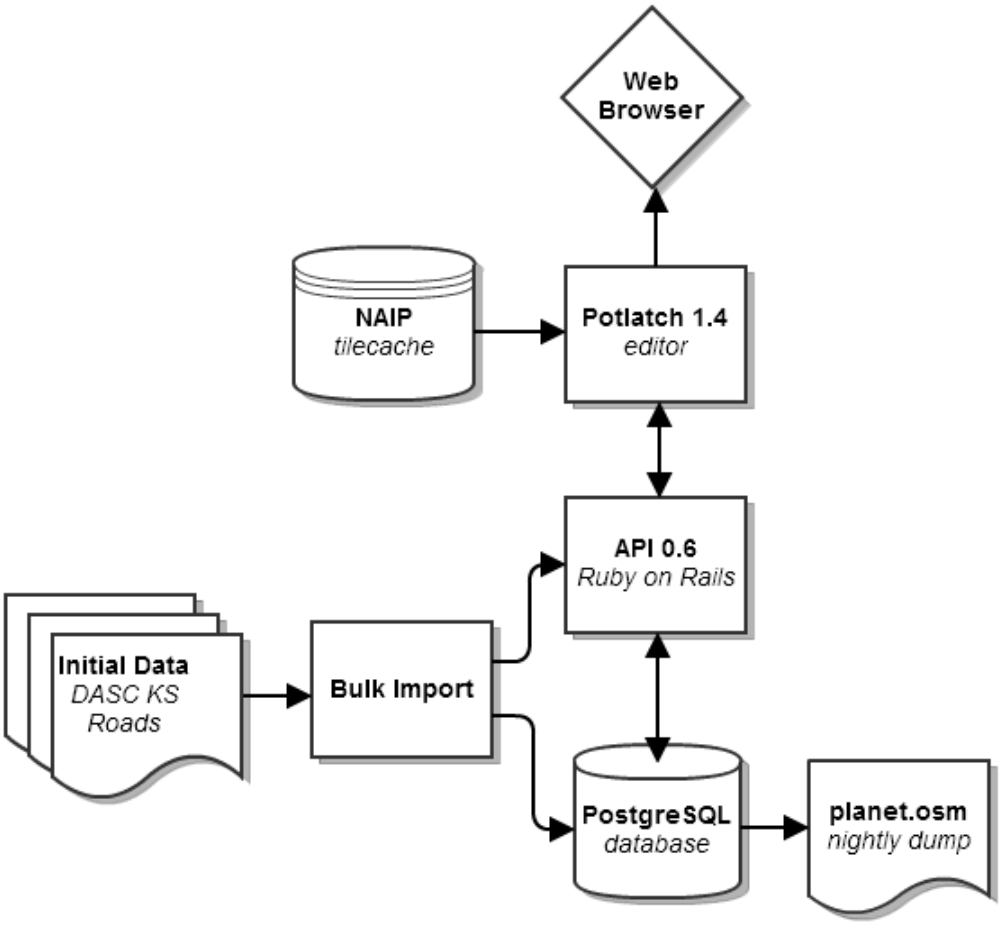

Figure 2. Components of the OSMCP systems architecture (subset of Figure 1).

Table 2. Files controlling Potlatch vector symbology (note British spelling of "colour" is correct).

File	Description
presets.txt	CSV-like list of common way and node key/value pairs
colours.txt	Tab-separated list of fill and stroke colours used when drawing the map
relation_colours.txt	Tab-separated list of highlight colours used when drawing relations on the map
autocomplete.txt	Tab-separated list of keys and values used for the autocomplete menus

