Luke Kaim
Masters Project
Final Paper
2/17/2015

Feature Similarity Metrics for Integrating Volunteered Geographic Information (VGI)

Contents

Abstract: Volunteered Geographic Information (VGI) is becoming a valuable source of geographic information that compliments government generated data sets. VGI can be generated by a number of new Web supported technologies such as Wikimapia, OpenStreetMap, and Flickr; involve different themes (bird counts, phenological events); and take a number of different forms. Currently the mechanisms for VGI generation result in separate archives of data. Tremendous potential value lies in being able to combine these independently generated sources into integrated databases. The goal of this research project was to investigate an aspect of this integration problem. We assume a set of polygons generated by different sources with different geometries and different attribute sets and the objective is to discern if the polygons are sufficiently similar and could refer to the same feature. This project implemented a set of metrics for assessing whether two polygons are similar enough to be identified as the same feature. The similarity metrics use both geometry and attribute comparisons. The metrics are tested on simulated data sets that varied by geometry and name attributes, mimicking variations that might

occur in VGI contributions. The hope is that this work can contribute to an overall integration strategy for the growing volumes of VGI data.

Introduction:

Twenty years ago, geographic data creation was a very highly skilled endeavor. Most people were not familiar with the term GIS and had no experience collecting or working with geographic data. A simple geovisualization 15 years ago was reserved for highly skilled cartographers. With several new technologies and the reach of the World Wide Web this is a very exciting time for research in Geographic Information Science (GIS). The advent of smart phones and new Web 2.0 technologies has greatly changed the landscape of what is possible and who can utilize GIS. Web 2.0 technologies not only allow more people to use geospatial data, but allow people to create their own geographic information. Now an average 10 year old can use web based tools to collect data and create a geovisualization anywhere in the world. Goodchild (2007) coined the term Volunteered Geographic Information (VGI) to refer to geographic data collected by general citizen participants. Other terms referring to the same phenomena include Neogeography (Turner,2006; Sui 2008) and user contributed data, but the consensus seems to be that VGI is the preferred term.

Before VGI, creating geospatial data was a long and typically costly enterprise. Most of the geospatial data that was created before VGI was done by large National Mapping Agencies. Two of the largest National Mapping Agencies were the Ordnance Survey of Great Britain, and the United States Geological Survey. Because the National Mapping Agencies were government entities they had very strict metadata and standards associated with data collection.

There are currently no such metadata standards for VGI contributions, though over time we may see certain standards being adopted. Another issue with large governmental agency data production is the time lag between updates, often extending to years. VGI could help alleviate this short coming. With volunteered contributions, the time lag for edits is less than with National Mapping Agencies and often more detailed and thematically varied data are contributed by volunteers. My research looks at how we might go about utilizing this interesting and rich data. For example an entry for a lake in the National Hydrography Dataset (NHD), has a polygon boundary description, a name and a standard feature class identifier. VGI data however might contribute images of the lake, information on fish caught in the lake, wildlife seen on the lake, or invasive species seen in the lake, among other contributions.

This project work investigates methods for combining independently contributed VGI datasets together. A key challenge in integrating VGI data is in determining which VGI contributions refer to the same features. For this work we assumed that the VGI contributions included polygon representations and the objective was to identify polygons that appeared to refer to the same feature. Potential differences in VGI contributed polygons could be due to a number of factors. For example a polygon created in OpenStreetMap and another polygon created in Wikimedia is not going to be exactly the same since OpenStreetMap uses Bing imagery and Wikimapia uses Google imagery. Also two different people are digitizing the same feature, a lake for example, are unlikely to construct the exact same polygons. The two polygons are likely to be similar, but not exactly the same.

For the project research I implemented an approach for determining if two polygons created from different sources by different users might refer to the same feature. The project research goal was to determine if two polygons representing independent VGI contributions are similar based on geometric and attribute comparisons. A combination of geographic and attribute similarities over a certain threshold would suggest that two polygons likely represent the same feature.

Humans using manual checks would likely have little difficulty in determining if two representations referred to the same feature. Developing an automated approach, however, is quite challenging. The work is important since more and more user generated data is contributed, we have to find more efficient and automated ways to process the information.

The human eye is quite good at discerning changes between shapes, but with 1000 or millions of polygons to check this becomes an impractical way of comparing polygons. The goal of this work is to devise automated comparisons that can be done quickly and efficiently. Geometric measurements are used in combination with attribute comparisons since as Sehgal et al. (2006) writes "In our experiments, we find that using spatial and non-spatial together improves both recall and precision".

Geometrically the expectation is that polygons describing the same feature should be similar in their location and in their shape. For such comparisons one suggestion is to consider centroid to centroid distances (Fu et al 2005). This measure, however, does not really tell us how similar the two polygons are because the centroids could be very close in distance, but the areas and shapes of the polygons could vary greatly. Overlapping area is a measure that is pivotal. If the overlapping area is very high, then this could be a good indication that the two polygons are referring to the same feature. Perimeter comparison also should be an indicative measure. As illustrated in Figure 1, two polygons could have quite similar areas, but quite different perimeters suggesting that they might not be representing the same feature.

Figure 1: VGI representation compared to the reference polygon. The area is similar in both polygons, but the perimeter of the VGI polygon is larger.

The differences in distance between lower left and upper right coordinates of a polygon bounding boxes are helpful in determining the locational similarity between two polygons. The overlap in the bounding boxes between polygons is another metric for evaluating similarity that can be simpler to compute. For this project the ratio of overlapping area to the areas of each of the two comparison polygons was considered to be an effective metric and used in determining if the polygons were similar. For this metric, if the two input polygons are identical, the ratio is equal to 1.

The non-spatial information used for comparison was primarily the name field. Different VGI contributions have different tagging options, but most include a name field. Other options include an address field. Features can be assigned different names by different people in different contexts so names may not match exactly, but the expectation is that they should be similar. Various string comparisons were implemented to evaluate name similarity. The remainder of this paper described previous related research, describes the methodology in more detail, and presents results and evaluation. The last section makes conclusions and outlines future research.

