23 June 2017
English
Original: French

Eleventh United Nations Conference on the
Standardization of Geographical Names
New York, 8-17 August 2017
Item 12 (b) of the provisional agenda*
Toponymic data files and gazetteers: Data management and interoperability.

Comment UTF-8 a re volutionne l'e criture des toponymes autochtones

Submitted by Canada**

Summary

Statistics Canada's 2011 Census identified more than 60 Indigenous languages in Canada, and the number of officially recognized geographical names in these languages is constantly growing. The Secretariat of the Geographical Names Board of Canada (GNBC) manages and maintains the Canadian Geographical Names Database (CGNDB), and has developed solutions for displaying these geographical names. While many languages are properly represented using the Latin alphabet (used for English and French), other Indigenous languages require the use of diacritics or syllabics in order to properly spell and represent the geographical names used by Indigenous communities. This paper describes efforts by Natural Resources Canada (NRCan) to represent Indigenous geographical names in Canada. Efforts were expanded tremendously when the CGNDB was converted to the UTF-8 encoding (or *Universal Coded Character Set Transformation Format – 8-bit*).

Background

The GNBC is Canada's national coordinating body responsible for standards and policies regarding geographical names. It is made up of federal, provincial and territorial departments and agencies, each with specific responsibilities for their respective jurisdictions. Members of the GNBC coordinate their efforts in order to manage geographical names consistently. The GNBC is supported by a secretariat within Natural Resources Canada (NRCan), which also provides infrastructure and support for the CGNDB, a national reference database and key component of Canada's Spatial Data Infrastructure. The Secretariat inputs into the CGNDB geographical names and their descriptions, spatial delineations of named features in the form of geometries, and documents related to new naming decisions communicated by the GNBC geographical naming authorities.

The UTF-8 encoding revolution

Before the UTF-8 encoding revolutionized the Internet, programmers had to use the ASCII standard, limited to Basic English characters, and ISO-8859-1, also called Latin-1, which supports most European languages including French with its accented characters. The ASCII and ISO-8859-1 standards have only 128 and 191 characters respectively. To view diacritics, there was no other option but to create an image of these characters in bitmap format, and insert them at the desired location in the character chain.

For example, to display the character è, it was necessary to first create a geographical name supported by ISO-8859-1 by placing markers where the images of diacritics had to be inserted when the name was going to be displayed. The markers were made up of an integer between curly braces (brackets), corresponding to the sequential number of the image to be displayed.

Geographical name entered in the CGNDB:

Ch'in{24}kai Vàn

Corresponding geographical name published on Canada's geographical name query website after the image of the diacritic has been inserted, illustrating the alignment problems that this could entail:
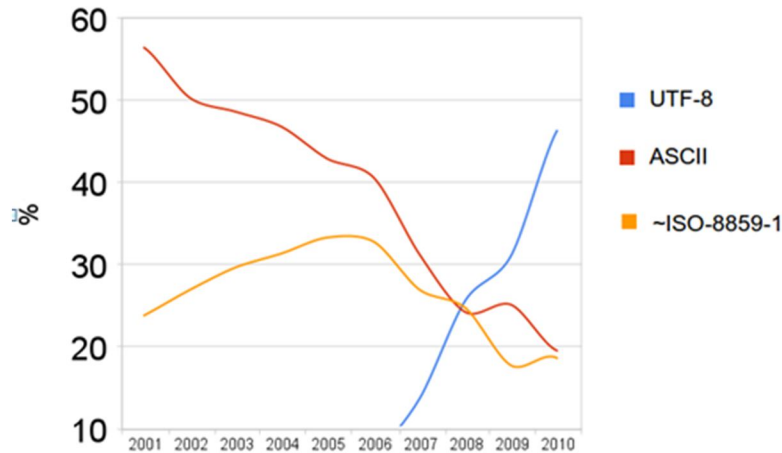
Ch'in ế kai Vàn

Using this process, it was not possible to display geographical names in the Inuktitut language in the syllabic form.

The Unicode standard and its UTF-8 encoding made its debut at the beginning of the 1990s, and makes it possible to display all known characters in any written language, including Indigenous and Asian syllabics. Since the UTF-8 encoding was entirely compatible with the ASCII standard, everything that already existed on the Internet could be quickly converted. Since 2007, the use of UTF-8 has expanded

and it has replaced all other standards in only a few years. It was at that time that NRCan converted the CGNDB to UTF-8 in order to support Indigenous languages in Canada.

