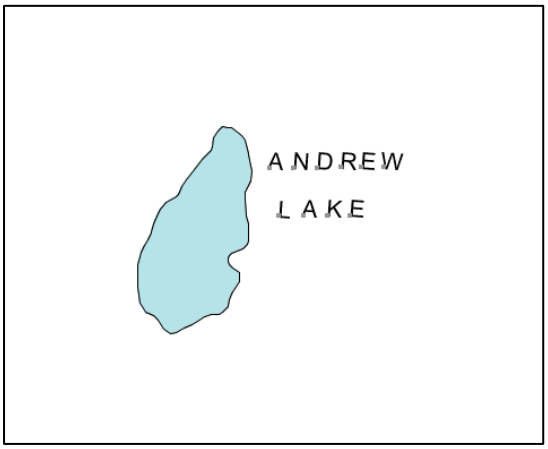
Water Body Toponym Matching

Submitted By: Justin Deoliveira, Programmer
Refractions Research Inc.
400 - 1207 Douglas Street
Victoria, BC, V8W-2E7
jdeolive@refractions.net
Phone: (250) 383-3022
Fax: (250) 383-2140

1 INTRODUCTION

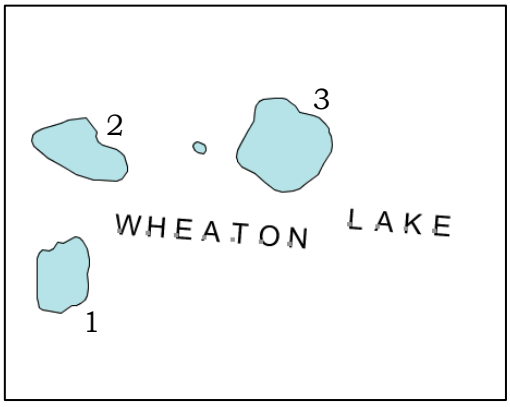
A *Water Body Toponym* is a collection of cartographic letters that are intended to name hydrographic features such as lakes, marshes, etc.

Currently there is no association between a water body and the toponym that names it. The only relationship between the two is derived from the spatial characteristics of the toponym: *A toponym should be close to the water body it is intended to name.*



In the figure above we can see from the placement of the toponym "ANDREWLAKE" that it is intended to name the lake shown.

However there are situations in which is not as clear as above which water body a toponym is intended to name.



In the figure above it is unclear as to which of the water bodies labeled 1, 2, or 3 is actually “WHEATONLAKE”.

In this situation we seek out the *best possible match* which is defined as the water body that is most likely to be named by the toponym. Obtaining the best possible match is described in the following section.

2 APPROACH

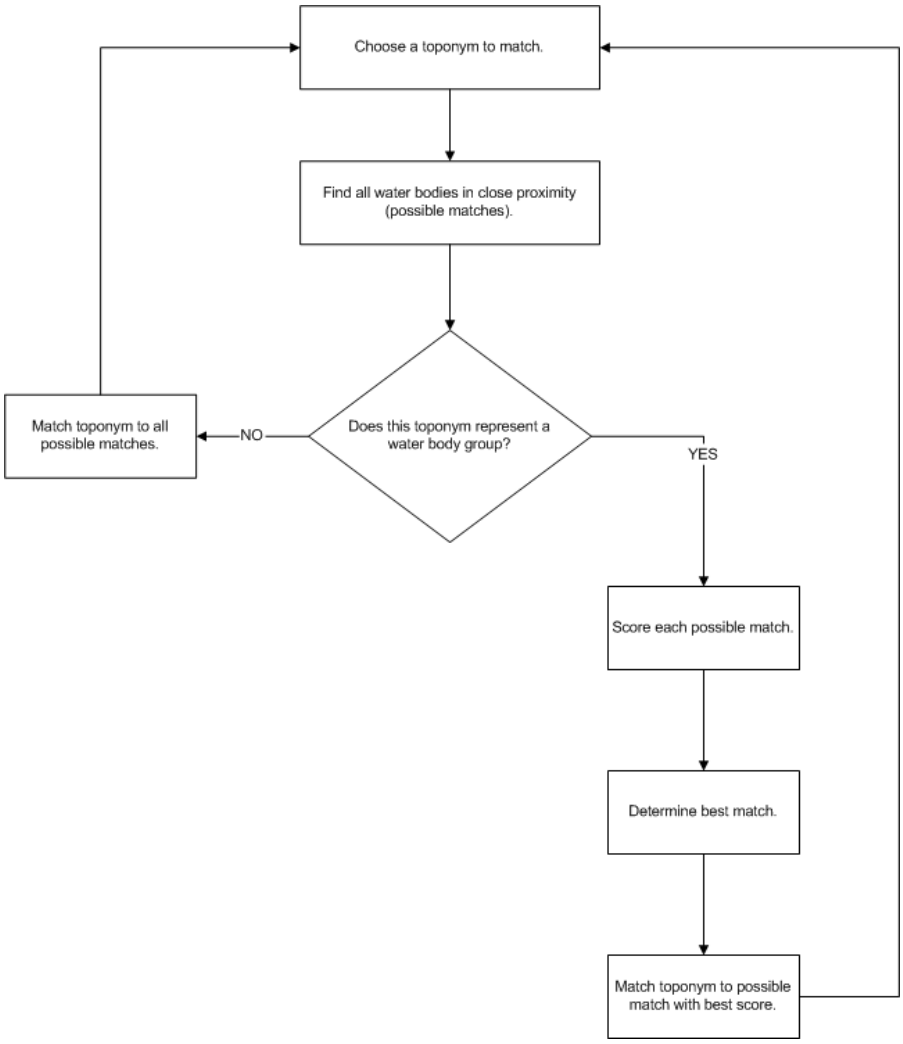
The Water Body Toponym Matching algorithm works on the following assumptions.