Related Research

Much previous work exists on similarity metrics and matching and research for this project reviewed and borrowed some concepts from this previous work. In the assessment of independently generated VGI data, the expectation is not so much that two polygon representations should be identical, but that they should have sufficient similarities to believe that they could represent the same feature. Detailed shape or vertex matching as proposed by some work (Arkin et al 1991, Latecki et al 2003, VeltKamp and Hagedoorne 2001) is not needed in this case. Scene similarity matching (Nedas and Egenhofer 2008) deals with multiple objects and relationships among them (e.g. topological, directional) which are also not necessary for this work. Geographic information retrieval addresses some aspects of the similarity matching problem in that the goal is to find within a set of data sources, geographic information that approximately matches a user's request. In their construction of a geo-ontology Fu et al. (2005) use a set of similarity measures to determine if two places are the same. They include measures for name similarity, feature type similarity, footprint matching and hierarchy matching. The geographic hierarchy metric is based on containment relationships between places. Hastings (2008) takes a similar approach in combining multiple sources in the construction of a digital gazetteer. He employs three metrics for conflating independently generated source data. These include placename matching, place type, and footprint matching. For footprint matching he used the ratio of the overlapping area to the average of individual areas, for two spatially extended features, P1 and P2.

Place or feature type or class similarity can also be an important dimension for determining feature similarity. Classification hierarchies are typically used to find similar classes in classification trees (Rodriguez and Egenhofer 2003). In this work we treat the name field as a compound text string and look for matching tokens in the string.

> Future scientific advances are likely to involve mining of multidimensional data sets and to require the kinds of data synthesis that can only be achieved if systems are to a large degree interoperable. Gober (2000) called for a new emphasis on synthesis in geography in her Presidential Address, and a recent paper in Bioscience called for a new effort to accelerate synthesis in and between ecology and the environmental sciences (Carpenter et al. 2009). Many forms of synthesis in the context of VGI applications can be described as mashups (e.g., Yee 2008). Borrowed from the music industry, the term originally refers to a song or composition created by blending two or more songs. Yet in the context of Web-based applications, a mashup might have multiple meanings (Sui 2009). At the functional or service level, a mashup might be a Web page or application that combines data or functionality from two or more external sources to create a new service. In terms of actual content, a mashup can be a digital media file containing a combination of text, maps, audio, video, and animation, which recombines and modifies existing digital works to create a derivative work. The term implies easy, fast integration, frequently using open APIs and data sources to produce something new (Sarah Elwood et. al, 2011).

The underlying objective of this work is essentially an automated form of VGI mashup.

Approach and test data

The basic approach taken was to assume the existence of a reference feature which in this case was a polygon generated by a national mapping agency. The reference feature is assumed to have a correct official name and also an accurate polygon footprint with respect to both location and shape. An assumed set of VGI polygons were then to be matched against the reference polygon. A set of VGI polygons was simulated by making variations on the reference polygon that users might typically make by using different VGI web tools, and digitizing more or less detailed representations. To create the VGI set, the reference polygon was subjected to rotations, scaling, and translations as well as subtractions and additions to the number of vertices. Names were assigned to the VGI polygons to replicate the types of naming differences VGI contributors might make. These included misspellings, dropping the feature type from the name or using a different feature type. Given the reference name, Megunticook Lake, for the simulated VGI data set the following name variations were applied: Megunticook Lake (for an exact match), Megunticook Pond, Megunticook River, Megunticook Range, Magunticook,… We assume that the VGI test polygons and the reference polygon are in the same datum and projection. The set of geometric measures used to evaluate location and shape similarity included a measure based on area of polygon overlap similar to that employed by Hastings (2008). For this work polygon area similarity was defined as:

$$A_s = Area(P_1 \cap P_2)/\frac{1}{2}(Area(P_1) + Area(P_2)$$

For test purposes, a perimeter metric was also employed which takes a similar form:

$$P_s = Perimeter(P_1 \cap P_2)/\frac{1}{2}(Perimeter(P_1) + Perimeter(P_2)$$

For small features whose polygons representations are small, the likelihood of them overlapping is reduced. While such cases might not overlap, we would expect them to be in close proximity. To address these cases, an additional metric was included to test for their spatial proximity. The metric compares the lower left and upper right bounding box coordinates of the two polygons. Relatively small distances in this case indicate that the features are in the same general location.

Methods

ArcGIS 10.1 service pack 1 was used for this work and Python 2.7 scripting language was used to create a suite of tools for carrying out the matching operations. The sequence of steps implemented by these tools includes the following:

Intersect a VGI test and reference polygon. The tool actually takes an input a set of VGI polygons and generates the union of this set with the reference polygon. VGI polygons that intersect with the reference polygon are flagged and for these polygons their area and perimeter of overlap are estimated.

Compute the Area similarity metric A_s

Compute the Perimeter similarity metric P_s

Compute the separation distances between the polygons' Bounding box lower left and upper right coordinates

The developed tool carries out some initial variable assignments. Then it executes a union operation on the two input polygon files. The result of the union operation is a set of intersected polygons. ArcGIS uses a -1 to denote if there is an intersection. This means that if there are two -1's then there is an overlapping polygon.

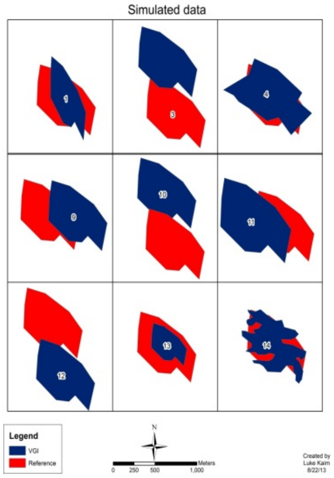


Figure 2: This is a set of simulated VGI polygons compared to a reference polygon.

The areas of these new polygons are then compared to the original polygon areas. This creates a ratio such that 1 equals a perfect match. The perimeter ratio is compared in the same fashion. The bottom right bounding box extent of VGI is compared to the bottom right bounding box extent for the reference polygon. If the bounding box point distances are small the polygons have a higher likelihood of being similar. This is also done for the top right extent. A minimum bounding box is created for every polygon. The bounding box area and perimeter ratios are compared in the same way as the polygon areas. The reason for computing both was to investigate how well the similarity metrics on the bounding box compare with respect to the similarity metrics on the polygons themselves. If they are consistently close, this would indicate that the similarity comparisons could be done on the bounding boxes alone.