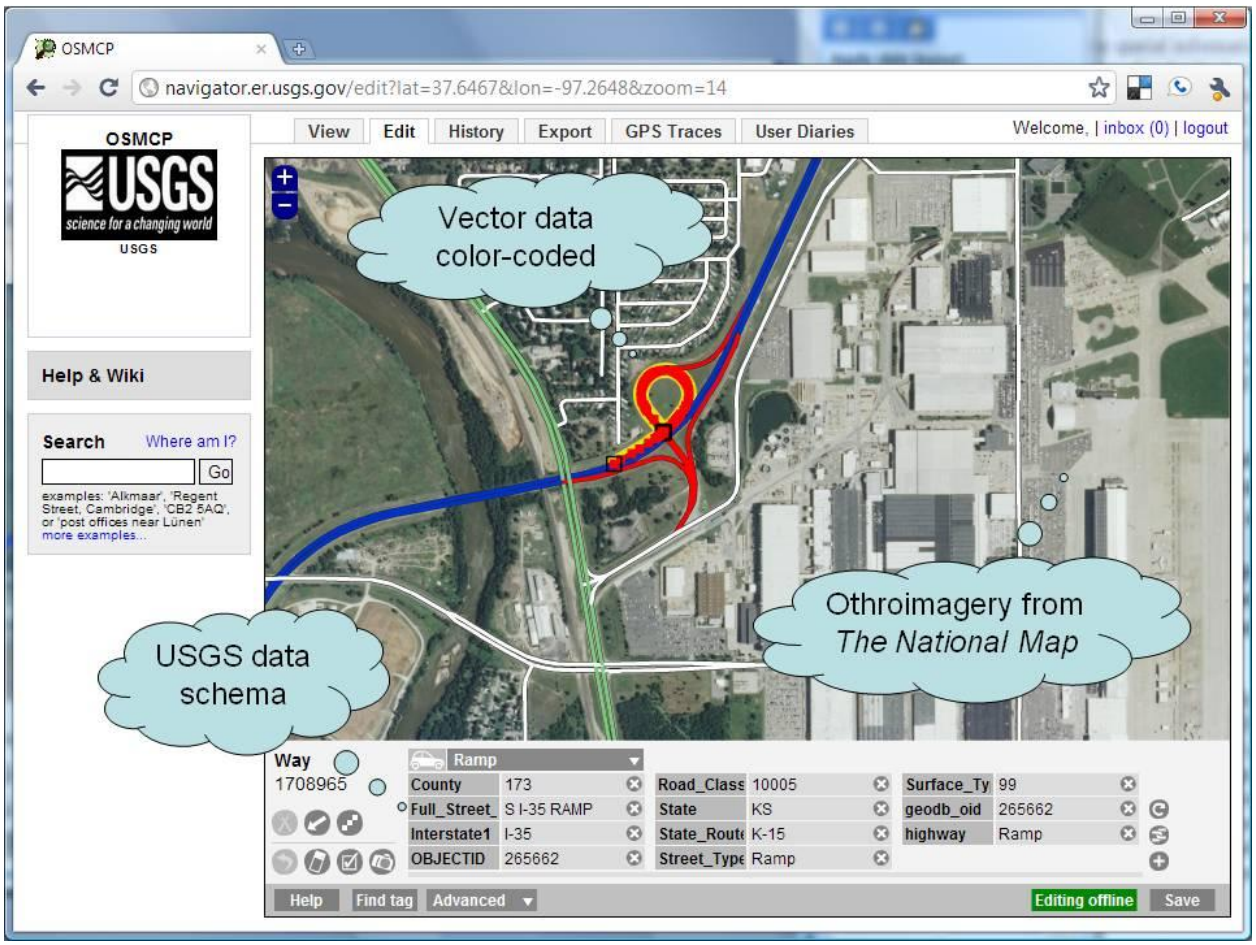

Figure 3. The Potlatch interface with vector data overlaid on Orthoimagery layer from *The National Map*. Symbology is tied to the data schema.

Editing in Potlatch essentially follows a heads-up digitizing practice. Although users are able to upload GPS tracks in a GPX file, the most common practice among OSM contributors in areas with good aerial imagery coverage is to trace over background imagery. On the OSM site (*http://www.openstreetmap.org*), Potlatch displays vector data overlaid on imagery donated by Microsoft Bing Maps that provide world-wide coverage at 0.3m color (Meyer, 2010).

In Phase One of the project, a local tile cache was configured to provide background imagery from the Orthoimagery layer provided by *The National Map* in place of the Microsoft imagery. This follows internal NGTOC data quality assurance practices, using the Orthoimagery layer that is mostly sourced from the National Aerial Imagery Program (NAIP) as the image data source for determining the acceptability of data. Potlatch required imagery to be provided in a Google-style tile cache where in tiles were directly addressed as graphic files served via HTTP. The tiles were extracted from a Web Mapping Service (WMS) provided by *The National Map*: Combined/TNM_Large_Scale_Imagery (MapServer) (see *http://raster.nationalmap.gov/ArcGIS/rest/services/Combined/TNM_Large_Scale_Imagery/MapServer* for metadata). While *The National Map* provides orthoimagery as a tiled service, the imagery provided by this service only provides zoom levels 1–11 (less than approximately

1:220,000 scale) which provides insufficient detail for tracing features in Potlatch. In order to visually identify features in background imagery, Potlatch editing occurs at zoom levels 16–20 (approximately 1:20,000 and larger scale). The Python script Tilecache 2.11 was used to generate the tiles on the server (see Appendix A: Tilecache.cfg). The tile cache script provided both the means to dynamically request tiles and "seed" the cache. Most of the Phase One study area was seeded beforehand, but the process was too slow to complete before the partner data collection effort in OSMCP Phase One.

Developing Collaborative Editing Data Specifications

The USGS maintains guidelines for roads data contributed to *The National Map*. These guidelines include information about spatial accuracy, representation of geometry, completeness, topology, and attribution rules (J. Walters and G. Matthews, USGS, unpub. report, 2010). In order for the USGS to incorporate data from others, including other agencies and/or volunteers, it is necessary for those entities to meet minimum specifications to ensure the data are of a sufficient quality to include in *The National Map*. For the Phase One project, the partner was required to meet the USGS minimum standards for roads. Reaching agreement on specifications creates similar methods of doing business across different levels of government and the public, which is the ideal situation. This type of convention may not always be possible with all potential partners and volunteers.

Selecting a Partner (and partner data) for Collaborative Editing

The Data Access and Support Center (DASC) of the State of Kansas was chosen as a partner for this project. The DASC was selected for two significant reasons. First, the DASC was willing to provide personnel time to the project. Second, the DASC state-wide road data holdings were only partially complete. The quality of these data relative to the Census road data was evaluated in an NGTOC internal report (G. Matthews, USGS, unpub. report 2009). The report concluded that the DASC data have similar positional accuracy to the Census data and have attribution closely matching the Best Practices Data Model. However, the DASC data lacked standard geometric representations for dual carriageways and interchanges. This specific deficiency in geometric representation provided a focus for the co-editing process in the OSMCP.

