Assessment of Data Quality Improvements in *The National Map* Corps Project for 2012 Colorado Pilot (unpublished)

Erin Korris, Greg Matthews, Larry Moore

December, 2012

Introduction

Background

2012 was the third year of *The National Map* Corps project, an ongoing effort to determine how Volunteered Geographic Information (VGI) can contribute to *The National Map*. In 2011 the project studied a VGI model for collecting point structures, using a relatively controlled volunteer workforce of university students. In 2012 the model was expanded to a true VGI effort, but over a limited area of one state.

Data for 10 types of structures are being collected by volunteers covering the state of Colorado. Volunteers contribute to this project by editing or verifying existing structure points, collecting new structure points and removing points that are incorrect. Volunteers also perform quality control, called Adopt-a-Quad, testing whether the volunteer peer-review process - like that of Wikipedia - is effective. One of the goals of this project is to understand the overall quality of volunteered data collected. To help determine if volunteers can be used in the data editing and quality control process, a random statistical sampling of data is being reviewed by USGS staff to evaluate quality and completeness.

Purpose of data analysis

There are four "data status phases" in *The National Map* Corps project model. In this paper, these will be referred to as *phases*:

1. **GAZ** - Baseline data from USGS databases, not yet touched by volunteers.
2. **Volunteer Edit** - Features have been verified, edited, added, or deleted by volunteers.
3. **Adopt-A-Quad** (AAQ) - Features have been peer-reviewed by another volunteer. Volunteers who have edited at least 20 points can sign up as AAQ volunteer reviewers. AAQ reviewers check all of the points within a quad to assess positional accuracy and attribute accuracy (see appendix A for review guidelines). AAQ volunteers have responsibility for editorial formatting that goes beyond the expectations for other volunteers.
4. **Quality assurance** (QA) by USGS employees on statistically selected samples of data.

As part of the FY12 Colorado project, the results of this model were studied and these hypotheses tested:

> Hypothesis 1: As data move from phase 1 to 3, the overall quality and completeness improves with each step.
>
> Hypothesis 2: Phase 4 can monitor and measure these quality increases at acceptable internal cost.

Methodology

The Colorado Pilot Project domain is 10 structure types[1]. A random sample was selected from the population of all 10 structure types in the state of Colorado, 4,159 points in all. 100 points were sampled from this population for detailed phase 4 inspection. For each sampled point, USGS project staff evaluated the validity of the identification, and the position and attribute accuracy. This allowed overall accuracy of phase 1 (GAZ) to be compared to phase 2 (Volunteer Edit), and phase 2 to be compared to phase 3 (AAQ).

Safe Software FME 2012 was used to track each structures point contributed to the map, recording user login information, changes to location, attributes, and the time the change was made. Changes are stored in an OSM XML standard planet file that the USGS downloads nightly. The file from the previous day is not overwritten so a complete history of every point is available for a detailed analysis of the multiple changes to each structure as it was edited and quality checked.

Esri Data Reviewer was used to select a random sample of points. Using a confidence level of 95% and a margin of error of 10%, 100 points were randomly selected. Esri's data reviewer uses this standard formula:

$$sample\ size = \frac{Z^2 \times p \times (1-p)}{m}$$

where

Z = *Z* value (e.g., 1.96 for a 95% confidence interval)

p = Estimated incidence of the characteristic of interest in the population, expressed as a decimal percent (e.g., 0.3=30%). 0.5 maximizes the value p(1-p), so is commonly used as a default.

m = acceptable margin of error, expressed as a decimal (e.g., 0.10 = ± 10%)

This is consistent with, though not quite identical to, normal NGTOC procedures for certifying structures obtained from contractors and partners for the National Structures Dataset (USGS, 2012), which follow classic data quality measures used in the mapping community (Goodchild, 2007).

Four types of errors are evaluated in this study:

1. **Horizontal position errors**. Points are positioned relative to NAIP orthoimagery in the VGI interface. A point passes the horizontal accuracy test if it falls within the visual footprint of the correct building. The results of this evaluation are in columns 3 and 4 of Table 1.
2. **Attribute errors**. Name is a required attribute, street address is optional. For a point to pass the attribute accuracy test, the name must agree exactly with the name found in an independent authoritative source (such as GNIS or a Web site for the facility). The address, if provided, must also match independent authoritative sources if any are available. The results of this evaluation are in columns 5 and 6 of Table 1.

[1] See https://my.usgs.gov/confluence/download/attachments/155025503/Structure_Def_table.pdf for the list and definitions of structures.

