Process Summary for the TRIM Toponymy ⇔ BCGNIS Linking Process

Submitted By: Paul Ramsey (pramsey@refractions.net)
David Blasby (dblasby@refractions.net)
Refractions Research Inc.
400 - 1207 Douglas Street
Victoria, BC, V8W-2E7
Phone: (250) 383-3022
Fax: (250) 383-2140

REVISION NUMBER	DATE OF ISSUE	AUTHOR	DESCRIPTION OF CHANGE
1.0	2003/08/01	Paul Ramsey	First Draft
1.1	2003/08/13	Paul Ramsey	Internal Revisions

1 BACKGROUND

1.1 BC GNIS

BC GNIS is the British Columbia Geographical Names Information System, the repository of official gazetted names for the province of British Columbia. BCGNIS is stored as an Oracle relational database, with detailed information about provenance, location, and history of geographical names in the province.

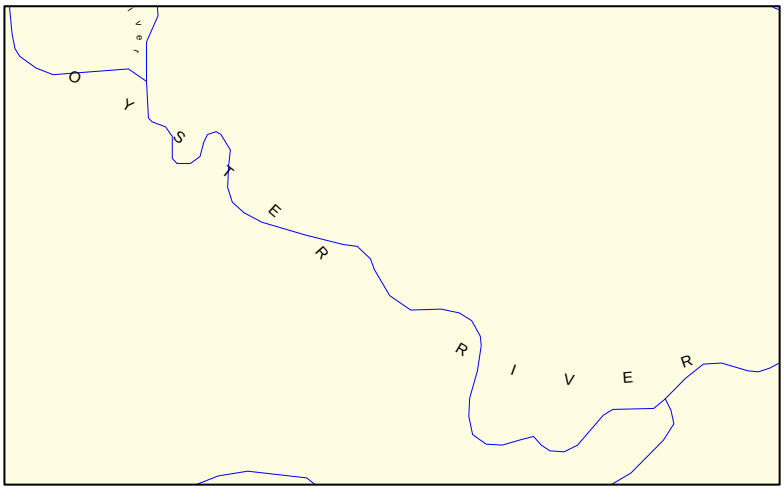
The BCGNIS data model includes the concept of a "feature", a physical land form to which a name (or many names in one or more languages) can be attached. For each "feature", BCGNIS stores the following information (along with a great deal of other information not important to this discussion):

- The name (or names) associated with the feature.
- The approximate location of the feature (and possibly a headwaters location if the feature is a river).
- The type of landform which describes the feature (e.g. Bay, Lake, Mountain, Community).
- The unique identifier used to identify the feature in the BCGNIS system.

So each BCGNIS "feature" can be spatially modeled as a point, with attribution about the name, typology and unique identifier.

1.2 TRIM Toponymy

TRIM toponymy is cartographic lettering, placed on maps to indicate the names of features. The lettering is placed so as to be legible at a particular representational scale (1:20000) and for a particular representational partitioning of the province (NTS 1:20000 mapsheets).



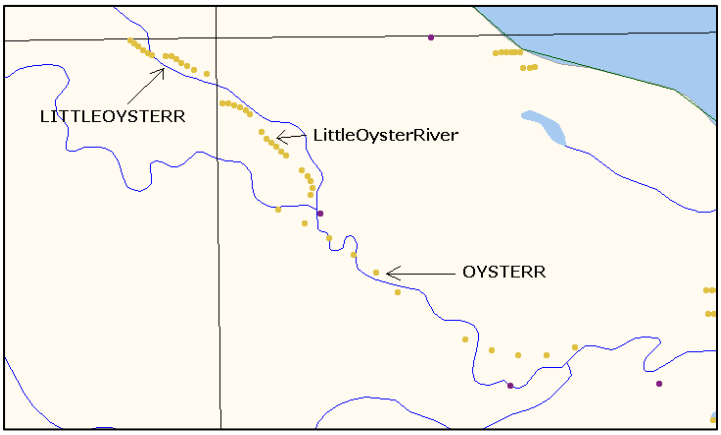
TRIM toponymy was originally compiled as independently placed letters. As viewed by someone reading a map, these letters happened to form words, but there was no explicit relationship between the individual letters in the machine representation.

The Smart Toponymy project processed all the TRIM toponymy letters and formed them into "toponyms", groups of letters which spelled out words. Each of

the words was checked against the BCGNIS database to ensure that it existed as a gazetted name.

The "toponyms" produced by the Smart Toponymy project can be spatially modeled as "multi-point" features, with the following attributes:

- The string which results from forming all the letters in the toponymy together (e.g. "FraserR", "BlueLake", "JohnsonCr")
- A unique identifier for toponym.



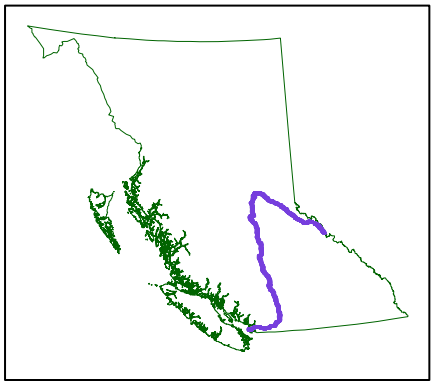
TRIM toponymy will tend to be "redundant" for larger features. Because TRIM toponymy is compiled on a mapsheet basis, large features which pass through several different mapsheets will be labeled several times – once for each mapsheet the feature passes through.

1.3 Joining TRIM Toponymy and BCGNIS

Joining toponymy and BCGNIS is an important first step in "named feature delineation", the transfer of naming information onto the lines and polygons which represent the extents of physical features in a spatial database.

Named feature delineation has many practical uses in cartography and data management, but the immediate purpose of delineation for this project is to provide named rivers into the Corporate Watershed Base project. Named rivers will form the initial basis of a river hierarchy for the province.

BNGNIS records only give one piece of location information – a single point to represent the entire feature. For a small contained feature, like a lake, this information would be sufficient for delineation. But for most larger features, it is not.