1. A toponym *contained* with a water body indicates a very good match.
2. A toponym in close *proximity* to a water body indicates a good match.
3. A water body with a large *area* is more likely to be named than a water body with a smaller area.

Each water body that is being considered as a possible match for a toponym is scored using the properties mentioned above. A weighted average of the properties is then calculated and the water body with the highest average is deemed the *best possible match*.

3 THE ALGORITHM

The following flowchart outlines the Water Body Toponym matching algorithm.



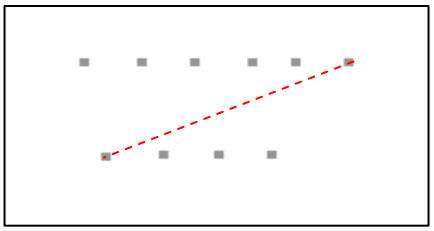
3.1 Finding Water Bodies in Close Proximity.

In order to determine which water bodies to include in the set of possible matches for the toponym, the spatial orientation of the toponym is analyzed.

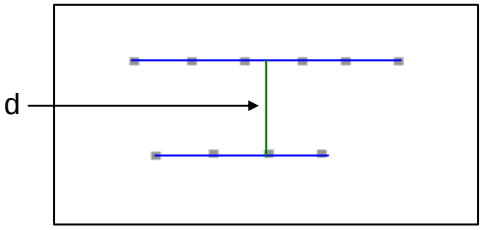
First the *name break* of the toponym is found. A name break is defined as the position in the toponym where one word ends and another begins. As an example consider the toponym “ANDREWLAKE”. This toponym is made up of the words

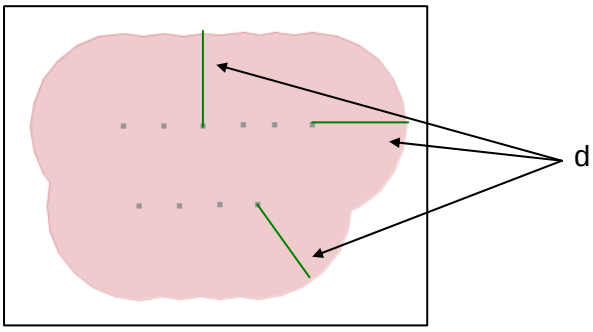
"ANDREW" and "LAKE". The name break exists between the 6th and 7th letters of the toponym.

In order to determine the name break of a toponym, the greatest distance between two *consecutive* points in the toponym is calculated. This is illustrated below.