Douglas and Johnson Counties in Kansas were selected as areas of interest for editing. The DASC is located within Douglas County. Johnson County is adjacent to Douglas and covers the metropolitan areas of Lawrence and much of Kansas City as well. This area was chosen due to the variety of road feature types and because the DASC staff had a reasonable working knowledge of this area (Figure 4).

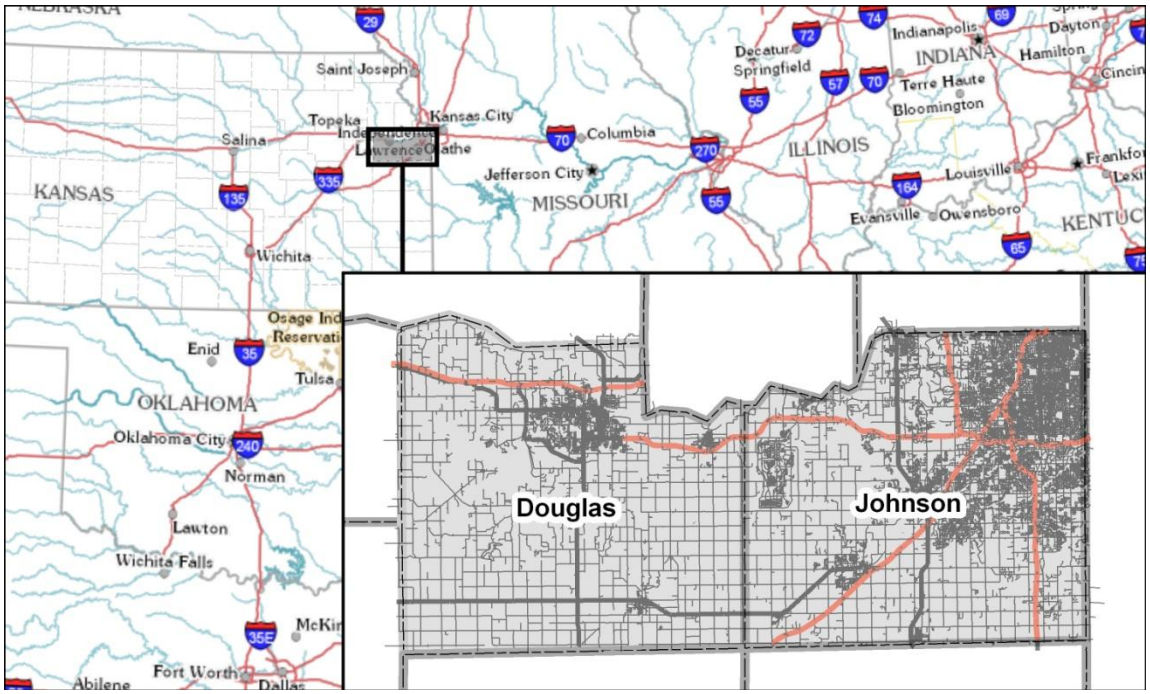

Figure 4. Area of interest for OSMCP Phase 1.

Preprocessing and Loading Data into the OSM Environment

The pre-processing steps for the Phase One project were carried out in two stages. First, the DASC assimilated county roads data to the state level to create a consistent state level roads dataset. Next, the USGS processed the DASC roads by using Safe Software Feature Manipulation Engine (FME) to crosswalk the data from the Kansas data schema into the USGS Best Practices data schema for Transportation and convert it from Personal Geodatabase into OSM XML format (Figure 5). This provided a baseline set of roads that had consistent attribution.