3. **Errors of commission** – points identified falsely, where no actual feature exists. It is usually possible, though often expensive, to determine whether or not a given feature exists in the real world. Independent confirmation of feature existence was attempted for every point sampled in this study. The removal of a false point by a volunteer is a data improvement, and counts as a "pass" during a phase 4 evaluation. The results of this evaluation are in Table 2.
4. **Errors of omission** – not identifying a point where a relevant real-world feature does exist. These errors are difficult to measure because it is generally not possible to establish an absolute baseline for completeness of structures data. However, the 10 structures within the domain of this project include one feature type for which a complete baseline can be closely approximated. The U.S. Postal Service publishes an online directory of post office locations, so for any zip code area it is possible (though time consuming) to find the locations of all post offices. This was done for selected quads after they had been through AAQ to evaluate the effects of the VGI process on feature class completeness. The study of errors of omission was therefore done over a **different population** than the other three types of errors. The population of interest in this case is the set of all real-world post offices as identified by the U.S. Postal Service, over the area of three 7.5-minute quadrangles that had completed the AAQ process (17 post offices, as shown in Table 3). These points were not sampled; every post office location was inspected for a corresponding USGS data point. For the data point to pass, the post office feature had to be present, positioned accurately, and correctly named. The results of this evaluation are in Table 3.

Results

The table and charts below provide an overview of the quality analysis. Table 1 shows attribute and positional accuracy at each of the data phase. Each of these two types of accuracy was improved by each successive phase, reaching 100% for points that had been checked by AAQ volunteers.

Table 1. Attribute and Positional Accuracy by phase

Phase	Number of Points	Passed Points			
		Attribute Name		Position	
		Number	%	Number	%
GAZ	54	40	74%	41	76%
Volunteer Edit	32	24	75%	26	81%
Adopt-A-Quad	14	14	100%	14	100%

Errors of commission (Table 2) are the number of points that needed to be removed during USGS inspection. The same trend as in Table 1 is apparent, again validating hypothesis 1. The data in tables 1 and 2 also confirm hypothesis 2, that an additional phase of internal NGTOC quality control can monitor the data quality at each of the other phases. (Note that Table 1 measures accuracy, while Table 2 measures errors; in Table 2 low numbers are good.)

Table 2. Errors of Commission - Structures that do not really exist.

Phase	Number of Points	Errors of Commission	
		Number	%
GAZ	54	7	13%
Volunteer Edit	32	4	13%
Adopt-A-Quad	14	0	0%

It is impossible to obtain a 100% complete dataset for most feature types. Post offices are an exception (or nearly so) within the domain of this project, because the U.S. Postal Service publishes post office locations. The official U.S. Postal Service Web site has a locator tool (https://tools.usps.com/go/POLocatorAction_input) that was used to determine a complete dataset for quads that had been through AAQ. For these quads, the set of USPS locations was compared to the USGS set of post office points. This population was not sampled, but was completely inspected. Table 3 shows the results of this evaluation. The true complete dataset for this area, established by internal QA for the sample area, includes 49 post offices. Again, the accuracy of the data improved with each successive phase, though even the AAQ phase achieved only 82% completeness (in contrast to 100% position and attribute accuracy). We speculate this is due to the greater inherent difficulty of achieving feature completeness.

Table 3. Error of omission study results.

Phase	Identified Post Offices	Errors of Omission	Completeness %
GAZ	29	20	59%
Volunteer Edit	45	4	92%
Adopt-A-Quad	49	0	100%

Figure 1. Bar chart representation of several quality improvement measures.

Figure 1 is a different representation of selected quality measures from Tables 1-4. The figure illustrates the consistent pattern: for all quality measures, the Volunteer Edit phase is higher quality than the baseline data, and data that has reached Adopt-A-Quad phase is still higher quality.

Conclusions

This analysis of the FY12 Colorado pilot of the VGI project confirms the overall data collection model of the project. For all structure feature types, volunteer involvement improves postional accuracy, attribute accuracy, and reduces errors of commission. Errors of omission are more difficult to study and quantify, but the the study of post offices provides some evidence that the volunteer model improves completeness as well.

The cornerstone of this model is a Wikipedia-like hierarchy of editors and reviewers. The Colorado pilot demonstrated that volunteer edits improve our baseline strucutures data; that further review by advanced volunteers willing to adopt full quadrangles improves the data further; and that sample-based inspection by USGS personnel can monitor these processes. This study indicates that the model and procedures will scale to national scope.

References

Goodchild, M., 2007, Beyond metadata: Towards user-centric description of data quality: Proceedings International Symposium on Spatial Data Quality, Enschede, Netherlands, p. 1-11.

U.S. Geological Survey, 2012, Guidelines for Contributing Structures Data to *The National Map*: Reston, VA, U.S. Geological Survey.