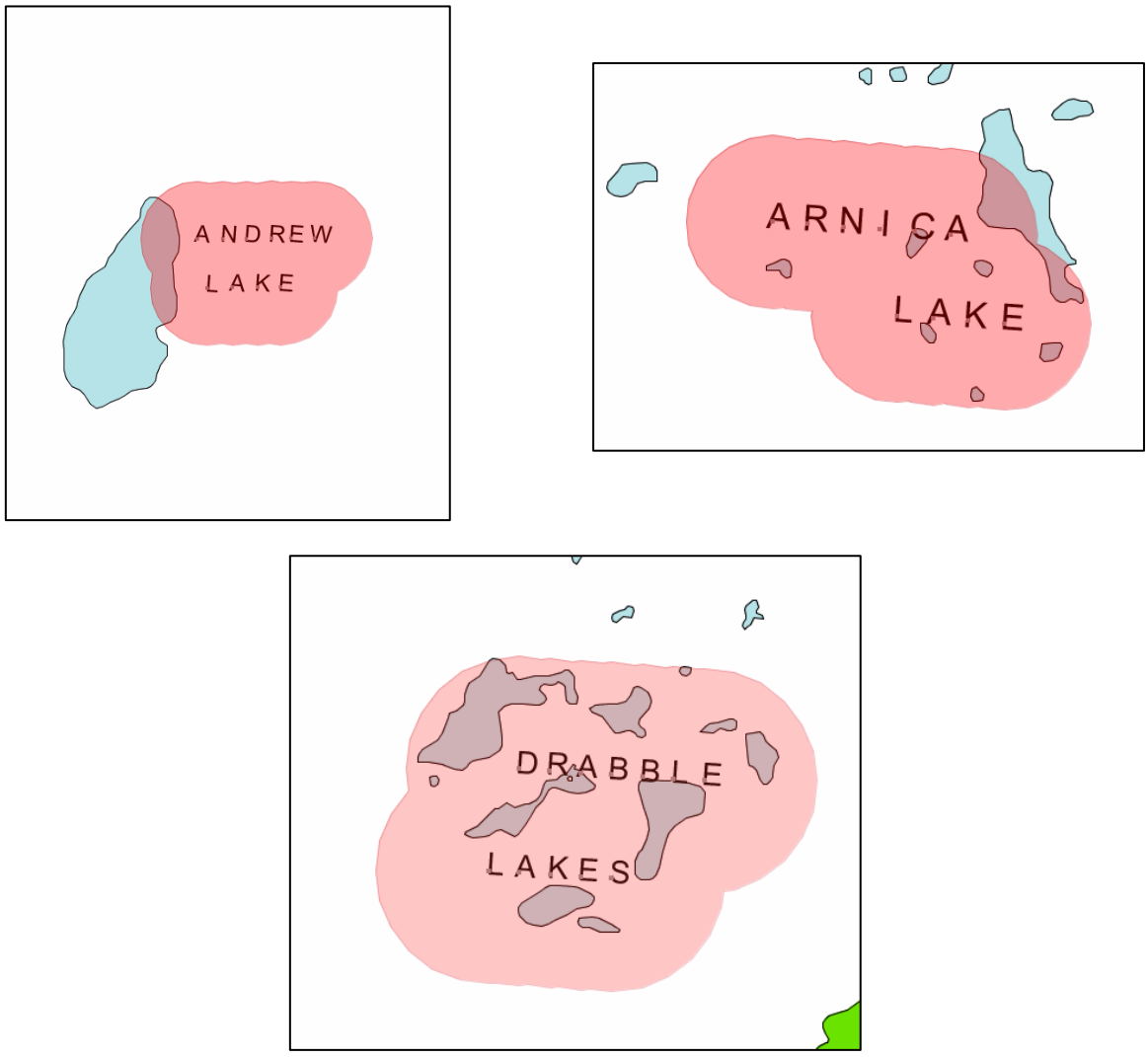


Once the name break has been calculated, there exists two sets of letters, "ANDREW" and "LAKE". The distance between the two sets is then calculated by joining the letters in each set with a line, then calculating the distance d between the lines as shown below.





This distance d is then used to create a buffer around the toponym as shown above. Any water body that intersects the buffer is considered to be a possible match. The following are examples.



Any water body that intersects the red area is considered to be a *possible match*.