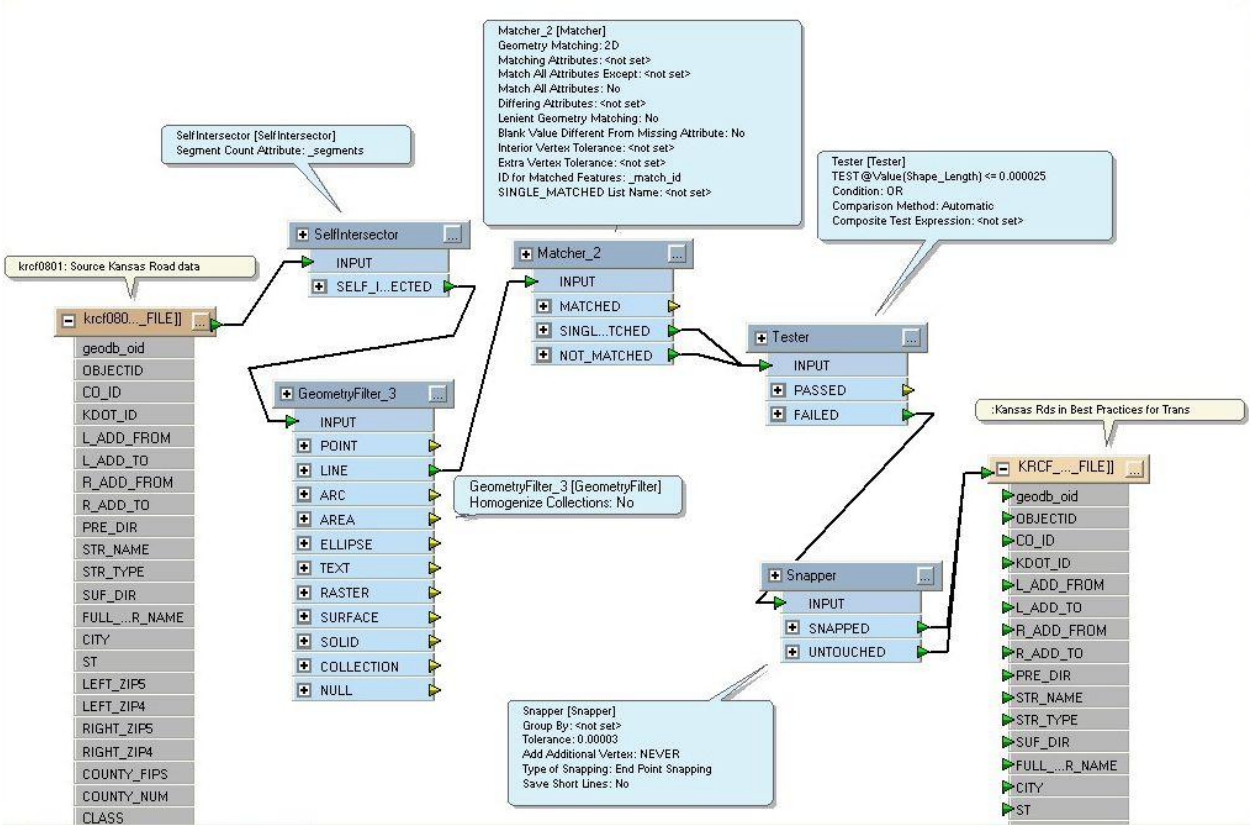

Figure 5. Generalized cross-walk of DASC road data into the BP schema in FME 2011 beta.

Geoprocessing steps included evaluating self-intersecting lines; running a geometry filter to check feature type; running a matching process to fix geometry problems, such as duplicate lines; testing to remove line slivers; running a snapping process to ensure connectivity in the new network; and, finally, converting the KS roads geometry into the USGS schema.

After the initial crosswalk was completed, the resulting file was converted into an OSM format that could be loaded into the OSM PostgreSQL database. This was a simple process that involved copying attributes, flagging certain text strings to help identify road types, and then writing to the final OSM file type (Figure 6).