Next the attribute information is compared. The name of the reference polygon is compared to the name of a VGI polygon. The string matching algorithms that were used to compare the names were adapted from the Python module difflib and from a library called FuzzyWuzzy. One commonly used string comparison algorithm is the Levenshtein distance. This algorithm returns the number of edits (character insertion, deletions or substitutions) that must occur to turn one string into the other (Baeza- Yates and Ribeiro-Neto 1999). This algorithm has some short comings for this feature name comparison task. First, it is case sensitive. There are ways to get around this short coming, namely to make all strings either lowercase or uppercase to start. Another shortcoming of this string matching algorithm is it does not work well with strings consisting of multiple words (e.g. strings 3-10 words long (Cohen). Feature names are typically two words incorporating a feature type class with a name, such as Branch Lake, or Union River. The Levenshtein distance creates extra penalties for omitting or using the wrong feature class. The developed matching tool includes a Levenstein distance metric, but to account for its shortcomings several other text string matching algorithms were tested. The SequenceMatcher algorithm returns a value between 0 and 1, with 1 representing an exact match. For this metric, X is the total number of elements in the two string sequences being compared, and M is the number of matches, therefore, 2.0*M/T. This is a simple string comparison algorithm. There is an algorithm from the FuzzyWuzzy library that is very close to the SequenceMatcher algorithm in this algorithm it matches the longest partial string in both. New York Giants and New York Jets would have a fairly high score using this partial string comparison in this approach in that the order of the string tokens matters. This is one short coming that might not be the optimal solution. One approach to overcome this is to tokenize the string and then sort them alphabetically and then join them back into the string. This is known as the Token Sort. Another string comparison called the Token Set, is similar in that it tokenizes the strings before comparing them and they are broken into two groups the intersection and the reminder. One thing that also is being tested is a mean average and weighted average of all the FuzzyWuzzy algorithms. All of the FuzzyWuzzy algorithms are normalized and comparable so averaging them all should produce an interesting string match. This will untimely be used to choose a threshold for which name information is similar enough to be considered the same.

Results

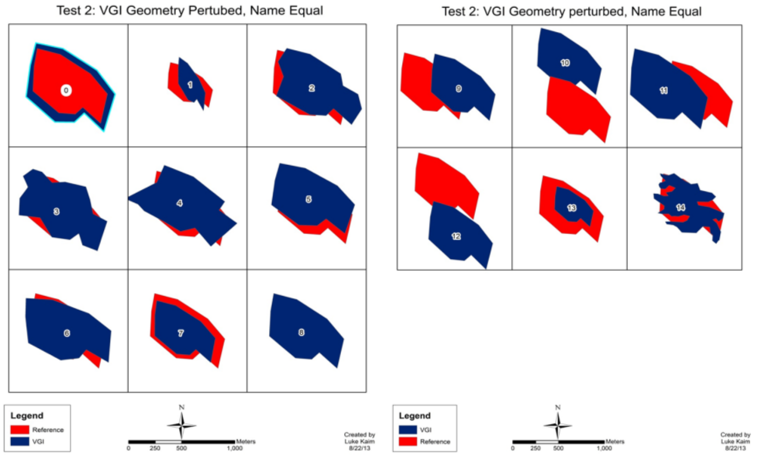


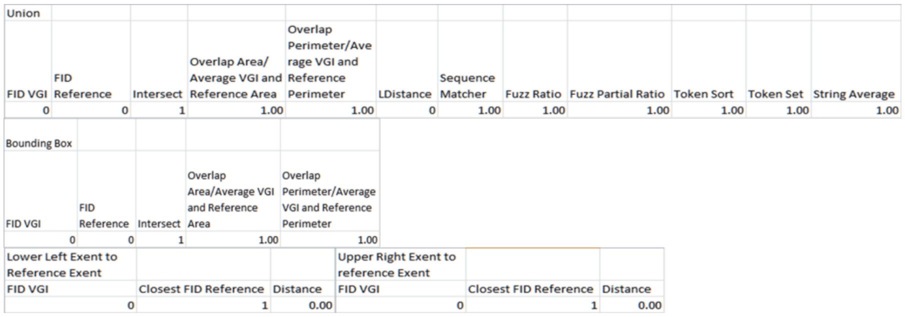
Figure 3: This is to show the 14 different VGI polygons that were tested against the reference polygon.

There were four tests that were conducted. Test 1 is where the VGI polygon is identical to the reference polygon, both in geometry and name. Test 2 is where the VGI geometry is perturbed and the name equals the reference polygon. Test 3 is where the VGI name is perturbed and the geometry equals the reference polygon. Test 4 is when both the VGI geometry and name are perturbed.

Test1

Test 1 was conducted where the VGI record was an exact match to the reference polygon. This is polygon 8 in figure 3. This is the ideal scenario.

Table 1: Test 1 is where the VGI polygon geometry and name equal the reference polygon.



Union												
FID VGI	FID Reference	Intersect	Overlap Area/ Average VGI and Reference Area	Overlap Perimeter/Average VGI and Reference Perimeter	LDistance	Sequence Matcher	Fuzz Ratio	Fuzz Partial Ratio	Token Sort	Token Set	String Average
0	0	1	1.00	1.00	0	1.00	1.00	1.00	1.00	1.00	1.00

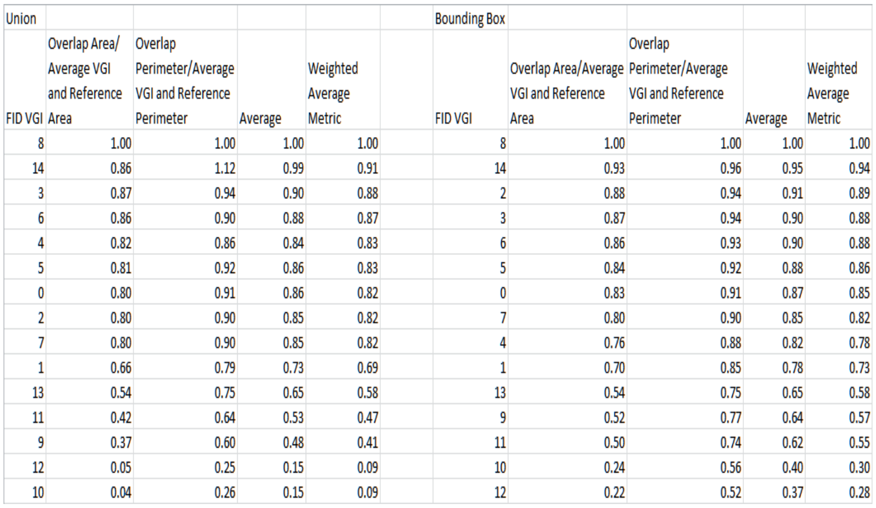
Bounding Box				
FID VGI	FID Reference	Intersect	Overlap Area/Average VGI and Reference Area	Overlap Perimeter/Average VGI and Reference Perimeter
0	0	1	1.00	1.00

Lower Left Exent to Reference Exent			Upper Right Exent to reference Exent		
FID VGI	Closest FID Reference	Distance	FID VGI	Closest FID Reference	Distance
0	1	0.00	0	1	0.00