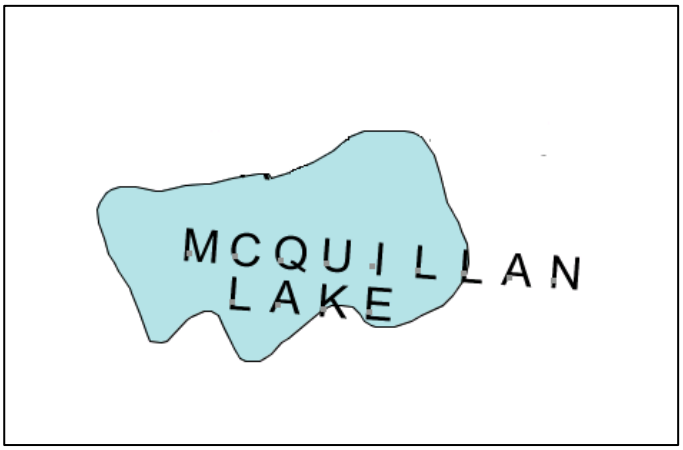
The motivation for using this method is as follows. Toponyms that name large water bodies tend to be spread out more then toponyms that name smaller water bodies. Using this method allows the algorithm to dynamically change its definition of "close" based on the spatial properties of the toponym.

3.2 Scoring matches

Each possible match is scored using the measures of *containment*, *proximity*, and relative *area*.

3.2.1 Containment

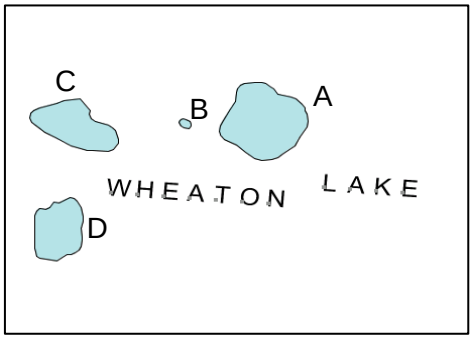
This is an absolute measure that is simply the number of letters inside the water body divided by the number of letters in the toponym. The following is an example



In the above example, containment = 10/13 = 77%.

3.2.2 Proximity

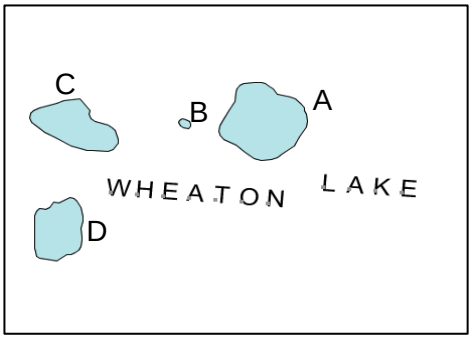
This is a relative measure of how close the polygon is to the toponym compared to the other polygons in the set of possible matches. It is calculated by taking the cumulative of distance of every point in the toponym from the water body, and dividing it by the largest measure of cumulative distance in the set.



Cumulative Distance	Proximity
A = 1829.53	A = 1829.53 / 3665.43 = 0.499
B = 2597.41	B = 2597.41 / 3665.43 = 0.708
C = 3623.78	C = 3623.78 / 3665.43 = 0.989
D = 3665.43	D = 3665.43 / 3665.43 = 1.000

3.2.3 Relative Area

This is a relative measure of how big the polygon is compared to the other polygons in the set of possible matches.



Area	Relative Area
A = 20615.364	A = 20615.364 / 20615.364 = 1.000
B = 373.807	B = 373.807 / 20615.364 = 0.018
C = 18052.783	C = 18052.783 / 20615.364 = 0.876
D = 16813.124	D = 16813.124 / 20615.364 = 0.815

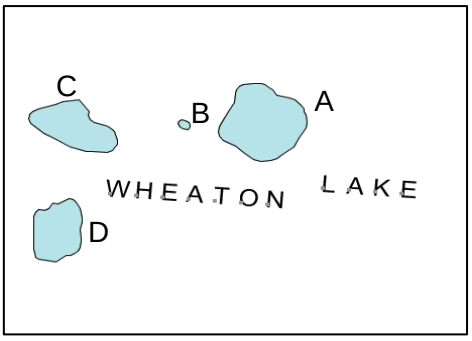
3.3 Determining the best possible match

For each possible match, a weighted average of the properties mentioned in section 3.2 is calculated.

```
Average = a*(containment) + b*(1 – proximity) + c*(area)

a,b,c = weights
```

The best possible match is the polygon with the highest average. In the following example we use the weights a = 0.10, b = 0.50, c = 0.40.



Weighted average
A = 0.655
B = 0.153
C = 0.356
D = 0.326

Therefore, “WHEATONLAKE” would be matched to water body A.