Toponyms provide name information across the entire length of a feature, spreading out letters along the whole extent, making delineation relatively straight forward – if the grouping of toponyms can be ascertained. For example, "FraserRiver" and "FraserR" occur hundreds of times in the provincial toponymy data set. If all the "FraserRiver" and "FraserR" toponyms are joined together in a group, the delineation of the Fraser River is a matter of following the path of the toponyms. The figure above shows a picture of all the "Fraser River" toponyms drawn in a line.

The BCGNIS concept of a "feature" is what makes joining toponymy and BCGNIS attractive. If every toponym can be assigned to the BCGNIS "feature" it is supposed to be labeling, then delineation is a matter of taking all the toponyms for a particular BCGNIS feature and looking for the spatial lines or polygons which provide the best match for that feature. It also allows delineation to proceed in typologically sensitive ways – matching the toponymy for the city of Campbell River to the spatial lines which make up the fluid Campbell River could confuse a delineation algorithm. Using BCGNIS typology to subset the toponym data into particular types (e.g. freshwater lakes, rivers and streams, coastal features) allows different algorithms to be used for different delineation problems.

2 MATCHING PROCESS

2.1 Basic Principles

TRIM toponymy has a natural affinity to the BCGNIS database – it is a detailed spatial representation of the name records that occur in BCGNIS. However, **there is no common key between the two data sets**.

In order to create a join between the data sets, we leveraged the similarities in the data using the following assumptions:

- TRIM and BCGNIS both describe features on the ground, so the positional information in the two data sets should be similar. (e.g. Toponyms labeling "Blue Lake" should be in a similar location to the BCGNIS location point for "Blue Lake".)
- TRIM and BCGNIS both provide names for the features, which should be lexically similar. (e.g. BCGNIS will have a record for "Blue Lake" and TRIM may have a toponym for "BLUE LK".)
- TRIM toponymy provides hints regarding the typology of feature names via letter positioning. BCGNIS gives typology explicitly in the database. (e.g. BCGNIS will describe "Blue Lake" as a "Lake" and the TRIM toponym will fall inside the water area of "Blue Lake".)

These commonalties are sufficient to match up the TRIM toponymy and BCGNIS features except for one issue – there are vastly more TRIM toponyms than BCGNIS features. Because TRIM labels the same feature over and over, a single BCGNIS feature can be represented by hundreds (in the case of the Fraser River) or several (in most river and creek cases) separate toponyms.

In order to try and turn the matching process into a one-to-one problem, we grouped the TRIM toponyms using the following assumption:

- In general, TRIM toponyms with the same lexical signature and in the same spatial area describe the same physical feature.

We know that there are exceptions to this rule (for example, the "Campbell River" city/river duality) but we count on there being relatively few exceptions compared to the larger number of cases in which it holds true.

The matching process proceeded through the following steps, which will be described in more detail below:

1. Process the BCGNIS and TRIM name strings and ensure that all TRIM strings exist in BCGNIS.
2. Create groups ("bags") of lexically identical and spatially proximate TRIM toponyms.
3. Match toponym bags with BCGNIS to form spatially proximate and lexically identical pairings.
4. Identify toponym bags with no BCGNIS counterpart and manually determine why they had no match.
5. Identify toponym bags with multiple BCGNIS counterparts and manually determine which BCGNIS records they should match.
6. Rematch toponym bags and BCGNIS, using inputs from manual process to automatically match difficult cases.

2.2 Processing the TRIM and BCGNIS Strings

TRIM toponyms and BCGNIS names are quite different lexically, even when they describe the same feature. The table below gives some examples:

TRIM String	**GNIS String**
PowellR	Powell River
MtRobson	Robson, Mount
LakeoftheWoods	Woods, Lake of the
RockPt	Rock Point

In order to make a lexical match between TRIM and BCGNIS, the name strings needed to be normalized into the same form. The following transforms were applied to the strings:

1. BCGNIS "inverted" names ("Robson, Mount") were reversed ("Mount Robson").
2. All typological identifiers ("River", "Creek", "Point") were abbreviated ("R", "Cr", "Pt"). We abbreviated rather than expanded the typologies because an expansion algorithm would have made more mistakes (e.g. expanding the TRIM toponym for the community of "Stoner" into "Stone River").
3. All strings were capitalized and had white space removed ("Fraser R" became "FRASERR").

Once the name strings were normalized, the TRIM strings were compared with the BCGNIS database to ensure that there were no extraneous cases which did not exist in BCGNIS. Several TRIM toponyms are for features outside of the province – these toponyms where detected by looking for "bags" (lexically identical and spatially proximate TRIM toponyms) that were completely outside the province with no nearby GNIS entry.

This quality control step yielded several TRIM toponyms which had been retired from the BCGNIS database since the completion of the Smart Toponymy project (e.g. “Squaw River”, “Jap Mountain”).

Once it was confirmed that these toponyms were no longer official names, they were removed from consideration for further matching.

2.3 Grouping the TRIM Toponyms

Grouping the TRIM toponyms was called creating “bags” of toponyms. Each “bag” contained toponyms which were lexically identical and spatially close together.

The “bagging” logic is as follows:

- If a named feature passes through several mapsheets, then by the rules of toponymy compilation, there should be one toponym for each mapsheet the feature passes through.
- Furthermore, if there is one toponym per mapsheet, then the farthest one should have to look to find a nearby match for a toponym is about the size of one mapsheet.

For the purposes of bagging, a fixed distance was used as the tolerance to find “nearby” toponyms of the same name, 13 km. 13 km is the size of a TRIM sheet near the southern border of the province. The bagging algorithm proceeded as follows:

1. Find an isolated toponym and add it to a new bag.
2. Find any toponyms within 13 km of the bag that have the same lexical signature and add them to the bag.
3. Repeat 2 until no toponyms are found.
4. Store the bag and go back to 1.

NOTE: Bags are merged together if they are less than 33km apart and there is exactly one GNIS lexical match for the name. This helps in cases where a major river passes through a large lake.