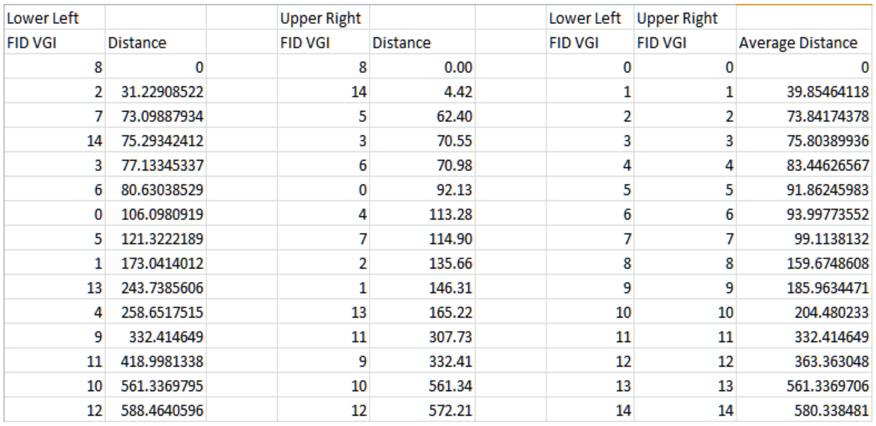
The overlapping area is equal to 1 and the overlapping perimeter is equal to 1. The name comparison algorithms are 1 also. This means that this VGI contribution is a perfect match to the reference polygon. The lower left and upper right extent distance is 0. This is the bench mark test that the other polygons will be compared to.

Test 2

Table 2: Test 2 is where the VGI polygons geometry is perturbed and the name equals the reference polygon.



Union						Bounding Box				
FID VGI	Overlap Area/ Average VGI and Reference Area	Overlap Perimeter/Average VGI and Reference Perimeter	Average	Weighted Average Metric		FID VGI	Overlap Area/Average VGI and Reference Area	Overlap Perimeter/Average VGI and Reference Perimeter	Average	Weighted Average Metric
8	1.00	1.00	1.00	1.00		8	1.00	1.00	1.00	1.00
14	0.86	1.12	0.99	0.91		14	0.93	0.96	0.95	0.94
3	0.87	0.94	0.90	0.88		2	0.88	0.94	0.91	0.89
6	0.86	0.90	0.88	0.87		3	0.87	0.94	0.90	0.88
4	0.82	0.86	0.84	0.83		6	0.86	0.93	0.90	0.88
5	0.81	0.92	0.86	0.83		5	0.84	0.92	0.88	0.86
0	0.80	0.91	0.86	0.82		0	0.83	0.91	0.87	0.85
2	0.80	0.90	0.85	0.82		7	0.80	0.90	0.85	0.82
7	0.80	0.90	0.85	0.82		4	0.76	0.88	0.82	0.78
1	0.66	0.79	0.73	0.69		1	0.70	0.85	0.78	0.73
13	0.54	0.75	0.65	0.58		13	0.54	0.75	0.65	0.58
11	0.42	0.64	0.53	0.47		9	0.52	0.77	0.64	0.57
9	0.37	0.60	0.48	0.41		11	0.50	0.74	0.62	0.55
12	0.05	0.25	0.15	0.09		10	0.24	0.56	0.40	0.30
10	0.04	0.26	0.15	0.09		12	0.22	0.52	0.37	0.28

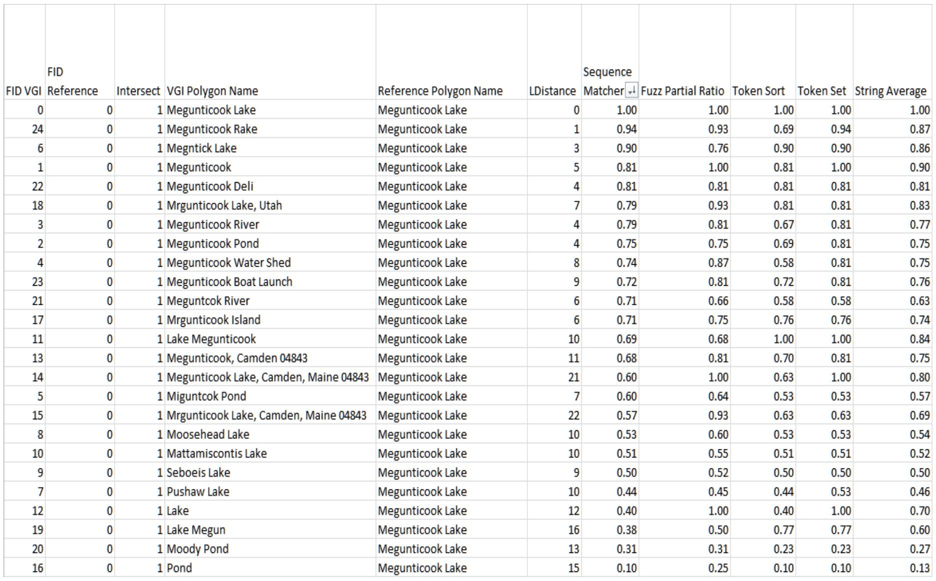


Lower Left			Upper Right			Lower Left	Upper Right	
FID VGI	Distance		FID VGI	Distance		FID VGI	FID VGI	Average Distance
8	0		8	0.00		0	0	0
2	31.22908522		14	4.42		1	1	39.85464118
7	73.09887934		5	62.40		2	2	73.84174378
14	75.29342412		3	70.55		3	3	75.80389936
3	77.13345337		6	70.98		4	4	83.44626567
6	80.63038529		0	92.13		5	5	91.86245983
0	106.0980919		4	113.28		6	6	93.99773552
5	121.3222189		7	114.90		7	7	99.1138132
1	173.0414012		2	135.66		8	8	159.6748608
13	243.7385606		1	146.31		9	9	185.9634471
4	258.6517515		13	165.22		10	10	204.480233
9	332.414649		11	307.73		11	11	332.414649
11	418.9981338		9	332.41		12	12	363.363048
10	561.3369795		10	561.34		13	13	561.3369706
12	588.4640596		12	572.21		14	14	580.338481

Here we see that the overlapping area and perimeter are different then 1. The perimeter of polygon 14 is greater than the average of both the reference and VGI polygon and that is why the perimeter calculation is greater than 1. The average for overlapping area and perimeter is also computed. The weighted average for overlapping area and perimeter was also calculated. The area calculation is weighted higher. This is because area has an absolute range, but perimeter does not.

