E/CONF.105/100/CRP.100

23 June 2017

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**

* E/CONF.105/1

** Prepared by Cindy Doyle Centre canadien de cartographie et d'observation de la Terre, Ressources Naturelles Canada (Canada)

11e Conférence des Nations Unies sur la normalisation des noms géographiques.
New York, 8 au 17 août 2017

Point 12b de l’ordre du jour provisoire
Fichiers de données toponymiques et répertoires géographiques - Gestion de données et interopérabilité

Comment UTF-8 a révolutionné l'écriture des toponymes autochtones

Document préparé par Cindy Doyle
Centre canadien de cartographie et d’observation de la Terre, Ressources Naturelles Canada (Canada)
2017-05-30

Résumé

Selon le recensement de Statistiques Canada, plus de 60 langues autochtones ont été déclarées au Canada en 2011 et la reconnaissance officielle de toponymes dans ces langues est en augmentation constante. Le secrétariat de la Commission de toponymie du Canada (CTC) qui gère et maintient la Base de données des toponymes du Canada (BDTC) a dû s'adapter pour répondre au défi que pose l'écriture de ces toponymes. Si la plupart de ces langues se représentent bien dans l'alphabet latin communément utilisé pour l'anglais et le français, certaines langues requièrent l'usage de caractères diacritiques ou de caractères syllabiques afin d'orthographier et de représenter fidèlement les toponymes tels qu'ils sont utilisés dans les communautés autochtones. Cet article souligne les efforts déployés au fil des ans par le ministère des Ressources naturelles du Canada (RNCan) pour représenter les toponymes autochtones du Canada. Cet effort a connu un essor formidable lors de la conversion de la BDTC en encodage UTF-8, de l'abréviation anglaise *Universal Character Set Transformation Format - 8 bits*.

Contexte

La CTC est l'organisme de coordination national chargé des normes et des politiques en matière de toponymes canadiens. Il est composé de ministères et d'organismes fédéraux, provinciaux et territoriaux, ayant chacun des responsabilités particulières au sein de leur juridiction. Les membres de la CTC coordonnent leurs efforts afin d'assurer la gestion cohérente des toponymes. La CTC est appuyée par un secrétariat au sein du ministère des Ressources naturelles Canada (RNCan) qui fournit l'infrastructure et le soutien à la BDTC, un élément clé de l'infrastructure des données géospatiales du Canada à titre de base de données nationale de référence. Le Secrétariat conserve dans la BDTC les toponymes et leurs descriptions, les délimitations spatiales des phénomènes nommés sous forme de géométries et les documents décisionnels relatifs aux nouvelles dénominations communiquées par les autorités toponymiques de la CTC.

La révolution de l'encodage UTF-8

Avant que l'encodage UTF-8 ne vienne révolutionner internet, il fallait utiliser la norme ASCII limitée aux caractères de l'anglais de base, et la norme ISO-8859-1, aussi appelé Latin-1, qui supporte la plupart des langues européennes dont le français avec ses caractères accentués. La norme ASCII et la norme ISO-8859-1 ne comportent respectivement que 128 et 191 caractères. Pour afficher les caractères diacritiques, il n'y avait d'autre choix que de créer des images de ces caractères en format bitmap et de les insérer dans la chaîne de caractères à l'endroit voulu.

Par exemple pour afficher le caractère è, il fallait d'abord créer un toponyme supporté par la norme ISO-8859-1 en y plaçant des marqueurs là où les images des caractères diacritiques allaient être insérés au moment de l'affichage. Les marqueurs étaient constitués d'un nombre entier entre accolades correspondant au numéro séquentiel de l'image à afficher.