The bagging algorithm is simple, and will make mistakes in several cases:

- Spelling errors in the TRIM toponymy can cause the requirement for a lexical match to fail and toponyms to be excluded.
- Multiple toponyms referring to different features that share a name can be added to the same bag (e.g. “Campbell River” (river) and “Campbell River” (city))
- Gaps in the labeling of a feature can cause two bags to be created where there should be only one.
    - A mapsheet can be missing a label because the TRIM compilers simply forgot to put one in.
    - A river can pass through a large lake and continue on a large distance from where it entered. There will be a corresponding lake-sized gap in the labeling.

We used a simple bagging process and tested it against the data to see how many "exceptional" cases would be generated. The goal was not to completely automate the matching process, but to create a process which produced an acceptably low number of cases requiring manual resolution.

2.4 Toponym / BCGNIS Pairing

The goal of the overall matching process is to ensure that for every TRIM toponym the BCGNIS feature to which it refers is known, to leave no TRIM toponym behind. So the pairing process works against the TRIM toponym bags, taking each bag in turn and looking for potential BCGNIS matches.

One of the assumptions of the matching process is that BCGNIS location points are at least somewhat correct. That is, they may not land precisely **on** the feature they are supposed to represent, but they land **near** to the feature.

GNIS location points are placed differently relative to different kinds of features.

- For lakes, bays, and other "area" features, they are placed near the middle of the feature.
- For mountain ranges and communities and other notional areas, they are placed near the middle off the area.
- For rivers, creeks and other flowing water, they are placed at the mouth of the feature (and a second point is sometimes placed at the headwaters).

If the TRIM toponyms were not "bagged" ahead of time, matching the toponyms to flowing water features or very large area features would be extremely difficult. The "Fraser River" toponym which is placed on a map near Prince George has absolutely no spatial proximity to the "Fraser River" BCGNIS point, which is at the mouth of the river, near Richmond.

Because toponyms are bagged before the matching is done, only one toponym from a bag needs to be acceptably close to a BCGNIS location point in order for a match to be made.

Spatial proximity for the purposes of matching BCGNIS points and toponym bags was considered to be within 13 km. This number was chosen for the same reasons as the 13 km TRIM bagging distance (approximate size of a mapsheet), plus it is large enough to find cases where there are multiple nearby lexical GNIS matches (this forces manual checking of the cases).

Ideally, the pairing process would turn out one BCGNIS point for each toponym bag, and the process would be complete. However, because of various issues outlined below, the following two exception categories were created and populated in the initial pairing process:

- Toponym bags which had no BCGNIS counterpart.
- Toponym bags which had multiple BCGNIS counterparts.

2.5 Toponyms with No BCGNIS Pairings

Toponyms with no BCGNIS pairings ("unmatched toponyms") were investigated manually. There was one valid case in which a toponym bag could be expected to have no BCGNIS pairing – if the toponym was extrajurisdictional.

An ArcView tool was used to quickly cycle through the unmatched toponyms and resolve why they had no pairings.

There were several reasons a toponym could have no pairing:

- **It was extrajurisdictional**. If the toponym fell outside the border it was marked as extrajurisdictional.
- **It was misspelled, so there was no lexical match nearby**. Nearby toponyms and BCGNIS points were examined. Usually a pattern would emerge where all the nearby toponyms would spell a feature name one way, and the toponym under examination spelled it a different way. The mis-spelled toponym was associated with the correct BCGNIS identifier, and a notation was added indicating the spelling error for later reporting.
- **It should have been part of a nearby toponym bag but was left out because of a labeling gap**. The flow of nearby rivers was examined and toponyms of the same lexical spelling were highlighted to identify these cases. The toponym was associated with the correct BCGNIS identifier.
- **The BCGNIS point it should have matched to was improperly located**. All BCGNIS toponyms within two mapsheets were examined to see of there was a BCGNIS identifier with the same or similar name nearby. The toponym was associated with the correct BCGNIS identifier, and a notation was added indicating the BCGNIS location was wrong.
- **It simply did not have a match**. No unpaired candidate BCGNIS feature could be found, neither nearby nor in the BCGNIS database anywhere. No spelling mistakes were noticeable through the proximity of similarly spelled toponyms or BCGNIS points. The toponym was given a code to indicate its general typology (coastal, lake, community, or flowing water).

The problems with BCGNIS and TRIM toponymy identified through this process will be resolved over time (some of the problems have already been resolved as of this writing). The misspellings should be repaired, the BCGNIS locations should be corrected, and BCGNIS records should be added for those toponyms which do not have a BCGNIS match (or the TRIM toponym should be deleted).

2.6 Toponyms with Many BCGNIS Pairings

Toponyms with multiple BCGNIS pairings ("multiply matched toponyms") were investigated manually. There were no cases in which a toponym could be expected to have multiple BCGNIS pairings – all toponyms were expected to refer to a single BCGNIS feature.

An ArcView tool was used to quickly cycle through the multiply matched toponyms and resolve why they had several (usually two) pairings.

There were only a few reasons for multiply matched toponyms:

- **There were two BCGNIS features with the same name and different typologies in the same general area**. This was the most common occurrence. Communities are commonly named after the physical features they are nearby, so most of the cases consisted of a community and a physical feature. For example, "Fanny Bay", "Campbell River", and "Fraser Lake" are all both water features and community names. The cases were resolved by assigning the community BCGNIS identifier to the toponyms which labeled the community and the physical feature BCGNIS identifier to the toponyms which labeled the physical feature.
- **There were two BCGNIS features with the same name and same typologies in the same general area**. This as an error in the BCGNIS database. One of the BCGNIS records was deleted by the BCGNIS maintainer, and the identifier of the remaining entry was assigned to the toponym.

2.7 Completing the Matching Process

Once the manual stages were complete the pairing process was re-run, with a new step added.

When a toponym was found to have been assigned a BCGNIS identifier through one of the manual processes, it was given that identifier immediately without any pairing checks and the process moved to the next toponym.

This allowed the manual products to be integrated into the matching process without editing the source data directly. For example, all the mis-spellings still existed in the toponymy file, but the manual data allowed the matching process to simply ignore them.

3 NEXT STEPS

3.1 Test for Completeness

Now that the matching process is complete, the matched toponyms can be tested for completeness against the original set of toponyms.