Figure 1. Evolution of the UTF-8 encoding on the Internet

Graphic showing the use of UTF-8 exceeding the use of other main character encodings on the Web. In 2010, UTF-8 was used almost 50% of the time. Today, UTF-8 is used on 90% of websites. Source: w3techs.

Conversion of the CGNDB to UTF-8 encoding

In 2008, NRCan undertook a redesign of the geographical names management system in order to adopt the UTF-8 encoding, and to make it possible to disseminate names in Indigenous language. Efforts were first deployed for the database to encode geographical names with UTF-8 in their original form, and second, to convert NRCan websites to display geographical names without having to insert bitmap images. Thus, the geographical name below, designating a lake in the Kutchin-Gwich'in (Loucheux) language, could be displayed in the CGNDB:

> # Ch'inèkai Vàn

To see the name on the NRCan web site, click here.

It was not until 2015, however, that NRCan developed a tool for collaborative management of geographical names completely converted to UTF-8. Today, all provincial, territorial and federal bodies that manage and produce geographical names have access to this collaborative tool and can enter Indigenous names as adopted by the communities into the CGNDB.

The pitfalls of "private use areas"

The Unicode standard is made up of an inventory of 128 172 characters, covering over 100 scripts. Each code point is normally represented by writing "U+" followed by a 4- to 6-character hexadecimal digit. All characters described in this article belong to the Basic Multilingual Plane from `U+0000` to `U+FFFF`.

This plane has a private use area (`U+E000`–`U+F8FF`) for characters which are not standardized but rather left intentionally undefined so that third parties can define their own characters. For example, an Indigenous community may utilize a character to represent a very specific term that does not yet exist in Unicode. That character would be defined in the private use area, and would be assigned a font that must be installed locally in order to correctly display the character.

That is exactly what happened for some Indigenous geographical names in Canada. The name below is that of a bay, written in Tlingit. The name includes a character from the private use area, the *underlined K*, the code point for which is (`U+ EDC4`).

Ch'âk' Kúdi K̲utá

The character displays correctly here because the font used in the text above is *Aboriginal Sans*, developed specifically to display Indigenous languages in Canada. Display problems still occur, however, if the proper font has not been installed. In most cases, as with the very common Arial and Times New Roman fonts, a blank box or a question mark will be displayed instead of the diacritic, because its Unicode code point is not standardized and unknown to most fonts:

Ch'âk' Kúdi □utá

Sometimes another third party has defined a different character for the same code point and the geographical name is altered. Below, the name is displayed with the MingLiU_HKSCS font designed to show traditional Chinese characters.

Ch' âk' Kúdi 僭utá

These display variations cause interoperability and accessibility problems that do not meet Government of Canada standards. A solution had to be found in order to display the geographical names consistently and so that users did not have to download and locally install a specialized font. The answer was composite characters, which are characters that combine and produce basically the same result as private use area characters but through the use of standardized Unicode code points that are not subject to regional variations seen above.

The composite equivalent of K̲ (`U+ EDC4`) is the following sequence (`U+004B` + `U+0332`) where the first code point represents the upper-case K and the second code point represents the line under the K, which is added to the preceding character. As many composite characters can be combined as are

needed to create the final diacritic. Thus, the following sequence is needed to create the character **ą̀**:

(a + U+0328 + U+0300)

Below are examples of geographical names found in the CGNDB that include a composite character without needing to use a specific font. The first line shows the name with the character from the private use area.

Font			
Aboriginal Sans	Ch'âk' Kúdi K̲utá	Shǎr Lūa	Behchokǫ̀
Calibri	Ch'âk' Kúdi K̲utá	Shǎr Lūa	Behchokǫ̀
Arial	Ch'âk' Kúdi K̲utá	Shǎr Lūa	Behchokǫ̀
Times New Roman	Ch'âk' Kúdi K̲utá	Shǎr Lūa	Behchokǫ̀
Verdana	Ch'âk' Kúdi K̲utá	Shǎr Lūa	Behchokǫ̀

Geographical names containing diacritics that are part of the private use area are systematically converted into composite characters in order to ensure that they are displayed consistently for users of geographical name products.