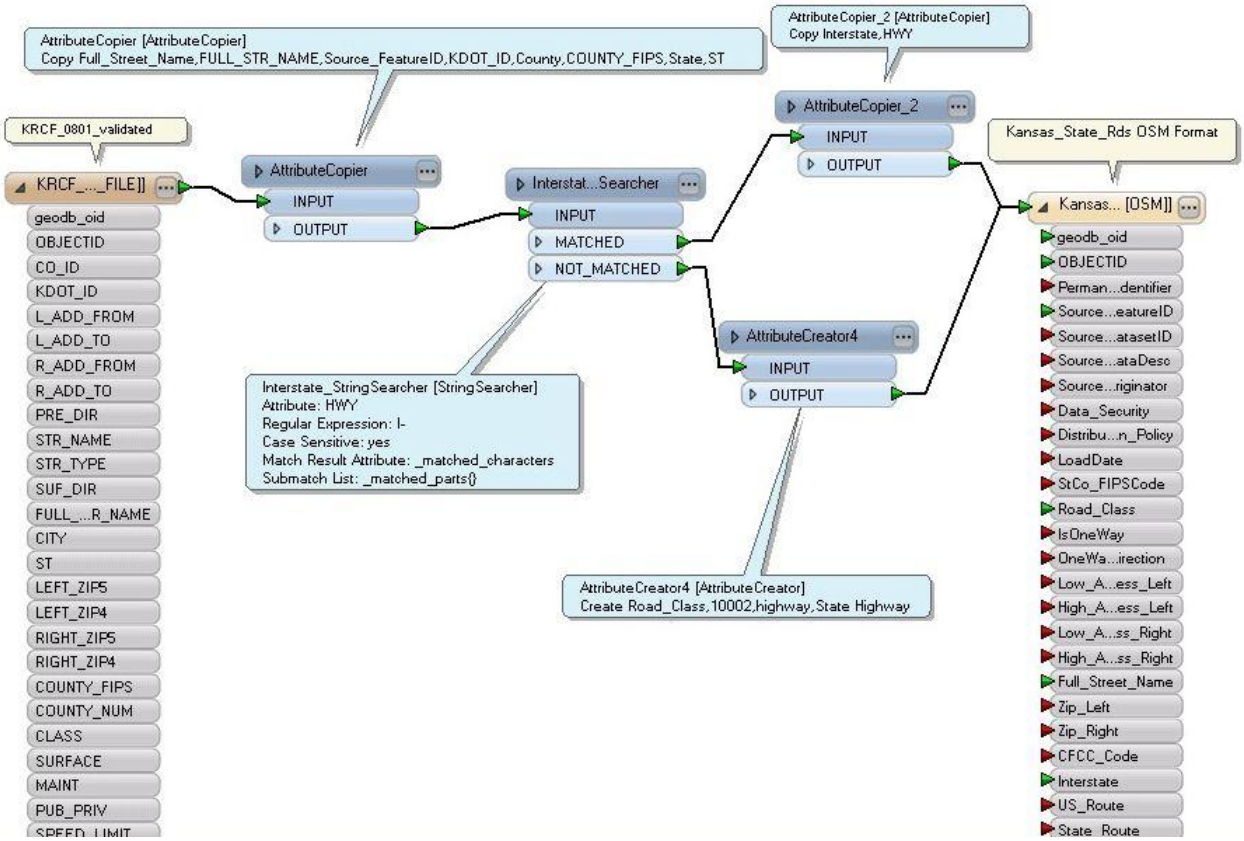

Figure 6. Conversion of crosswalk results into OSM-native file format.

Collaborative Editing of Kansas Roads Data in the OSM Environment

An NGTOC staff member oversaw the road editing with the DASC and verified that the edited data met all current USGS standards. Over the project area of interest, the USGS was responsible for editing Interstates, US Routes, and associated ramps and service roads. This editing covered a total of 270 miles at an approximate cost of 88 hours. This preliminary USGS editing was completed prior to involvement of the partner.

In order to assist the DASC in becoming familiar with the OSMCP web interface, a USGS staff member traveled to the DASC facility and spent time with personnel; helping them understand the software and the processes to be used in editing the data. The state routes the DASC was responsible for covered 72 miles of roads at a cost of approximately 40 hours. They accessed the Potlach editor interface through a standard browser, such as Mozilla Firefox or Microsoft Internet Explorer. While the editors at both the USGS and the DASC reported that the Web interface was easy to use and efficient, there was an editing problem that occurred for the DASC. A section of source roads data was hanging up the editing interface due to an unusually high number of points. This problem originally could have occurred due to “point streaming,” in which an editor may have neglected to close the session in a timely manner. This type of issue may even cause a desktop system to bog down while trying to load thousands of points along a single section of line. Future data collection error issues like this can be avoided by the USGS simplifying the data in FME before loading it into the OSM database.

Phase One OSMCP Data Results

The editing process focused on specific deficiencies in the DASC data identified in an internal report (G. Matthews, USGS, unpub. report 2009). The road data were already determined to meet NGP mapping requirements for positional accuracy and the attribution was cross-walked to NGP Best Practices. Table 3 shows a change summary of the resulting data after Phase One was complete. The increase in feature count and represented mileage for Interstates was largely due to the improper or missing representation of dual carriage-ways in the KS source data. Edits to interchange ramps were not captured properly by this statistic summary, but they are visually evident as seen in Figure 7.

Table 3. DASC roads data comparison for Douglas and Johnson counties.

		Road Type				
		Interstate	US Route	State Route	Ramps	Total
Original Data	Features	22	310	136	48	516
	Mileage	74	114	59	9	256
Final Data	Features	37	341	179	48	605
	Mileage	122	138	72	10	342
Changes	Feature	15	31	43	0	91
	Mileage	48	24	13	1	86
Percent Change	Features	68%	10%	32%	0%	18%
	Mileage	65%	21%	22%	11%	34%

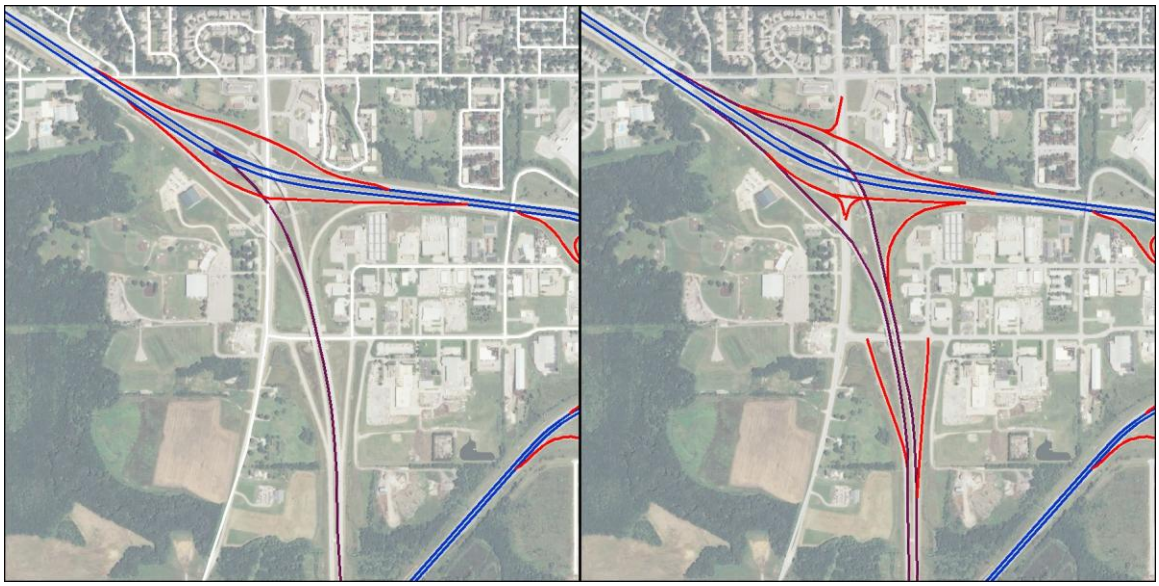

Figure 7. Unedited highway intersection, left. Corrected intersection with access ramps, right.