Test 3

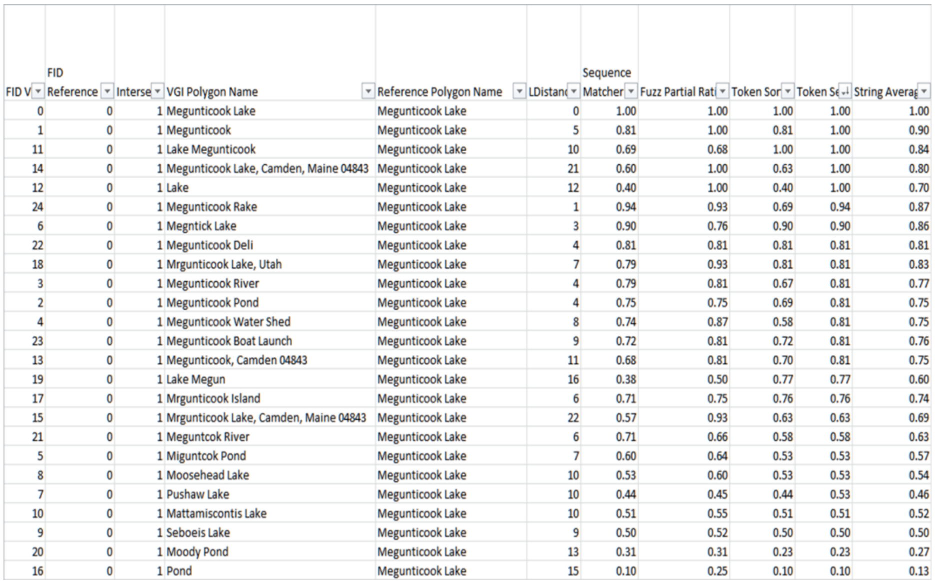
Table 3: Test 3 is where VGI geometry is equal to the reference polygon and the name is perturbed. There are a number of miss spelled words and other variations for the VGI polygon name. This table is sorted by sequence matcher score.



FID VGI	FID Reference	Intersect	VGI Polygon Name	Reference Polygon Name	LDistance	Sequence Matcher	Fuzz Partial Ratio	Token Sort	Token Set	String Average
0	0	1	Meguntícook Lake	Megunticook Lake	0	1.00	1.00	1.00	1.00	1.00
24	0	1	Megunticook Rake	Megunticook Lake	1	0.94	0.93	0.69	0.94	0.87
6	0	1	Megntick Lake	Megunticook Lake	3	0.90	0.76	0.90	0.90	0.86
1	0	1	Megunticook	Megunticook Lake	5	0.81	1.00	0.81	1.00	0.90
22	0	1	Megunticook Deli	Megunticook Lake	4	0.81	0.81	0.81	0.81	0.81
18	0	1	Mrgunticook Lake, Utah	Megunticook Lake	7	0.79	0.93	0.81	0.81	0.83
3	0	1	Megunticook River	Megunticook Lake	4	0.79	0.81	0.67	0.81	0.77
2	0	1	Megunticook Pond	Megunticook Lake	4	0.75	0.75	0.69	0.81	0.75
4	0	1	Megunticook Water Shed	Megunticook Lake	8	0.74	0.87	0.58	0.81	0.75
23	0	1	Megunticook Boat Launch	Megunticook Lake	9	0.72	0.81	0.72	0.81	0.76
21	0	1	Meguntcok River	Megunticook Lake	6	0.71	0.66	0.58	0.58	0.63
17	0	1	Mrgunticook Island	Megunticook Lake	6	0.71	0.75	0.76	0.76	0.74
11	0	1	Lake Megunticook	Megunticook Lake	10	0.69	0.68	1.00	1.00	0.84
13	0	1	Megunticook, Camden 04843	Megunticook Lake	11	0.68	0.81	0.70	0.81	0.75
14	0	1	Megunticook Lake, Camden, Maine 04843	Megunticook Lake	21	0.60	1.00	0.63	1.00	0.80
5	0	1	Miguntcok Pond	Megunticook Lake	7	0.60	0.64	0.53	0.53	0.57
15	0	1	Mrgunticook Lake, Camden, Maine 04843	Megunticook Lake	22	0.57	0.93	0.63	0.63	0.69
8	0	1	Moosehead Lake	Megunticook Lake	10	0.53	0.60	0.53	0.53	0.54
10	0	1	Mattamiscontis Lake	Megunticook Lake	10	0.51	0.55	0.51	0.51	0.52
9	0	1	Seboeis Lake	Megunticook Lake	9	0.50	0.52	0.50	0.50	0.50
7	0	1	Pushaw Lake	Megunticook Lake	10	0.44	0.45	0.44	0.53	0.46
12	0	1	Lake	Megunticook Lake	12	0.40	1.00	0.40	1.00	0.70
19	0	1	Lake Megun	Megunticook Lake	16	0.38	0.50	0.77	0.77	0.60
20	0	1	Moody Pond	Megunticook Lake	13	0.31	0.31	0.23	0.23	0.27
16	0	1	Pond	Megunticook Lake	15	0.10	0.25	0.10	0.10	0.13

Table 3

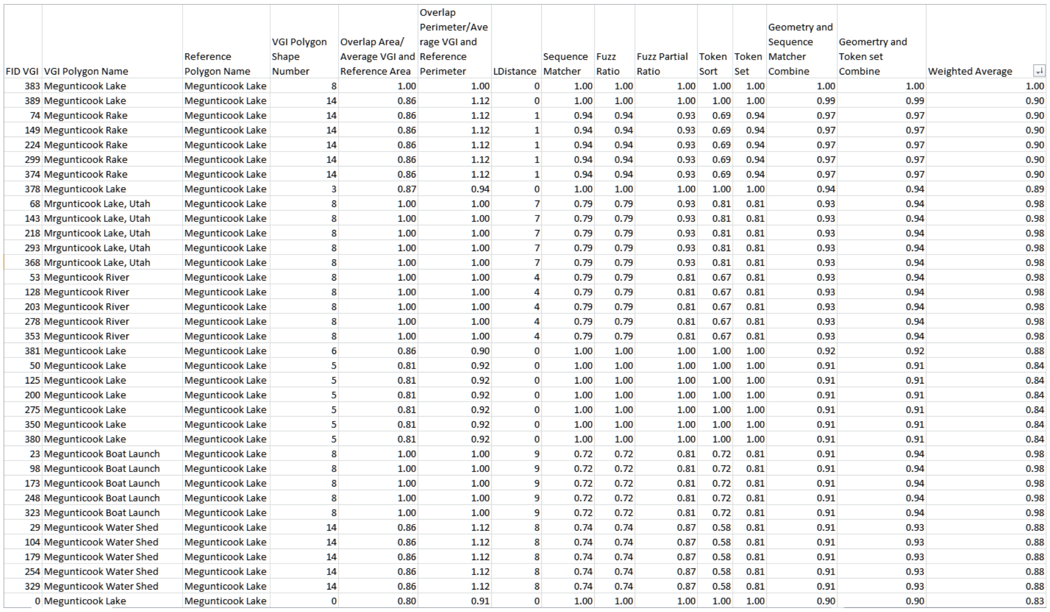
Table 4: Test 3 is where VGI geometry is equal to the reference polygon and the name is perturbed. This table is sorted by token set.