Every toponym which exists in the raw, unmatched toponym file must exist in the new matched toponym file, except for extrajurisdictional toponyms.

The raw and matched toponym files will be compared spatially to ensure that no original toponym has been left behind.

3.2 Feature Delineation and Spatial Consistency

The matching of TRIM toponyms and BCGNIS features has been carried out to make feature delineation more practical. As a side benefit, feature delineation will itself feed back into the quality control of the toponyms.

Once toponyms are attached to river and streams, it will be possible to test whether the toponyms meet a number of conditions:

- Is there a toponym in every mapsheet a named river or stream passes through?
- Do all the toponyms on named rivers and streams agree with respect to the flow path of the stream?

Attaching toponyms to lakes will turn up many cases where a single lake feature has multiple names attached to it. The delineation process will provide dividing lines between named lakes to divide single mapped features into multiple named features.

3.3 Potential Process Enhancements

There are a few extra pieces of information which could be used to reduce the number of "no match" and "multiple match" cases.

- **For community / physical feature "multiple match" cases**: community names are almost always placed on a single linear baseline. Therefore, if one toponym is on a linear baseline, and the other is on a curved baseline, the linear baseline toponym can be associated with the BCGNIS community record.
- **Extrajurisdictional determination shortcut**: if a toponym falls more than 150m outside the official boundary of the province, **and** it cannot be matched to any nearby BCGNIS record, then it can be automatically flagged as extrajurisdictional.

4 EXAMPLES OF ERRORS ENCOUNTERED

In all of the example pictures below, the following color conventions apply:

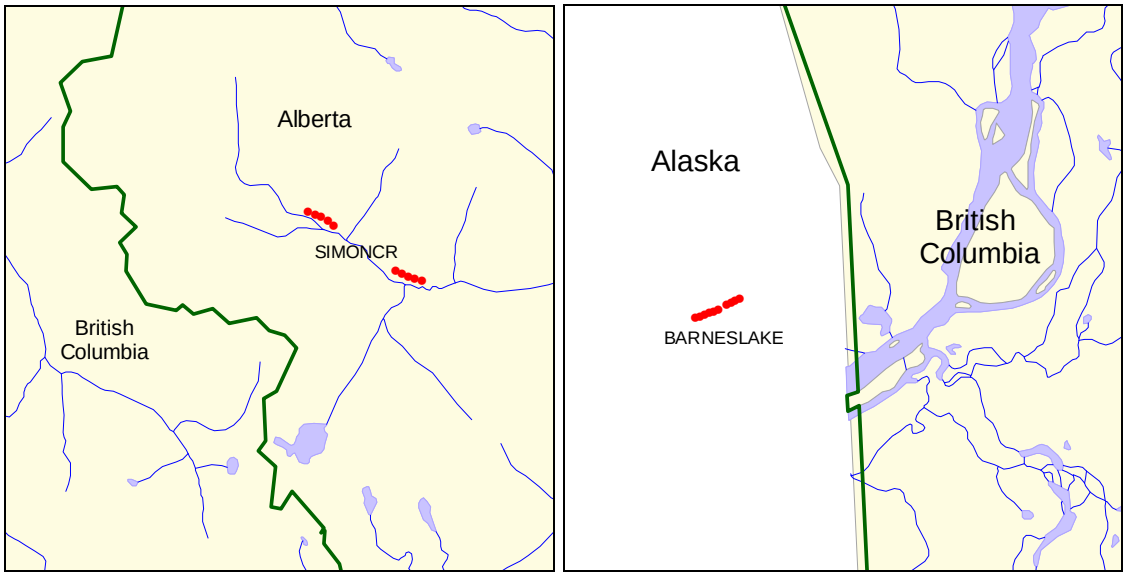
- • : Toponym with an error.
- • : TRIM toponyms.
- • : GNIS location points.
- – : 1:20K NTS mapsheet.
- – : 1:50K NTS mapsheet.

4.1 Unmatched Toponyms

Unmatched toponyms were toponyms for which a nearby BCGNIS entry could not be found.

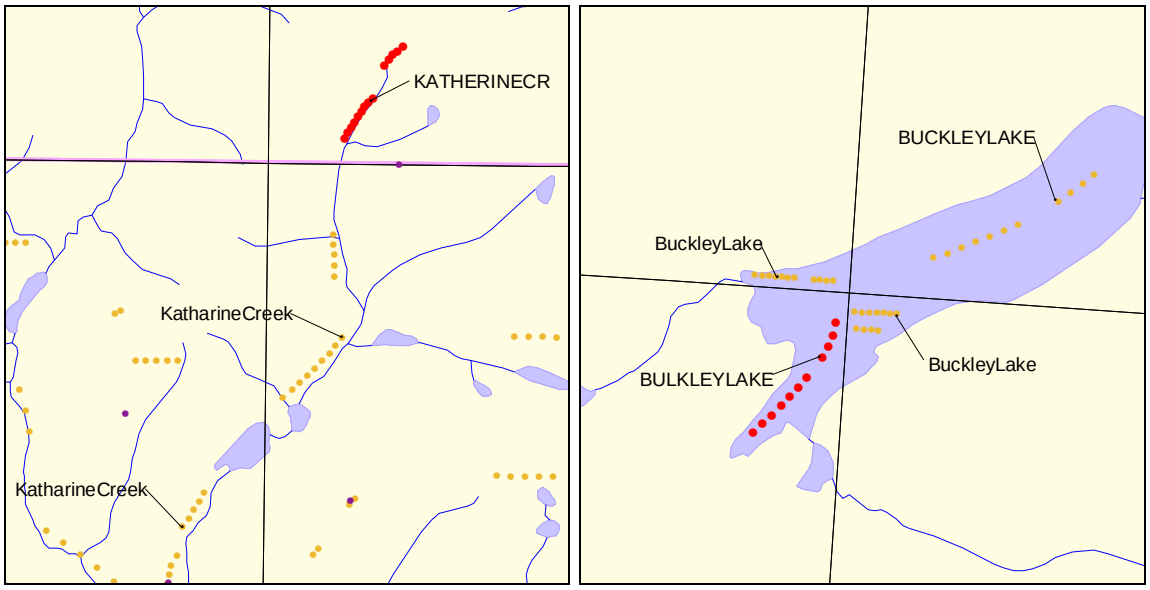
4.1.1 Extrajurisdictional

Extrajurisdictional toponyms exist in the TRIM mapping where a TRIM mapsheet includes areas outside the province of British Columbia. The extrajurisdictional toponyms provide cartographic consistency to the mapsheet. However, because the features they label are outside the province, there is no official gazetted name for them registered in BCGNIS. Therefore there is no possible match for extrajurisdictional toponyms.