Appendix A. Guidelines for populating the "Name" and "Address" Attributes in The National Structures Dataset

ATTRIBUTE NAME	EDITORIAL GUIDELINE FOR POPULATING THE ATTRIBUTE FIELD	EXAMPLES OF THE GUIDELINE		EXCEPTIONS TO THE GUIDELINE	GUIDELINE EXCEPTION EXAMPLES	
		INCORRECT	CORRECT		INCORRECT	CORRECT
NAME	NO ABBREVIATIONS	Bldg	Building	ABBREVIATIONS THAT REFER TO THE NAME OF A PERSON:		
		Co	County or Company			
		Ctr	Center			
		Cty	County or City			
		Dept	Department	**Jr** for **Junior**	Martin Luther King Junior Junior High School	Martin Luther King Jr Junior High School
		Dr	Drive or Doctor*			
		E	East			
		Est	Estate			
		FD	Fire Department	**Dr** for **Doctor**	Doctor William H Edison Memorial Hospital	Dr William H Edison Memorial Hospital
		Hosp	Hospital			
		Jr	Junior*			
		Med Center	Medical Center			
		Mt	Mount	**Sr** for **Senior**	Martin Luther King Senior Senior High School	Martin Luther King Sr Senior High School
		Mtn	Mountain			
		PS	Police Station			
		Sch	School			
		St	Saint			
		Sr	Senior*			
	NO APOSTROPHES FOR **NATURAL GEOGRAPHIC** FEATURES (mountain peak, natural water body, etc.)	Henry's Fork	Henrys Fork	APOSTROPHES **DENOTING MISSING LETTER**	Lake O the Woods	Lake O' the Woods
		Pike's Peak	Pikes Peak	APOSTROPHES NORMALLY USED TO **DENOTE SURNAME**	O Malley Hollow	O'Malley Hollow
		Long's Peak	Longs Peak	APOSTROPHES IN NAMES OF **ADMINISTRATIVE** OR **MAN-MADE GEOGRAPHIC** FEATURES	Sheriffs Office Marshal Office	Sheriff's Office Marshal's Office
	NO SPECIAL CHARACTERS[1]	"&" (ampersand)	the word "and"	--	--	--
		"@" (commercial at)	the word "at"	--	--	--
		"#" (number sign)	no replacement word	--	--	--
		"*" (asterisks)	remove without replacement	--	--	--
		"." (periods)	replace abbreviated word with with spelled out word for the noun (for example, spell out "Mount" instead of using the abbreviation "Mt.")	--	--	--
	TITLE CASE CAPITALIZATION (Each word begins with a capital letter.)	Denver art museum	Denver Art Museum	LOWER CASE FOR **PREPOSITIONS** (and, at, by, for, in, of, on)	Red River Of The North	Red River of the North
					Canyon De La Sierras Middle School	Canyon de la Sierras Middle School

* In most cases this word is not abbreviated. See exceptions to the rule columns for examples when this is not the case.

ADDRESS	NO ABBREVIATIONS (Spell out street types, prefixes, suffixes and directionals).	Ave	Avenue	--	--	--
		Blvd	Boulevard	--	--	--
		Cnty	County	--	--	--
		CR	County Road	--	--	--
		Dr	Drive	--	--	--
		E	East	--	--	--
		Hwy	Highway	--	--	--
		I	Interstate	--	--	--
		Int	Interstate	--	--	--
		NW	Northwest	--	--	--
		Pkwy	Parkway	--	--	--
		RT	Route	--	--	--
		Rte	Route	--	--	--
		SR	State Road	--	--	--
		St	Street	--	--	--
		SW	Southwest	--	--	--
		TSR	Township Road	--	--	--
		W	West	--	--	--
		Rd	Road	--	--	--
		Byp	Bypass	--	--	--
	NO APOSTROPHES IN STREET NAMES	Johnson's Way	Johnsons Way	--	--	--
		Sherman's Pike	Shermans Pike	--	--	--
	STREET NAME VARIATIONS: Use the name exactly as it is given by the local naming authority.[2]	--	· Third Street · 3rd Street · 3 Street	--	--	--
	SPELLING OF COUNTY, STATE, LOCAL HIGHWAY AND INTERSTATE ROUTES: Spell out consistent with the US Postal Service Address Standard[3].	Interstate Highway 680	Interstate 680	--	--	--
		I 55	Interstate 55	--	--	--
		US HWY 44	US Highway 44	--	--	--
		US 41 SW	US Highway 41 Southwest	--	--	--
		Hwy 64	County Highway 64	--	--	--
		KY ST HWY 1	State Highway 1	--	--	--
		Township RD 20	Township Road 40	--	--	--
	TITLE CASE CAPITALIZATION (Each word begins with a capital letter.)	US highway 40	US Highway 40	--	--	--
		County road 441	County Road 441			

[1] Hyphens ("-") and slashes ("/") are allowed in limited quantities, depending on the source of information.
[2] Different jurisdictions follow different practices for numbered street names. For example, Pittsburgh spells out "First" through "Tenth" in numbered street names.
[3] US Postal Service standard. Appendix F - Address Standardization - County, State, Local Highways (See pages 79-80 of the document, .pdf file pages 82-83).
http://nd911.homestead.com/USPSpub28.pdf (accessed October 2010)