FID V	FID Reference	Interse	VGI Polygon Name	Reference Polygon Name	LDistanc	Sequence Matcher	Fuzz Partial Rati	Token Sor	Token Se	String Averag
0	0	1	Megunticook Lake	Megunticook Lake	0	1.00	1.00	1.00	1.00	1.00
1	0	1	Megunticook	Megunticook Lake	5	0.81	1.00	0.81	1.00	0.90
11	0	1	Lake Megunticook	Megunticook Lake	10	0.69	0.68	1.00	1.00	0.84
14	0	1	Megunticook Lake, Camden, Maine 04843	Megunticook Lake	21	0.60	1.00	0.63	1.00	0.80
12	0	1	Lake	Megunticook Lake	12	0.40	1.00	0.40	1.00	0.70
24	0	1	Megunticook Rake	Megunticook Lake	1	0.94	0.93	0.69	0.94	0.87
6	0	1	Megntick Lake	Megunticook Lake	3	0.90	0.76	0.90	0.90	0.86
22	0	1	Megunticook Deli	Megunticook Lake	4	0.81	0.81	0.81	0.81	0.81
18	0	1	Mrgunticook Lake, Utah	Megunticook Lake	7	0.79	0.93	0.81	0.81	0.83
3	0	1	Megunticook River	Megunticook Lake	4	0.79	0.81	0.67	0.81	0.77
2	0	1	Megunticook Pond	Megunticook Lake	4	0.75	0.75	0.69	0.81	0.75
4	0	1	Megunticook Water Shed	Megunticook Lake	8	0.74	0.87	0.58	0.81	0.75
23	0	1	Megunticook Boat Launch	Megunticook Lake	9	0.72	0.81	0.72	0.81	0.76
13	0	1	Megunticook, Camden 04843	Megunticook Lake	11	0.68	0.81	0.70	0.81	0.75
19	0	1	Lake Megun	Megunticook Lake	16	0.38	0.50	0.77	0.77	0.60
17	0	1	Mrgunticook Island	Megunticook Lake	6	0.71	0.75	0.76	0.76	0.74
15	0	1	Mrgunticook Lake, Camden, Maine 04843	Megunticook Lake	22	0.57	0.93	0.63	0.63	0.69
21	0	1	Meguntcok River	Megunticook Lake	6	0.71	0.66	0.58	0.58	0.63
5	0	1	Miguntcok Pond	Megunticook Lake	7	0.60	0.64	0.53	0.53	0.57
8	0	1	Moosehead Lake	Megunticook Lake	10	0.53	0.60	0.53	0.53	0.54
7	0	1	Pushaw Lake	Megunticook Lake	10	0.44	0.45	0.44	0.53	0.46
10	0	1	Mattamiscontis Lake	Megunticook Lake	10	0.51	0.55	0.51	0.51	0.52
9	0	1	Seboeis Lake	Megunticook Lake	9	0.50	0.52	0.50	0.50	0.50
20	0	1	Moody Pond	Megunticook Lake	13	0.31	0.31	0.23	0.23	0.27
16	0	1	Pond	Megunticook Lake	15	0.10	0.25	0.10	0.10	0.13

Token set seems to do a reasonable job sorting the name field. The first 4 records seem to be the same lake. This is why location matters and then name matters. If the location was close and one volunteer called the feature Lake Megunticook and another person called it Megunticook Lake it is likely they were depicting the same feature.

Table 5: Test 4 is where the Geometry and name are both perturbed when compared to the reference polygon. This was sorted by the average geometry and sequence match field.