In addition to edits focusing on dual carriage-way and interchange representation and attribution, many minor data improvements were made both by NGTOC and DASC personnel that are not captured in the above table. Many features exhibited minor representational discrepancies that were out of alignment with the feature as represented in the source imagery, such as short jogs in Interstates. For the most part, these problems were within the error

tolerances specified by NGP Best Practices. However, the nature of the OSMCP interface allows the user to quickly identify and correct these minor positional issues. Similarly, many small improvements were made to attribution when noticed by the user, such as consistent use of abbreviations and capitalization. These minor edits were not directly tracked.

Analysis of the OSM Software for Collaborative Editing

The OSM software provided an easy-to-use platform running in a web browser that allowed co-editing by geographically dispersed personnel. The following positive and negative results were noted in the process.

Positive

- OSM and the software that runs it is well-supported by the VGI community relative to other software systems considered.
  - ArcGIS 10 supports editing data via the API 0.6 interface.
  - FME 2011 supports the reading and writing of OSM XML.
  - There are many free and open source editors beyond the Potlatch editor that are considered part of the core architecture.
- OSM software was reasonably easy to set up and use.
  - Software is free of cost and license restriction.
  - Software can be used on multiple machines without negotiating new software licenses, such as a cloud-based deployment.
- Multiple agencies can edit to the same specifications using the same data.
  - Conflict detection works well.
  - Data do not need to be checked out and are automatically versioned.
  - The platform supports thousands of simultaneous edit sessions.
  - The editing process in Potlatch (in browser) is much faster and easier than ArcGIS.
  - Potlatch allows complex geospatial data creation without technical expertise.

Negative

- Software setup requires staff with understanding of Linux, Ruby on Rails, and PostgreSQL.

  Few training opportunities exist for developing a level of competence with these tools. Staff working on the project needed to be comfortable self-teaching in the tools and materials. (Steve Coast, written communication, 2010)

- Some data artifacts can choke the OSM database.

  The OSM software cannot handle extremely large numbers of objects in any one screen of data. For instance, some of the road features in the DASC data contained unusually large numbers of points (thousands of points in a kilometer long, straight section of road). The resulting API calls and SQL queries were extremely long and tended to cause timeouts.

- Working in the USGS DMZ can create serious problems.

  The OSM API 0.6 is a Ruby on Rails application. It uses HTTP for client-server communications. The eSAS proxy includes filters to catch "unusual" HTTP traffic to and from servers in the DMZ (U.S. Geological Survey, 2010). Communication between the Potlatch client and the API was continually hampered by the eSAS filter because the URLs passed to the server were not easily predicted. Unlike the web proxy and filter used by desktop users in the USGS, the eSAS proxy blocks traffic, not providing specific errors like "blocked by firewall". Aggressive firewall filtering created what appeared to be application configuration errors. Initially, it was difficult to determine if the problem was in the server configuration or elsewhere. A significant amount of effort was spent verifying server configurations and querying the OSM community before suspecting the firewall.

- OSM editors are designed to support a qualitative rather than quantitative QC process

  The philosophy behind OSM relies on implicit quality rather than explicit quality. OSM does not track any quantitative measures of data quality such as positional accuracy or currentness. Instead, it is expected that OSM users will identify quality problems on the map, especially for features that are local to the user, and correct them to the best possible representation. The tools developed by the OSM community reflect an open-source notion of quality control following Linus' Law that "given enough eyeballs, all bugs are shallow (Raymond, 2001). This was demonstrated by the minor edits of positional character and attribution that were outside the major editing tasks of Phase One.

Discussion and Future Direction

Phase One of the OSMCP proved that creating and improving data to USGS specifications through a VGI prototype system at the USGS is possible. The effort resulted in the creation of a workable system for partner contributions through collaborative editing of spatial data by multiple organizations in a distributed environment. This prototype VGI platform provides a starting point for future efforts in more fully evaluating the usefulness and effectiveness of VGI for *The National Map*.

Phase Two

In Phase One, partners outside the USGS contributing edits were part of a state geographic information office. The DASC is a formally structured organization with professional GIS experience. In Phase Two, the USGS will start working with less organized contributors with less formal GIS experience. In Phase Two, the USGS will also shift focus from the roads data theme to the structures data theme. Historically, *The National Map Corps* has focused on structures data. It is somewhat less complex than roads data, consisting of points with no topological relationships instead of lines or arcs with fairly complex topology. Phase Two will build on the existing OSM-based platform, but will upgrade the user interface using Potlatch version 2. This new version of Potlatch provides greater flexibility in incorporating organization-specific data specifications. Additional background tiles will be created from web mapping services provided as part of *The National Map:*

La*http://services.nationalmap.gov/ArcGIS/rest/services/TNM_Vector_Large/MapServer*) and Scanned Topo Maps (DRG, *http://raster.nationalmap.gov/ArcGIS/rest/services/DRG/TNM_Digital_Raster_Graphics/MapServer*).