Toponyme entré dans la BDTC

Ch'in{24}kai Vàn

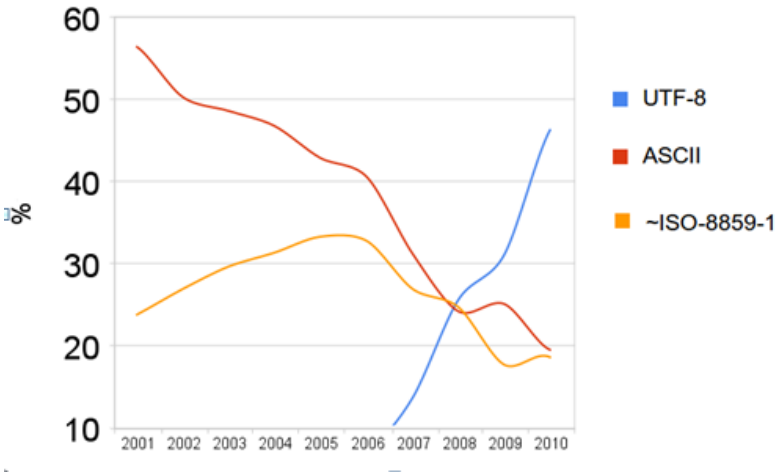
Toponyme correspondant affiché sur le site de découverte des toponymes du Canada après l'insertion de l'image du caractère diacritique avec tout ce que cela pouvait comporter de problèmes d'enlignement.

Ch'in$\overset{\grave{}}{\underset{\breve{}}{e}}$kai Vàn

Quant aux toponymes inuktitut, il était tout simplement impossible de les afficher dans leurs formes syllabiques.

Le standard Unicode et son encodage UTF-8 permettant d'afficher pratiquement tous les caractères connus de toute langue écrite y compris les syllabaires autochtones et asiatiques fait son apparition au début des années 1990. L'encodage UTF-8 étant entièrement compatible avec la norme ASCII, tout ce qui existait déjà sur l'internet a pu être rapidement converti. À partir de 2007, il connait un véritable essor et il supplante toutes les autres normes en quelques années. C'est à cette époque que RNCan a converti la BDTC en UTF-8 afin de supporter les langues autochtones du Canada.

Figure 1. Évolution de l'encodage UTF-8 sur internet



Graphique montrant l'utilisation d'UTF-8 dépassant les principaux autres codages de caractères sur le Web. Vers 2010 la prévalence d'UTF-8 était de l'ordre de 50%. Aujourd'hui, l'encodage UTF-8 est employé par près de 90% des sites web. Source : w3techs.

La conversion de la BDTC selon l'encodage UTF-8

En 2008, RNCan entreprenait une refonte des systèmes de gestion de la toponymie pour adopter l'encodage UTF-8 et permettre dissémination des toponymes d'origine autochtone. L'effort a d'abord été alloué à la base de données afin de saisir les noms en UTF-8 dans leur forme d'origine et, dans un deuxième temps, à la conversion de nos sites web pour afficher les toponymes sans avoir recours à l'insertion d'images. Ainsi, le toponyme ci-dessus, désignant un lac nommé en langue Kutchin-Gwich'in (Loucheux), a pu être affiché à partir de la BDTC. Pour le voir sur le site de découverte de RNCan, cliquer sur ce lien.

Ch'inèkai Vàn

Ce n'est toutefois qu'en 2015 que RNCan s'est doté d'un véritable outil de gestion collaboratif des toponymes complètement converti en UTF-8. Aujourd'hui, tous les organismes provinciaux, territoriaux et fédéraux qui gèrent et produisent des toponymes ont accès à cet outil collaboratif et peuvent entrer dans la BDTC les noms autochtones tels qu'adoptés dans les communautés.

Le piège des Zones à usage privée

Le standard Unicode est constitué d'un répertoire de 128 172 caractères, couvrant plus d'une centaine d'écritures. Chaque point de code est normalement représenté en écrivant l'expression «U +» suivie d'un chiffre en hexadécimal de 4 à 6 caractères. Tous les caractères décrits dans cet article appartiennent au plan multilingue de base de `U+0000` à `U+FFFF`. Ce plan comporte une zone à usage privée (`U+E000`–`U+F8FF`) dont les caractères ne sont pas standards mais plutôt laissés indéfinis intentionnellement afin que les tiers puissent définir leurs propres caractères. Par exemple, une communauté autochtone pourrait développer un nouveau caractère pour représenter un vocable bien précis qui n'existe pas encore dans le standard Unicode. Ce caractère serait défini dans une zone à usage privée et s'accompagnerait d'une police devant être installée localement pour afficher correctement le nouveau caractère.

C'est exactement ce qui s'est produit pour certains toponymes autochtones du Canada. Le nom ci-dessous est celui d'une baie, écrite en Tlingit. Le toponyme comporte un caractère provenant d'une zone à usage privée, le *K souligné* dont le point de code est (`U+ EDC4`).