FID VGI	VGI Polygon Name	Reference Polygon Name	VGI Polygon Shape Number	Overlap Area/ Average VGI and Reference Area	Overlap Perimeter/Average VGI and Reference Perimeter	LDistance	Sequence Matcher	Fuzz Ratio	Fuzz Partial Ratio	Token Sort	Token Set	Geometry and Sequence Matcher Combine	Geomertry and Token set Combine	Weighted Average
383	Megunticook Lake	Megunticook Lake	8	1.00	1.00	0	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
389	Megunticook Lake	Megunticook Lake	14	0.86	1.12	0	1.00	1.00	1.00	1.00	1.00	0.99	0.99	0.90
74	Megunticook Rake	Megunticook Lake	14	0.86	1.12	1	0.94	0.94	0.93	0.69	0.94	0.97	0.97	0.90
149	Megunticook Rake	Megunticook Lake	14	0.86	1.12	1	0.94	0.94	0.93	0.69	0.94	0.97	0.97	0.90
224	Megunticook Rake	Megunticook Lake	14	0.86	1.12	1	0.94	0.94	0.93	0.69	0.94	0.97	0.97	0.90
299	Megunticook Rake	Megunticook Lake	14	0.86	1.12	1	0.94	0.94	0.93	0.69	0.94	0.97	0.97	0.90
374	Megunticook Rake	Megunticook Lake	14	0.86	1.12	1	0.94	0.94	0.93	0.69	0.94	0.97	0.97	0.90
378	Megunticook Lake	Megunticook Lake	3	0.87	0.94	0	1.00	1.00	1.00	1.00	1.00	0.94	0.94	0.89
68	Mrgunticook Lake, Utah	Megunticook Lake	8	1.00	1.00	7	0.79	0.79	0.93	0.81	0.81	0.93	0.94	0.98
143	Mrgunticook Lake, Utah	Megunticook Lake	8	1.00	1.00	7	0.79	0.79	0.93	0.81	0.81	0.93	0.94	0.98
218	Mrgunticook Lake, Utah	Megunticook Lake	8	1.00	1.00	7	0.79	0.79	0.93	0.81	0.81	0.93	0.94	0.98
293	Mrgunticook Lake, Utah	Megunticook Lake	8	1.00	1.00	7	0.79	0.79	0.93	0.81	0.81	0.93	0.94	0.98
368	Mrgunticook Lake, Utah	Megunticook Lake	8	1.00	1.00	7	0.79	0.79	0.93	0.81	0.81	0.93	0.94	0.98
53	Megunticook River	Megunticook Lake	8	1.00	1.00	4	0.79	0.79	0.81	0.67	0.81	0.93	0.94	0.98
128	Megunticook River	Megunticook Lake	8	1.00	1.00	4	0.79	0.79	0.81	0.67	0.81	0.93	0.94	0.98
203	Megunticook River	Megunticook Lake	8	1.00	1.00	4	0.79	0.79	0.81	0.67	0.81	0.93	0.94	0.98
278	Megunticook River	Megunticook Lake	8	1.00	1.00	4	0.79	0.79	0.81	0.67	0.81	0.93	0.94	0.98
353	Megunticook River	Megunticook Lake	8	1.00	1.00	4	0.79	0.79	0.81	0.67	0.81	0.93	0.94	0.98
381	Megunticook Lake	Megunticook Lake	6	0.86	0.90	0	1.00	1.00	1.00	1.00	1.00	0.92	0.92	0.88
50	Megunticook Lake	Megunticook Lake	5	0.81	0.92	0	1.00	1.00	1.00	1.00	1.00	0.91	0.91	0.84
125	Megunticook Lake	Megunticook Lake	5	0.81	0.92	0	1.00	1.00	1.00	1.00	1.00	0.91	0.91	0.84
200	Megunticook Lake	Megunticook Lake	5	0.81	0.92	0	1.00	1.00	1.00	1.00	1.00	0.91	0.91	0.84
275	Megunticook Lake	Megunticook Lake	5	0.81	0.92	0	1.00	1.00	1.00	1.00	1.00	0.91	0.91	0.84
350	Megunticook Lake	Megunticook Lake	5	0.81	0.92	0	1.00	1.00	1.00	1.00	1.00	0.91	0.91	0.84
380	Megunticook Lake	Megunticook Lake	5	0.81	0.92	0	1.00	1.00	1.00	1.00	1.00	0.91	0.91	0.84
23	Megunticook Boat Launch	Megunticook Lake	8	1.00	1.00	9	0.72	0.72	0.81	0.72	0.81	0.91	0.94	0.98
98	Megunticook Boat Launch	Megunticook Lake	8	1.00	1.00	9	0.72	0.72	0.81	0.72	0.81	0.91	0.94	0.98
173	Megunticook Boat Launch	Megunticook Lake	8	1.00	1.00	9	0.72	0.72	0.81	0.72	0.81	0.91	0.94	0.98
248	Megunticook Boat Launch	Megunticook Lake	8	1.00	1.00	9	0.72	0.72	0.81	0.72	0.81	0.91	0.94	0.98
323	Megunticook Boat Launch	Megunticook Lake	8	1.00	1.00	9	0.72	0.72	0.81	0.72	0.81	0.91	0.94	0.98
29	Megunticook Water Shed	Megunticook Lake	14	0.86	1.12	8	0.74	0.74	0.87	0.58	0.81	0.91	0.93	0.88
104	Megunticook Water Shed	Megunticook Lake	14	0.86	1.12	8	0.74	0.74	0.87	0.58	0.81	0.91	0.93	0.88
179	Megunticook Water Shed	Megunticook Lake	14	0.86	1.12	8	0.74	0.74	0.87	0.58	0.81	0.91	0.93	0.88
254	Megunticook Water Shed	Megunticook Lake	14	0.86	1.12	8	0.74	0.74	0.87	0.58	0.81	0.91	0.93	0.88
329	Megunticook Water Shed	Megunticook Lake	14	0.86	1.12	8	0.74	0.74	0.87	0.58	0.81	0.91	0.93	0.88
0	Megunticook Lake	Megunticook Lake	0	0.80	0.91	0	1.00	1.00	1.00	1.00	1.00	0.90	0.90	0.83

Polygon 14 is very high in this test because polygon 14 has a larger perimeter then the average perimeter of both the VGI input and the reference polygon. This is why that it is 1.12. This is raising the value above 1.

Summary

When using both geometry and name similarities the tool seems to be able to deduce the nuances of different polygons. The weighted average using both geometry and name similarities should be able to discern most polygons. The weights might need to be changed slightly because polygon 14 is skewing the results. More work will have to be done if the perimeters of the VGI contributions are at a higher detail then the reference polygon. The area computation maximum can only be 1 because it is the area of overlap and as such there cannot be more area then the input polygons. This is, however, not true with perimeter. More work could be done to see if there is a way to normalize the perimeter calculation. Another option is to weight the area calculation more heavily than the perimeter calculation. In the weighted average that is what was done.

Looking at the geometry values for the shape and the values for the bounding box it looks as if the bounding box is giving similar results. This could be an indication that the bounding box is enough to discern different polygon shapes.

The token set algorithm seems to do the best job when looking at all the different name variations. This is because order shouldn't matter as much; meaning Megunticook Lake should be the same as Lake Megunticook. If feature type was its own field, then token set would definitely be the best choice.

It is also important to keep in mind that this method of detecting similar polygons is an important step in automating this process. It is important to keep in mind though that humans will still need to be involved if there is no reference polygon to base a VGI contribution on. If volunteers collect the data, then it is not a far stretch to think that volunteers could also check the contributions as well. This methodology is not meant to replace the human component, but help the human check for similarities within VGI contributions. With being able to test similarities within different datasets this should lend itself to synthesizing geographic data and we might be able to learn more about the world around us because the whole is bigger than the sum of its parts (Aristotle).

Future work:

Future work could incorporate a feature thesaurus so that similar feature types are more similar then none similar feature types. This is something that my implementation does not take advantage of currently. This work could also be added to by testing this set of tools against complex polygons. Looking at very small polygons presents a challenge that was not looked at with this research. Building footprints are very small so two people could digitize buildings that do not overlap. Currently if the two polygons do not overlap then it would be hard to know that they are similar. Two building footprints could be very similar, but be disjoint. Future works could look at how to discern disjoint feature and if they are similar. Adding elevation could also be helpful. Just because two polygons overlap doesn't necessarily mean that they are similar if they are at largely different elevations.

Acknowledgments:

I want to take this time and thank my advisor Kate Beard, committee, colleagues, family and office mates.

References

Adam Cohen. “FuzzyWuzzy: Fuzzy String Matching in Python | SeatGeek.” Web. 22 July 2013.

Al-Bakri, Maythm, and David Fairbairn. “Assessing Similarity Matching for Possible Integration of Feature Classifications of Geospatial Data from Official and Informal Sources.” International Journal of Geographical Information Science 26.8 (2012): 1437–1456. Print.

Arkin, Esther M. et al. “An Efficiently Computable Metric for Comparing Polygonal Shapes.” Proceedings of the First Annual ACM-SIAM Symposium on Discrete Algorithms. Society for Industrial and Applied Mathematics, 1990. 129–137. Print.

Auer, Thomas, and Martin Held. “Rpg-heuristics for the Generation of Random Polygons.” Proc. 8th Canada Conf. Comput. Geom. Ottawa, Canada. Citeseer, 1996. 38–44. Print.