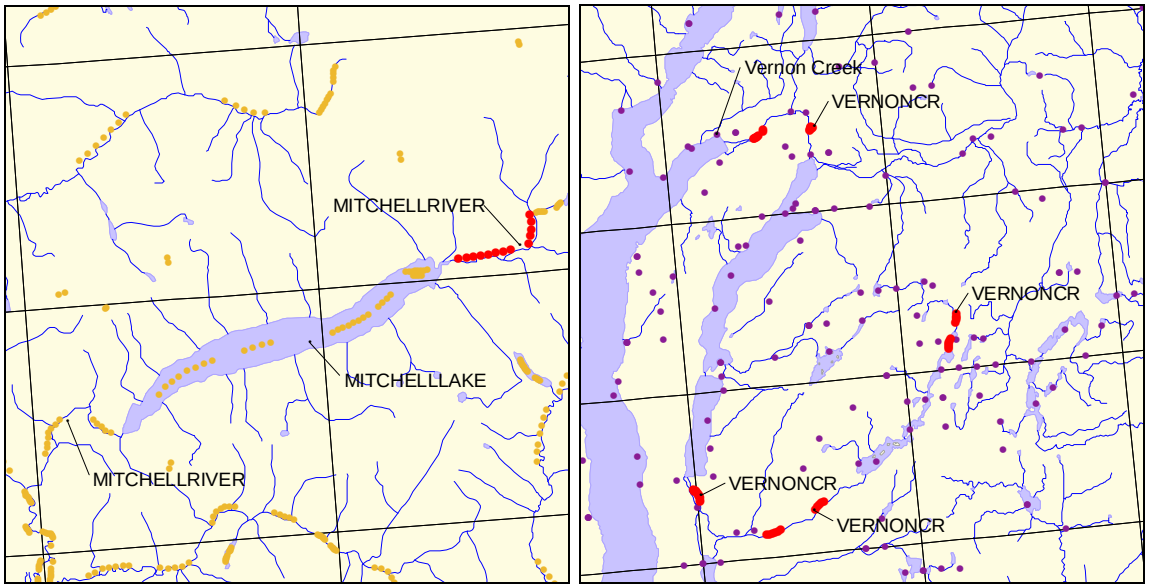
4.1.2 Misspelling

Misspellings were generally easily identifiable by reviewing the neighboring toponyms for consistency with the misspelled toponym.



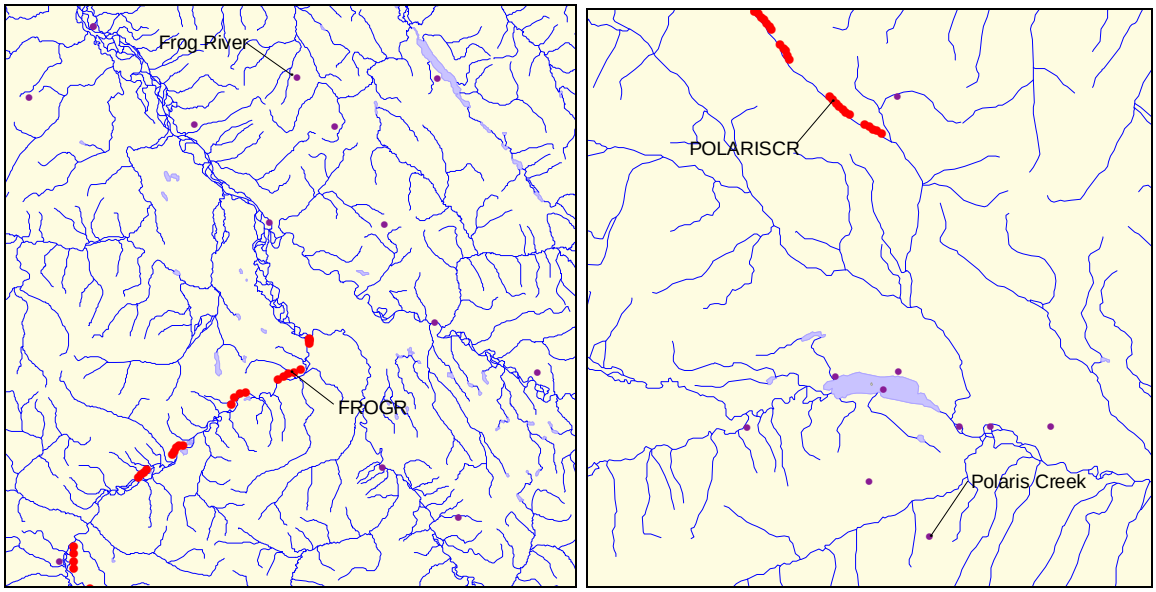
4.1.3 Labeling Gap

Labeling gaps occurred where a labeled feature passed through another feature and re-emerged on the far side. Commonly a creek or river would pass through a long lake, and continue on out the bottom end. The toponyms on the upstream end of the group would be too far away from the downstream toponyms to be included in a “bag”, and so would end up isolated an “unmatched”.