Phase Two will also focus on incorporating a more public, volunteer-based component. Rather than working with a state level GIS organization, the USGS will be engaging with GIS clubs at Denver-area colleges and universities to systematically collect structures data for four 7.5 minute USGS Topographic quadrangles in the Denver area: Arvada, Commerce City, Fort Logan, and Englewood. This will allow the USGS to gain an understanding of the quantity and quality of data collected by volunteers. Because the Phase 2 project will involve a wider, less formal audience, attention will be paid to elements of the OSMCP interface to ensure they more closely comply with USGS specifications. One of the basic concepts of user centered design is to build interfaces to guide the user naturally through the appropriate steps in a process. By paying closer attention to details in the user interface, the end user will have less trouble following the data collection guidelines.

Another important goal of Phase Two will be exploring methods for not only understanding the quality of data from volunteers, but also for managing quality control of the data being collected. Methods involving "user editors" will be explored. In other words, the project will include an investigation of the use of volunteers to perform quality control on data collected by other volunteers. Additionally, quantitative methods for determining completeness will be explored. Errors from incorrectly identified structures (commission or attribution errors) will be addressed as well as methods for estimating unidentified structures (omission). One possible technique involves comparing against existing datasets such as Geographic Names Information System (GNIS). These methods of estimating quality will be an important factor in determining the next steps for the use of volunteer data in *The National Map*.

Phase Three

If the results of Phase Two indicate the incorporation of volunteer-provided data provides quality data at lower cost than other methods of data collection, Phase Three of the OSMCP will expand the collection of structures data to more diverse areas. Using the 7.5 minute quads as an organizing structure for monitoring data quality, specific quads will be chosen to test the capability of user contributed data, especially in more rural areas. It has been demonstrated that OSM has less user contributions in rural areas (Brando, 2010; Girres, 2010; Haklay, 2010). Understanding how to encourage data collection activities in these areas will better meet the expectation of uniform coverage in *The National Map*. A wider range of volunteers will also be engaged (Table 4). Different volunteer groups provide different opportunities for sustainability. For instance, initial data capture of large or remote areas may require more skilled groups like GIS clubs. However, maintenance may be better managed through regular review by primary school groups where a fresh group of volunteers can review the data each year.

Table 4. Potential volunteer groups to engage in Phase 3.

Existing National Map Corps Members	Volunteer Fire Departments
OSM Community	4H Clubs
GIS Clubs	Primary School Classes
University Cartography/GIS Courses	

Phase Three will involve fewer technological changes to the platform with greater attention paid to the look and feel of the interface to match USGS standards.

Several models of data quality have been suggested for volunteered geographic information. One important model focuses on quality of data contributors (Grira and others, 2009). Phase Three will utilize automated methods for estimating quality of data, where possible, to create quality indicators for users. Different user communities may be compared to determine which are best engaged. The level of training effort may be varied across groups of volunteers to determine the necessary training to meet data quality standards.

Acknowledgments

This project would not have been successful without the active support of the Kansas Data Access and Support Center, efforts by the NGTOC Information Technology Support staff in Rolla, and the interest and enthusiasm of the OSM community. Austin Green provided valuable support during the creation of the cross-walked data in FME 2011. Jennifer Walters created the editing specifications document used in the project. Among the OSM community, we received input via mailing lists that directly benefitted the project from Steve Coast, Richard Weait, Brett Henderson, Ian Dees, Andrzej Zaborowski, Matt Amos, Tom Hughes, Andy Allan, Shaun McDonald, Richard Fairhurst, Steve Bennett and Brendan Morley.

References

Brando, C. and Bucher, B., 2010, Quality in user generated spatial content—A matter of specifications, *in* Painho, M., Santos, M.Y. and Pund, H. eds., Proceedings of the 13th AGILE International Conference on Geographic Information Science, 11–14 May 2010, Guimarães, Portugal, p. 1–8.

Cohn, J.P, 2008, Citizen science—Can volunteers do real research?: BioScience, v. 58, no. 3, p. 192–197.

Girres, J.F. and Touya, G., 2010, Quality assessment of the French OpenStreetMap dataset: Transactions in GIS, v. 14, no. 4, p. 435–459, also available at *http://dx.doi.org/10.1111/j.1467-9671.2010.01203.x*.

Goodchild, M.F, 2007, Citizens as sensors—The world of volunteered geography: GeoJournal, v. 69, no. 4, p. 211–221, also available at *http://dx.doi.org/10.1007/s10708-007-9111-y*.

Grira, J., Bedard, Y., and Roche, S., 2009, Spatial data uncertainty in the VGI world—Going from consumer to producer: Geomatica, v. 64, no. 1, p. 61–71.

Haklay, M., 2010, How good is volunteered geographical information? A comparative study of OpenStreetMap and ordnance survey datasets: Environment and Planning B—Planning and Design, v. 37, no. 4, p. 682–703, also available at *http://econpapers.repec.org/RePEc:pio:envirb:v:37:y:2010:i:4:p:682-703*.

