United Nations E/CONF.101/142

Economic and Social Council

Distr.: General
29 June 2012
Original: English

Tenth United Nations Conference on the Standardization of Geographical Names
New York, 31 July – 9 August 2012
Item 11(d) of the provisional agenda*
Toponymic data files and gazetteers:
Data standards and interoperability

Storage and display of special characters from Aboriginal languages in Canadian geographical names

Submitted by Canada **

* E/CONF.101/1.

** Prepared by Heather Ross, Toponymist, Geographical Names Board of Canada Secretariat

Two long-standing issues in the dissemination of Aboriginal names related to the ability to store and display special characters used in Canadian toponyms in Aboriginal languages.

For example, many of the names in Nunavut are in the Inuit language, often referred to as Inuktitut. The written form of the language uses a writing system based on symbols which represent syllables. Although an internationally-recognized standard font existed for the Inuktitut syllabic characters, these characters could not be stored in the data base, or displayed correctly on the Web site. This issue was resolved when the GNBC adopted the use of ISO standard codes for international languages, and migrated the data base, the Web Feature Service (WFS) and the Web site to new instances which support Unicode UTF-8 character encoding to allow the correct display of these names. Now, when the language code is present in the data base, it identifies the language of a toponym. When the language code attribute contains the code for Inuktitut, a conversion from Roman characters to syllabics is triggered. As new naming decisions are processed, this code is being added to the names records which are created when those decisions are entered in the data base. Not all names in the Inuit language display both writing systems as yet. Work is still needed to add the language code to pre-existing records. However, when all records are updated, the conversion will happen for all Inuit language names.

A slightly different challenge was the storage and display of special characters used in many Aboriginal languages. As most Aboriginal languages were not in a written form prior to European settlement, special writing systems were developed by linguists. These written forms consist mainly of standard Roman alphabet characters, but also include special characters, usually consisting of Roman characters combined with diacritical marks or symbols. The "hard to construct" or Modified Extended Roman Alphabet Characters (MERACs) used in many Aboriginal names presented a very difficult challenge for those wishing to Aboriginal names in databases, or to display them in digital environments. A number of different approaches were adopted at different times to deal with this problem. One solution was to create graphic images of the written names and present those names as an image, in .gif, .jpg, or other format. This was a partial solution, which allowed display on Web pages, but it did not allow the name to be searched easily in data bases or on the Web. A different approach was used in the Canadian Geographical Names Data Base (CGNDB) and on the Geographical Names of Canada Web site. Both used a system of numbers combined with brackets to represent the special characters. For example, a lower case barred L was represented by {2}. Each bracketed number was used to cross reference each character to the image of the character in a table. Although it was better than nothing, this solution was far from ideal, and did not solve the problem of using names containing special characters on maps.

The table below shows a few examples of the special “hard to construct” characters, and the bracket/number escape sequences which were used to represent them in the data base and on the Web site. (The Sorting Value column shows the Roman character which was used to represent the special character for alphabetization and search purposes.)

Table 1

Num Value	**Character**	**Sorting Value**
{1}	Ł	**L**
{2}	ł	**l**
{3}	X̲	**X**
{4}	x̲	**x**
{5}	K̲	**K**
{6}	K̲	**k**
{7}	G̲	**G**
{8}	g̲	**g**
{9}	&	**&**
{10}	į	**i**
{11}	ō	**o**
{12}	ē	**e**
{13}	ū	**u**
{14}	ā	**a**

The table below shows examples of some names containing special characters, and how they were displayed using the character substitutions shown above.

Table 2

Łutselk'e	{1}utselk’e
Délįne	Dél{10}ne
Tēle Lake	T{12}le Lake

As these examples demonstrate, these character substitutions still left much to be desired in displaying Aboriginal names containing special characters.

A new international standard known as Unicode has made it possible to incorporate these special characters into the data base, and to display them correctly on the Web and in documents. Previously, there were hundreds of different encoding systems which were used to assign numbers to special characters. Now Unicode provides a unique number for *every* character, regardless of platform or language. The GNBC has adopted the Unicode standard, and employs the UTF 8 character set. Some of the Web site pages had to be migrated to UTF-8 encoding and were implemented in such a way that they do not require users to download special fonts. These pages have been tested in all browsers. There have been very few issues identified.

The author wishes to thank André Mainville, Team Leader, Toponymy Applications, Natural Resources Canada, for his valuable assistance in the preparation of this paper.