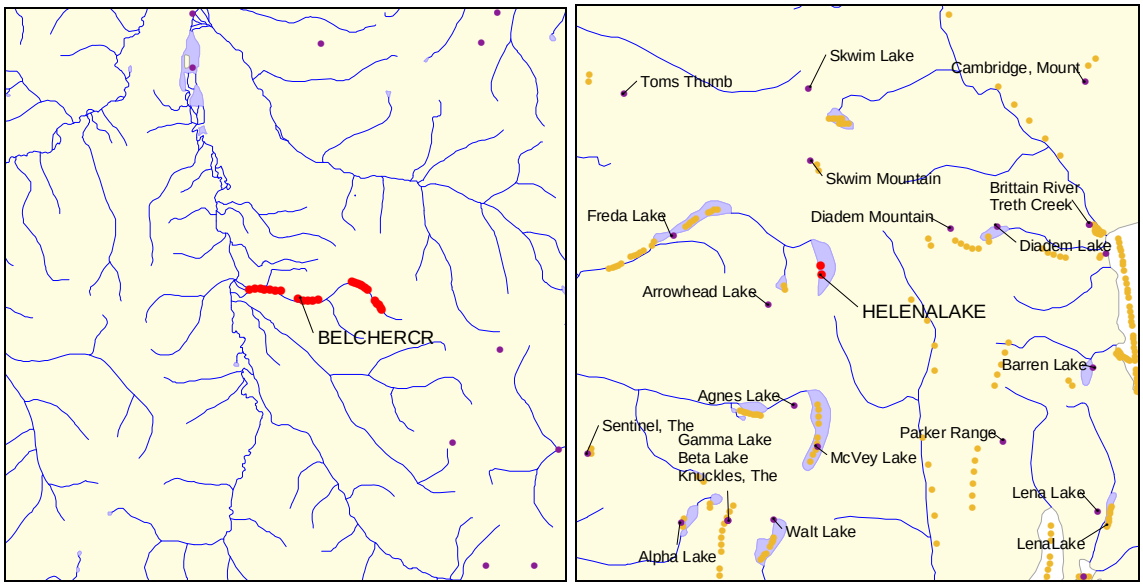
4.1.4 Bad BCGNIS Location

In some cases, the BCGNIS location itself was poorly placed. As a result, valid toponyms were too far away from their proper BCGNIS record for a match to be made. Ordinarily, the BCGNIS point for a creek or river should be near the mouth of the creek or river.



4.1.5 No GNIS Record

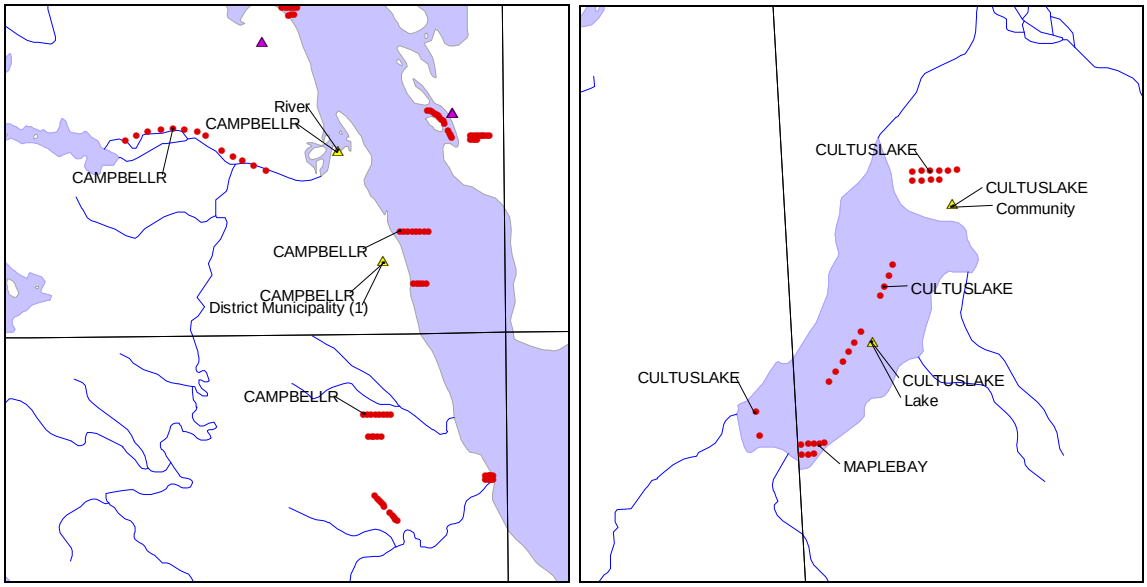
For some of the toponyms, no BCGNIS record existed at all! There are two "Helena Lake" toponyms in TRIM, in two different parts of the province, but only one BCGNIS record for "Helena Lake". As a result, only one "Helena Lake" (the close one) is matched, and the other is orphaned. The same thing occurs for "Belcher Creek".



4.2 Multiply Matched Toponyms

4.2.1 Community / Physical Features with Same Name

Communities are commonly (about 400 times) named after physical features near their location. The result is a and "overmatch", where multiple toponyms match multiple BCGNIS records. The problem is resolved by grouping the correct toponyms with each BCGNIS record, using the cartographic layout as a guide to which toponym represents the community, and which toponym represents the physical feature.