Table 1: Conversion chart for private use area characters used in Canada

Diacritic character	Code point in Unicode private use area	Composite character equiavalent
G̲	EDBC	G + U+0332
g̲	EDBD	g + U+0332
K̲	EDC4	K + U+0332
k̲	EDC5	k + U+0332
X̲	EDDE	X + U+0332
x̲	EDDF	x + U+0332
ą̀	F291	a + U+0328 + U+0300
ā̀	F293	a + U+0304 + U+0300
ä̀	F297	a + U+0308 + U+0300
ä́	F2B7	a + U+0308 + U+0301
ą̂	F2D1	a + U+0328 + U+0302
ę́	F351	e + U+0328 + U+0301
į̀	F3D1	i + U+0328 + U+0300
ī̀	F3D3	i + U+0304 + U+0300
į́	F3F1	i + U+012e + U+0301
ǫ̀	F471	o + U+0328 + U+0300
ǭ́	F495	o + U+0328 + U+0304 + U+0301

ų́	F531	u + U+0328 + U+0301
ų̄	F59B	u + U+0328 + U+0304
ʔ	F861	U+0242

Indigenous languages in the CGNDB

Place names are extremely important for Indigenous communities. They represent their culture and reflect their lives. The knowledge and use of traditional Indigenous names help preserve and strengthen Indigenous peoples' cultures and languages. The CGNDB enables geographical naming authorities to indicate the language of the geographical name that they adopted by choosing from a list of 74 languages established based on the ISO 639-3 standard. Table 2 below contains a list of 30 Aboriginal languages currently used in the CGNDB with some examples of geographical names that can be found on NRCan's query website.

Table 2: Indigenous languages used in the CGNDB

Language	Roman Alphabet	Diacritic	Inuktitut Syllabic
Babine	Det San Ecological Reserve		
Comox	Kwahtums Teeshohsum		
North Slave (Hare)	Deho		
South Slave	Dendale Lake	Nduchįʔelá	
Gitksan	Khutzeymateen Park		
Halkomelem	Slesse Mountain		
Upper Tanana	Kletsan Creek	Tąyh Chìį	
Hän	Chandindu River		
Eastern Canada Inuktitut	Tasikutaak	Kangiqłukuluk	ᐊᕐᑎᕆᐅᖕᑉ
Western Canada Inuktitut (Inuvialuktun)	Amitturyuaq		
Kaska	Itsi Lakes	Eghá' Dā'ǫ́li Lake	
Kutchin-Gwich'in (Loucheux)	Nothlah Hill	Chii Gho' Tł'ąįį	
Kwakiutl	Tsitika Mountain Ecological Reserve		
Michif	Grande Rivière		
Mi'kmaq	Île à Moyacs		
Mohawk	Wahta		
Montagnais	Utshimau-nipi		
Nishga	Gingietl Creek Ecological Reserve		
Nootka	Muqqiwn Park		
Okanagan	sxwexwnitkw park	sẁiẁs park	

Porteur	Bednesti Lake Ecological Reserve		
Salish	Chilliwack River Ecological Reserve		
Sekani	Sikanni Chief River Ecological Reserve		
Tagish	Shootamook Creek		
Tahltan	Ningunsaw River Ecological Reserve		
Thompson	Skwaha Lake Ecological Reserve		
Tlingit	Aishihik River	Tàsłèyi K'ídze Lake	
Tsimshian	Skeena River Ecological Reserve		
Northern Tutchone	Ghechuck Creek	Tawát Mä̀n	
Southern Tutchone	Kluane Lake		

Figure 2: Geographical names as shown on the NRCan query website

Nduchı̨ʔelá

Geographical Names
- Geographical Names Board of Canada
- Data
- Publications
- Search Place Names
  - By Geographical Name
  - By Coordinates
  - By Rectangular Area
  - By Unique Key
  - By Alphabetical List
- Tools and Applications

▶ Instructions: Map Navigation

505m
500m
Trout Lake

ℹ In some instances the feature boundary may not align with the base map due to the scale and datum at which the feature was collected.

Name	Nduchı̨ʔelá
Language	South Slave
Key	LCBWU
Status	Official
Feature Type	Cape
Feature Generic	Point
Location	
Provinces & Territories	Northwest Territories
Latitude - Longitude (DMS)	60° 33' 54" N, 121° 10' 32" W

Conclusion

The CGNDB identifies and represents just over 3500 geographical names of Indigenous origin. That number represents a small proportion of the 392 000 official geographical names in Canada, and some languages are barely or not at all present in geographical name publications and products.

The GNBC has identified increased engagement with Indigenous communities and organizations as a strategic goal, with the intention of accurately recording, storing and disseminating Indigenous place names in the national database. This will result in a higher proportion of Indigenous geographical names, which will be better defined and more representative of the vast diversity of languages spoken in Canada.