Ch'âk' Kúdi Kutá

Il s'affiche bien ici parce que la police utilisée dans le texte ci-dessus est *Aboriginal Sans,* développée spécifiquement pour afficher les langues autochtones du Canada. Des problèmes d'affichages sont toutefois créés si la police appropriée n'a pas été installée. Dans la plupart des cas, tels qu'avec les très communes polices Arial ou Time New Roman, une boite blanche ou un point d'interrogation sera affiché

en lieu et place du caractère diacritique puisque son point de code Unicode n'est pas standard et inconnu de la plupart des polices.

Ch'âk' Kúdi □utá

Pire encore, il pourrait arriver qu'une autre tierce partie ait défini un caractère différent pour ce même point de code et que le toponyme s'en trouve altéré. Ci-dessous, le toponyme est affiché avec la police MingLiU_HKSCS conçue pour montrer des caractères traditionnels chinois.

Ch'âk' Kúdi 僵utá

Ces variations dans l'affichage posaient un problème d'interopérabilité et d'accessibilité ne répondant pas aux normes du gouvernement du Canada. Il fallait trouver une solution pour afficher les toponymes de façon constante sans que les utilisateurs aient à télécharger et à installer localement une police spécialisée. La réponse est venue des caractères composés, qui sont des caractères qui s'additionnent et qui permettent d'obtenir sensiblement le même résultat qu'un caractère de la zone à usage privée mais, en utilisant uniquement des points de code Unicode standards qui ne sont pas sujet aux variations régionales observées ci-dessus.

L'équivalent composé du K̲ (U+EDC4) est la séquence suivante (U+004B + U+0332) où le premier point de code représente le K en majuscule et le deuxième point de code représente la barre sous le K qui s'additionne au caractère qui le précède. On peut additionner autant de caractères composés qu'il en faut pour constituer le caractère diacritique final. Ainsi, il faudra la séquence suivante : (a + U+0328 + U+0300) pour former ce caractère : ą̀

Exemples d'affichage de toponymes comprenant un caractère composé venant de la BDTC sans avoir recours à une police particulière. La première ligne contient le toponyme avec le caractère provenant de la zone à usage privée.

Police			
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

Encore aujourd'hui, nous recevons de la part des communautés des toponymes avec des caractères diacritiques faisant partie de la zone à usage privée. Ces caractères sont systématiquement convertis en caractères composés afin d'assurer leur affichage constant pour tous les utilisateurs de produits toponymiques.

Tableau 1 : Grille de conversion des caractères à usage privée utilisés au Canada

Caractère diacritique	Point de code Unicode de la zone à usage privée	Équivalence composée standard
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
ã̀	F293	a + U+0304 + U+0300
ą̈	F297	a + U+0308 + U+0300
ą́	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
ĩ̀	F3D3	i + U+0304 + U+0300
ı́	F3F1	i + U+012e + U+0301
ǫ̀	F471	o + U+0328 + U+0300
ǫ̃́	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