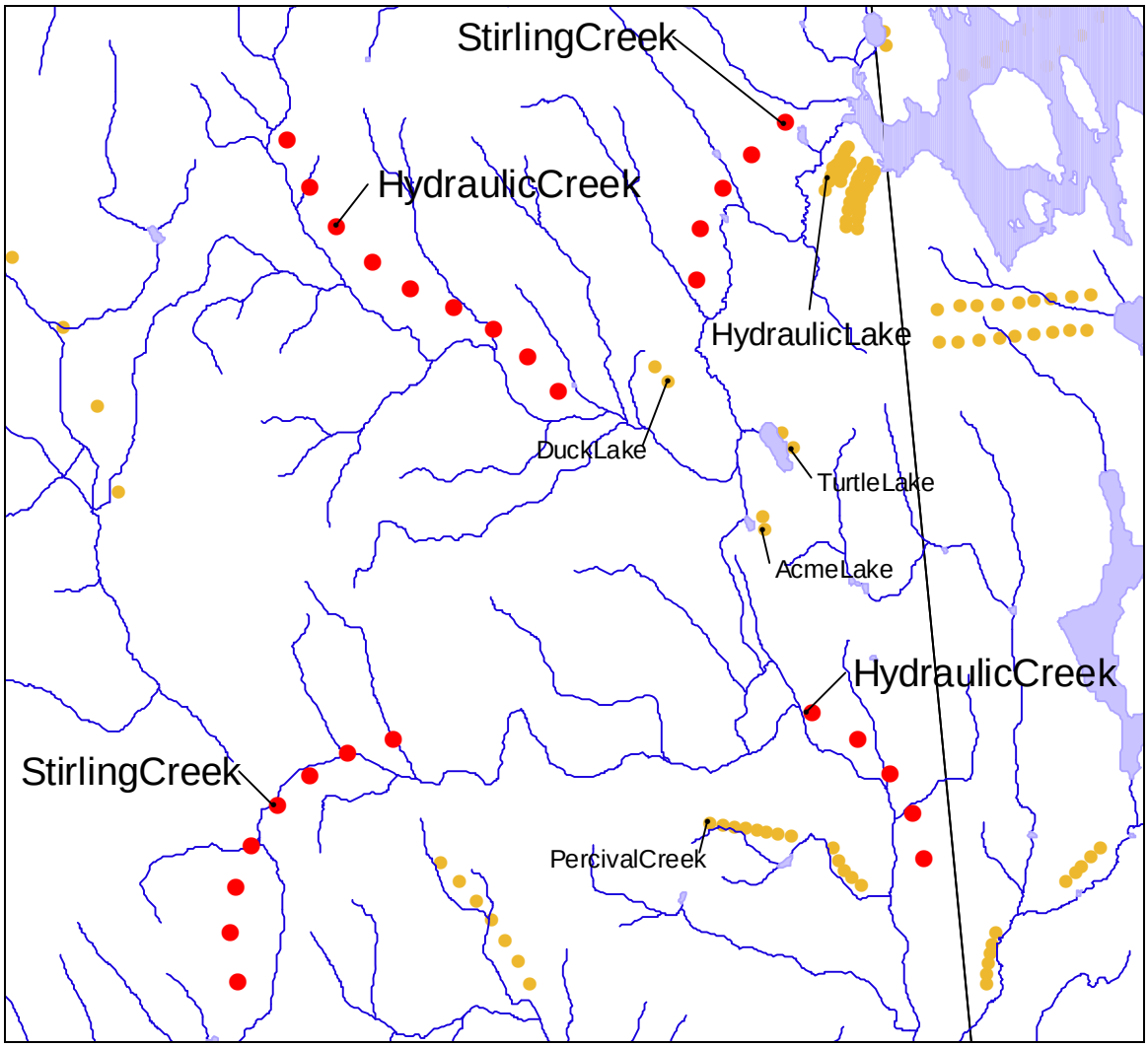


4.3 Other Toponym Errors

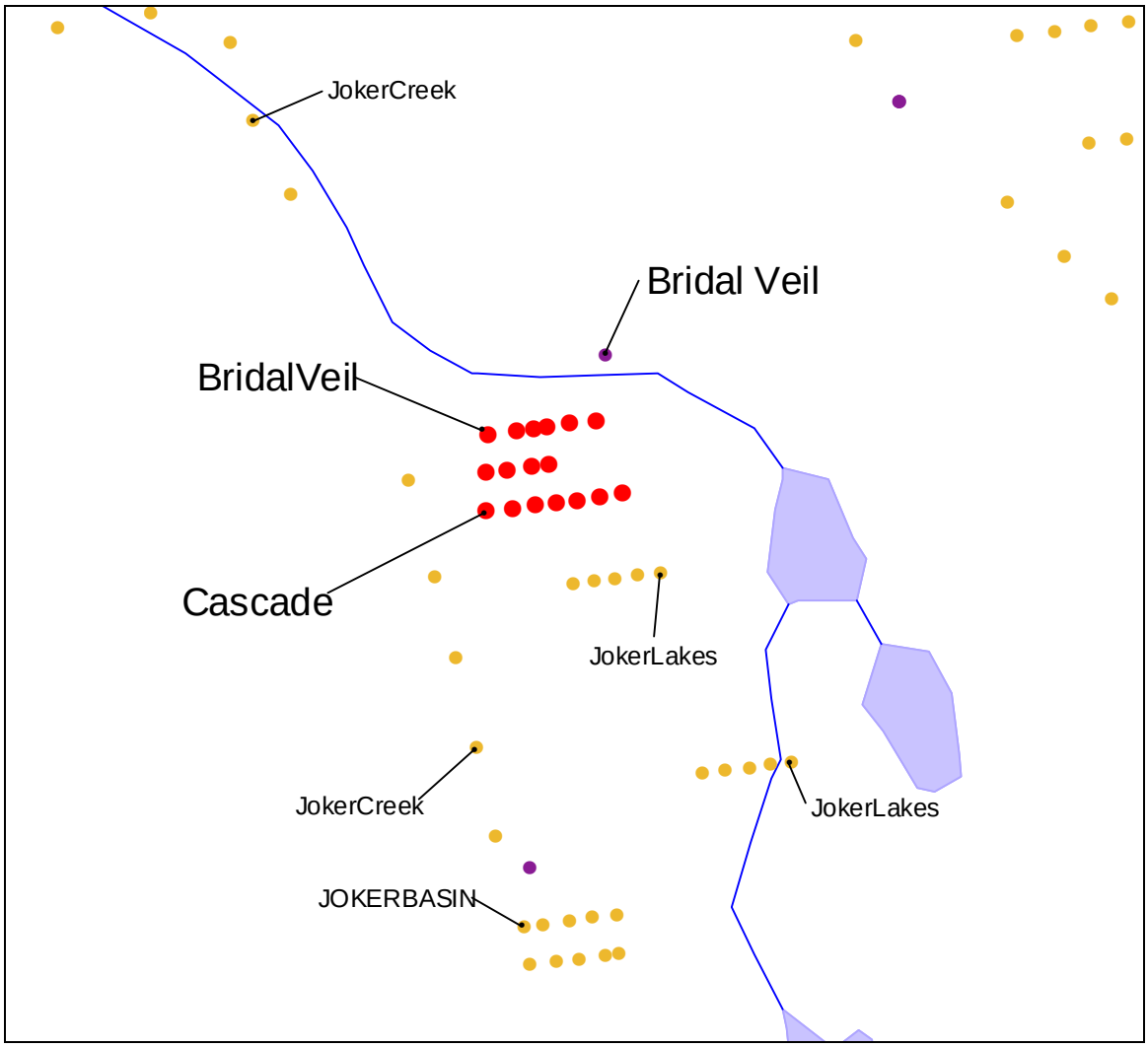
4.3.1 Toponymy Grouping Errors

Toponymy grouping errors are rare cases where the Smart Toponymy process incorrectly grouped letters. Because the results of the Smart Toponymy process were checked against the BCGNIS database, none of the groups is an illegal name. However, in rare cases it was possible for the algorithm to create several valid yet incorrect names from the letters available to it.