Beard, Kate. “A Semantic Web Based Gazetteer Model for VGI.” Proceedings of the 1st ACM SIGSPATIAL International Workshop on Crowdsourced and Volunteered Geographic Information. ACM, 2012. 54–61. Print.

Bennett, Brandon, Amar Isli, and Anthony G. Cohn. “A System Handling RCC-8 Queries on 2D Regions Representable in the Closure Algebra of Half-planes.” Methodology and Tools in Knowledge-Based Systems. Springer, 1998. 281–290. Print.

Bishr, Mohamed, and Lefteris Mantelas. “A Trust and Reputation Model for Filtering and Classifying Knowledge About Urban Growth.” GeoJournal 72.3-4 (2008): 229–237. Print.

Budhathoki, Nama Raj, and Zorica Nedovic-Budic. “Reconceptualizing the Role of the User of Spatial Data Infrastructure.” GeoJournal 72.3-4 (2008): 149–160. Print.

De Longueville, Bertrand et al. “Digital Earth’s Nervous System for Crisis Events: Real-time Sensor Web Enablement of Volunteered Geographic Information.” International Journal of Digital Earth 3.3 (2010): 242–259. Print.

Elwood, Sarah. “Volunteered Geographic Information: Future Research Directions Motivated by Critical, Participatory, and Feminist GIS.” GeoJournal 72.3-4 (2008): 173–183. Print.

---. “Volunteered Geographic Information: Key Questions, Concepts and Methods to Guide Emerging Research and Practice.” GeoJournal 72.3 (2008): 133–135. Print.

Elwood, Sarah, Michael F. Goodchild, and Daniel Z. Sui. “Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practice.” Annals of the Association of American Geographers 102.3 (2012): 571–590. Print.

Fowler, Amy et al. “How Reliable Are Citizen-derived Scientific Data? Assessing the Quality of Contrail Observations Made by the General Public.” Transactions in GIS (2013): n. pag. Print.

Fu, Gaihua, Christopher B. Jones, and Alia I. Abdelmoty. “Building a Geographical Ontology for Intelligent Spatial Search on the Web.” Databases and Applications. 2005. 167–172. Print.

Gerevini, Alfonso, and Bernhard Nebel. “Qualitative Spatio-temporal Reasoning with RCC-8 and Allen’s Interval Calculus: Computational Complexity.” ECAI. Vol. 2. 2002. 312–316. Print.

Goodchild, Michael F. “Citizens as Sensors: The World of Volunteered Geography.” GeoJournal 69.4 (2007): 211–221. Print.

---. “Commentary: Whither VGI?” GeoJournal 72.3 (2008): 239–244. Print.

---. “Geographic Information Systems and Science: Today and Tomorrow.” Annals of GIS 15.1 (2009): 3–9. Print.

Gouveia, Cristina, and Alexandra Fonseca. “New Approaches to Environmental Monitoring: The Use of ICT to Explore Volunteered Geographic Information.” GeoJournal 72.3-4 (2008): 185–197. Print.

Grira, Joel, Yvan Bédard, and S. Roche. “Spatial Data Uncertainty in the VGI World: Going from Consumer to Producer.” Geomatica 64.1 (2010): 61–71. Print.

Hastings, J. T. “Automated Conflation of Digital Gazetteer Data.” International Journal of Geographical Information Science 22.10 (2008): 1109–1127. Print.

Klyne, Graham, Jeremy J. Carroll, and Brian McBride. “Resource Description Framework (RDF): Concepts and Abstract Syntax.” W3C recommendation 10 (2004): n. pag. Print.

Latecki, Longin Jan, Rolf Lakämper, and Diedrich Wolter. "Shape Similarity and Visual Parts." Discrete Geometry for Computer Imagery. Springer, 2003. 34–51. Print.

Liu, Weiming, Sanjiang Li, and Jochen Renz. "Combining RCC-8 with Qualitative Direction Calculi: Algorithms and Complexity." IJCAI. 2009. 854–859. Print.

Maué, Patrick. "Reputation as Tool to Ensure Validity of VGI." Workshop on Volunteered Geographic Information. 2007. Print.

McCreath, Eric. "Partial Matching of Planar Polygons Under Translation and Rotation." CCCG. 2008. Print.

Miller, Eric. "An Introduction to the Resource Description Framework." Bulletin of the American Society for Information Science and Technology 25.1 (1998): 15–19. Print.

Mummidi, Lakshmi Narayana, and John Krumm. "Discovering Points of Interest from Users' Map Annotations." GeoJournal 72.3-4 (2008): 215–227. Print.

"OWLIM Primer - OWLIM42 - Ontotext Wiki." Web. 15 July 2013.

Pultar, Edward et al. "Dynamic GIS Case Studies: Wildfire Evacuation and Volunteered Geographic Information." Transactions in GIS 13.s1 (2009): 85–104. Print.

Qian, Xinlin et al. "Data Cleaning Approaches in Web2. 0 VGI Application." Geoinformatics, 2009 17th International Conference On. IEEE, 2009. 1–4. Print.

Rodríguez, M. Andrea, and Max J. Egenhofer. "Determining Semantic Similarity Among Entity Classes from Different Ontologies." Knowledge and Data Engineering, IEEE Transactions on 15.2 (2003): 442–456. Print.

Samal, Ashok, Sharad Seth, and Kevin Cueto 1. "A Feature-based Approach to Conflation of Geospatial Sources." International Journal of Geographical Information Science 18.5 (2004): 459–489. Print.

Seeger, Christopher J. "The Role of Facilitated Volunteered Geographic Information in the Landscape Planning and Site Design Process." GeoJournal 72.3-4 (2008): 199–213. Print.

Sehgal, Vivek, Lise Getoor, and Peter D. Viechnicki. "Entity Resolution in Geospatial Data Integration." Proceedings of the 14th Annual ACM International Symposium on Advances in Geographic Information Systems. ACM, 2006. 83–90. Print.

Smart, Philip D., Christopher B. Jones, and Florian A. Twaroch. "Multi-source Toponym Data Integration and Mediation for a Meta-gazetteer Service." Geographic Information Science. Springer, 2010. 234–248. Print.

Tulloch, David L. "Is VGI Participation? From Vernal Pools to Video Games." GeoJournal 72.3-4 (2008): 161–171. Print.

Veltkamp, Remco C., and Michiel Hagedoorn. State of the Art in Shape Matching. Springer, 2001. Print.

Wolter, Frank, and Michael Zakharyaschev. “Spatial Representation and Reasoning in RCC-8 with Boolean Region Terms.” Proceedings of the 14th European Conference on Artificial Intelligence (ECAI 2000). Citeseer, 2000. 244–248. Print.