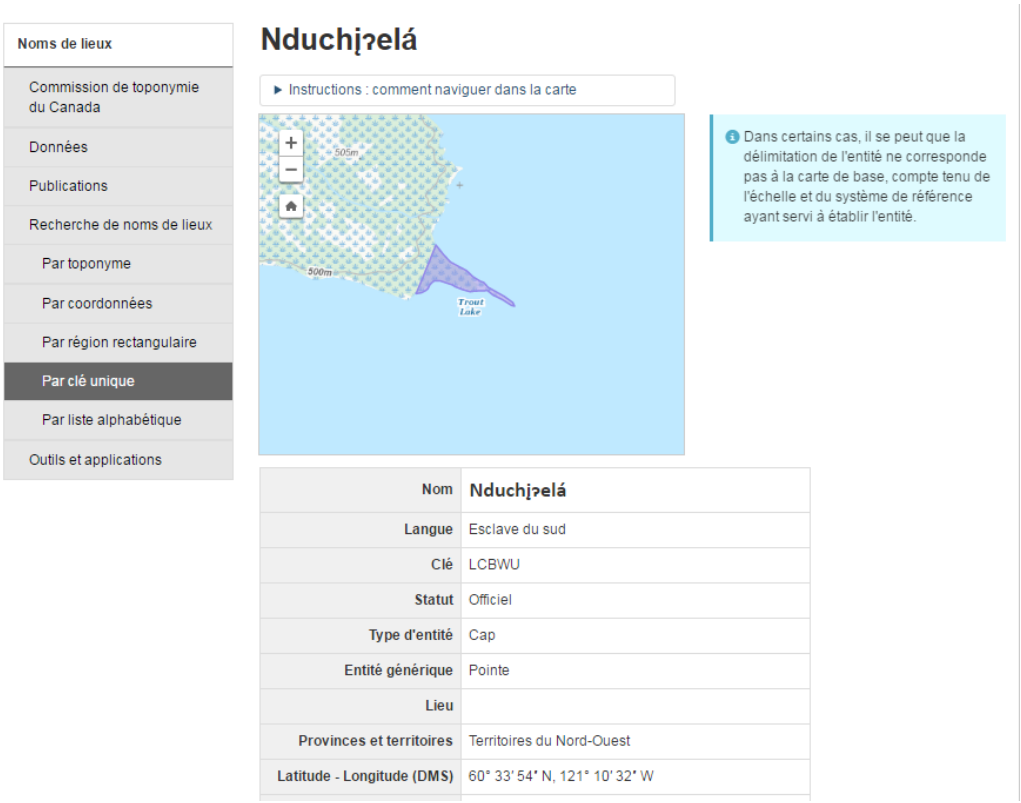
Langues autochtones de la BDTC

Les noms de lieux sont extrêmement importants pour les communautés autochtones, ils représentent leur culture et façonnent leur vie. La reconnaissance et l'usage de noms traditionnels autochtones aident à préserver et à renforcer les langues et les cultures des peuples autochtones. La BDTC permet aux autorités toponymiques d'indiquer la langue du toponyme qu'elles ont adopté en choisissant parmi une liste de 74 langues établie selon la norme ISO 639-3. Le tableau 2 ci-dessous contient la liste des 30 langues autochtones présentement en usage dans la BDTC avec quelques exemples de toponymes qu'il est possible de visualiser le site de découverte de RNCan. De ces 30 langues, 8 utilisent des caractères diacritiques et une seule utilise un syllabaire.

Tableau 2 : Les langues autochtones utilisées dans la BDTC

Langue de la BDTC	Alphabet romain	Caractère diacritique	Syllabaire Inuktitut
Babine	Det San Ecological Reserve		
Comox	Kwahtums Teeshohsum		
Esclave du nord (peau-de-lièvre)	Deho		
Esclave du sud	Dendale Lake	Nduchįʔelá	
Gitksan	Khutzeymateen Park		
Halkomelem	Slesse Mountain		
Haut Tanana	Kletsan Creek	Tąyh Chìį	
Hän	Chandindu River		
Inuktitut de l'est du Canada	Tasikutaak	Kangiqłukuluk	ᓇᖅᑎᓯᐅᖅᓱᒃ
Inuktitut de l'ouest du Canada (Inuvialuktun)	Amitturyuaq		
Kaska	Itsi Lakes	Eghá' Dā'ǿli Lake	
Kutchin-Gwich'in (Loucheux)	Nothlah Hill	Chii Gho' Tɬ'ąįį	
Kwakiutl	Tsitika Mountain Ecological Reserve		
Michif	Grande Rivière		
Micmac	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		
Tutchoni du nord	Ghechuck Creek	Tawā̀t Mān̄	
Tutchoni du sud	Kluane Lake		

Fig.2 Toponyme tel qu'affiché sur le site de découverte de RNCan



Conclusion

Un peu plus de 3500 toponymes d'origine autochtones sont représentés et identifiés dans la BDTC. Ce nombre représente une infime portion des quelque 392 000 noms officiels adoptés au Canada et certaines langues ne sont que peu ou pas du tout présentes dans les publications et les produits toponymiques.

La CTC a établi parmi ses objectifs stratégiques l'implication accrue des communautés autochtones dans le but de représenter et de diffuser fidèlement les toponymes autochtones de la base de données nationale. Il en résultera une proportion plus grande de toponymes autochtones, mieux définis et plus représentatifs de la très grande diversité des langues parlées au Canada.