In the example below (from 82e074), the "Creek" portion of "Hydraulic Creek" has be exchanged with the "Creek" portion of "Stirling Creek". The presence of "Hydraulic Lake" at the bottom of the "Hydraulic Creek" is a cue to the proper toponym grouping in this case.



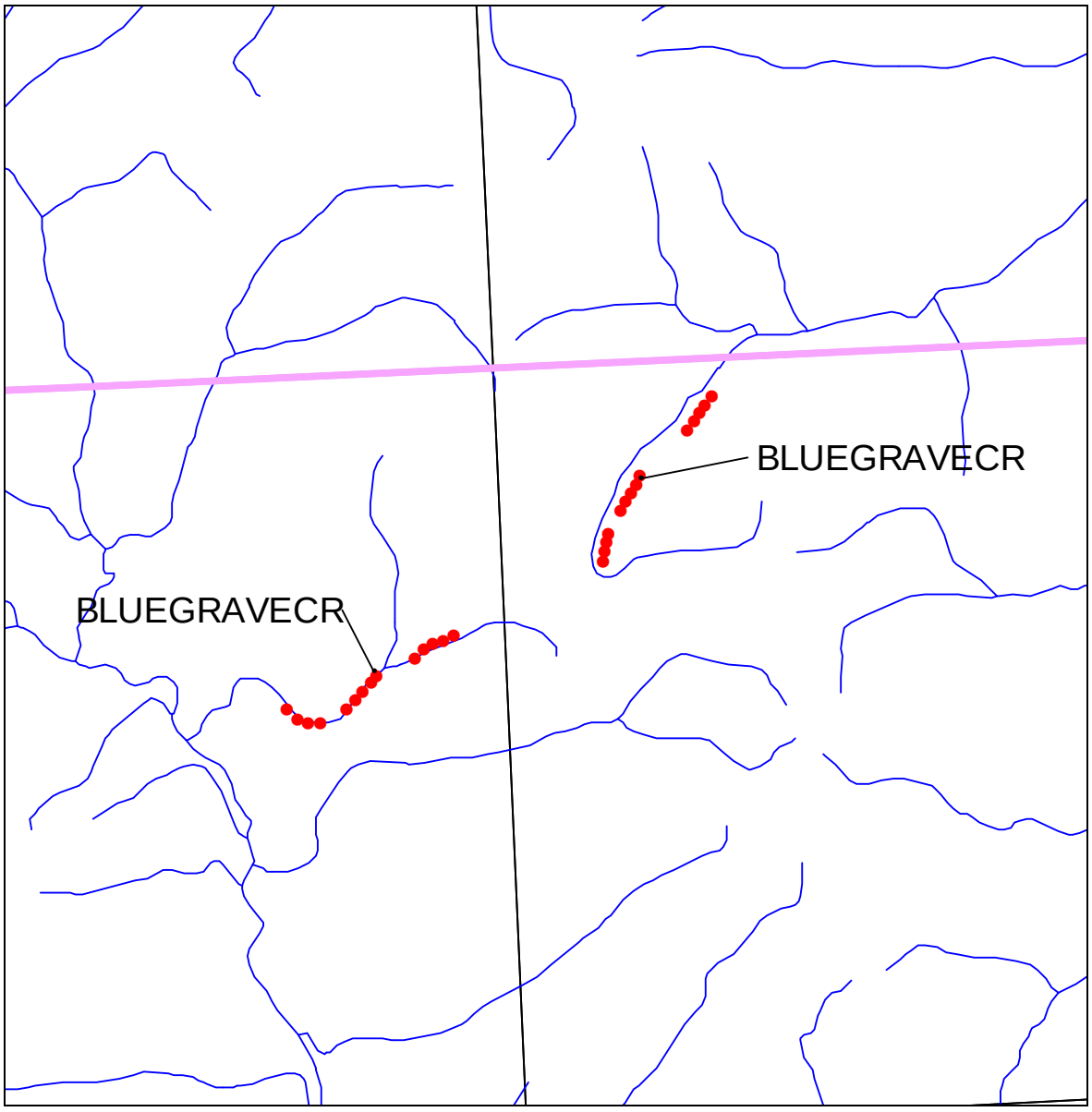
In the example below (from 82f075), the toponym “Bridal View Cascade” has been grouped into two separate toponyms, because the official BCGNIS name is in fact just “Bridal Veil”. In this case, the remaining “Cascade” toponym is extraneous and should be deleted. This case was not detected in the original Smart Toponymy compilation because “Cascade” is a valid BCGNIS entry – there is a community named “Cascade”.



4.3.2 Toponymy Correctness Errors

In some cases, the toponymy is just plain wrong. It is placed in the wrong location, or it carries a name which it should not.

In the example below, "Blue Grave Creek" has been placed on streams which flow in **opposite directions!** Examining the location of the GNIS point (which should adhere to the mouth) indicates that the label on the right is correct and the one on the left is misplaced.



In the example below, “Columbia River” labels have been placed along most of the length of “Lake Revelstoke”. The “Lake Revelstoke” labels can be seen at the north and south ends. Examining the placement of the river labels versus the 50K mapsheet grids yields the reason – the 50K mapsheets from which the TRIM toponymy was compiled pre-date the Revelstoke Dam in a couple cases. In this cases, sheets 82M/15 and 82M/10 predate the dam. The TRIM toponymy compilers placed river labels on the lake incorrectly, because they were following very old source material.