Meyer, D., 2010, Microsoft gives Bing Maps photos to OpenStreetMap: ZDNet-UK, last accessed 23 December 2010 at *http://www.zdnet.co.uk/blogs/communication-breakdown-10000030/microsoft-gives-bing-maps-photos-to-openstreetmap-10021150/* .

OpenStreetMap, 2010, OpenStreetMap Wiki—Component overview: last accessed 15 Aug 2010 at *http://wiki.openstreetmap.org/wiki/Component_overview*.

Raymond, E.S., 2001, The cathedral & the bazaar—Musings on Linux and Open Source by an accidental revolutionary: Sebastopol, CA, O'Reilly Media, 241 p.

Scharl, A. and Tochterman, K., 2007, The geospatial web: London, Springer, 282 p.

Sugarbaker, L., Coray, K.E., and Poore, B.S., 2009, The National Map customer requirements—Findings from interviews and surveys*:* U.S. Geological Survey Open-File Report 2009–1222, 34 p., also available at *http://pubs.usgs.gov/of/2009/1222/*.

U.S. Geological Survey, 2010, DMZ & Infrastructure Team—Enterprise Secure Applications Service (eSAS): last accessed 1 February 2011 at *http://itsot.usgs.gov/dmz/eweb/index.html*.

Appendix A: Tilecache.cfg

Tilecache 2.11 was used to build the local set of NAIP imagery tiles from a WMS service provided by *The National Map*. This appendix lists the configuration file used by Tilecache to generate the tiles.

```
[cache]
type=GoogleDisk
base=/osmcp/rails/public/naip

[0]
type=WMS
url=http://isse.cr.usgs.gov/ArcGIS/services/Combined/SDDS_Imagery/MapSe
rver/WMSServer
#layers=0
levels=19
bbox=-180,90,180,-90
srs=EPSG:102113
spherical_mercator=true
tms_type=google
```

Appendix B: Potlatch Configuration

Potlatch 1.4 can be configured to display vector data based on colors and symbology in the following files. Also, tags and attributes can be automatically requested from the user per the settings in autocomplete.txt. This appendix provides the portions of these files that were modified for OSMCP Phase One.

Colors.txt (excerpt)

```
	Potlatch colours file
	# each line must be tab-separated: tag, stroke colour, casing?, fill
colour
	# ** TODO: act on key=value, not just one or the other

	Interstate	0x0033CC	1	-
	USHighway	0xCC0099	1	-
	Statehighway	0x7FC97F	1	-
	StateHighway	0x7FC97F	1	-
	USRoute	0x0033CC	1	-
	CountyRoute	0xE46D71	1	-
	LocalRoad	0xFFFFFF	1	-
	4wd	0xE8E8E8	1	-
	Ramp	0xFF0000	1	-
	Serviceroad	0xCC0099	1	-
	Privateroad	0xE8E8E8	1	-
```

presets.txt (excerpt)

```
	way/road
	Interstate: highway=Interstate,ref=(type road number)
	USHighway: highway=USHighway,ref=(type road number)
	Statehighway: highway=Statehighway,ref=(type road number)
	StateHighway: highway=StateHighway,ref=(type road number)
	CountyRoute: highway=CountyRoute,ref=(type road number)
	LocalRoad: highway=LocalRoad,ref=(type road number)
	4wd: highway=4wd,ref=(type road number)
	Ramp: highway=Ramp,ref=(type road number)
	Serviceroad: highway=Serviceroad,ref=(type road number)
	Privateroad: highway=Privateroad,ref=(type road number)
```

autocomplete.txt (excerpt)

```
	# Potlatch autocomplete values
	# each line should be: key / way|point|POI (tab) list_of_values
	# '-' indicates no autocomplete for values
	highway/way	Interstate,US Route,State Route,County Route,Local
Road,Ramp,Service Road,4WD Road, US Highway
```

Appendix C: Overcoming OSM Data Import Problems – Negative IDs

Importing the cross-walked OSM-native file resulted in one spurious issue. The OSM community assigns negative values to the ID file in large datasets intended for bulk upload into OSM. The FME (2011) followed this convention, assigning negative values to the ID field for every database object. The osmosis tool, which is commonly used by the OSM community for bulk uploads, will input data directly into PostgreSQL, bypassing the API 0.6. When osmosis is run in this fashion, it uses the ID values assigned to objects in the file rather than requesting a new ID value when the object is inserted into the database. Thus our first upload of the DASC data used negative ID values. Some aspects of the OSM software treated those negative IDs as invalid but no error messages were encountered. When an object was changed, a new copy with a valid ID was generated but the old object persisted. The system had no way to remove an object with a negative ID.

The problem was confirmed by Steve Coast, the founder of OSM, and others in the OSM community. They suggested using a bulk upload utility that worked against the API. This proved too slow, likely requiring several days to load the DASC dataset. Since we were starting with an empty database, we were able to drop and rebuild the database. Then the negative IDs in the cross-walked OSM-native file were changed to positive values using a series of sed commands. The Linux command-line utility, sed, executes regular-expression search and replace operations on text files. The osmosis utility was then used to load the OSM native file with positive IDs into the database. Finally, the sequences in PostgreSQL used to generate IDs were reset to the highest ID value after the import plus one to avoid duplicate IDs.