UNITED
NATIONS

E

Economic and Social Council

Distr.
LIMITED

E/CONF.85/L.52
17 July 1992

ENGLISH ONLY

SIXTH UNITED NATIONS CONFERENCE
ON THE STANDARDIZATION OF
GEOGRAPHICAL NAMES
New York, 25 August-3 September 1992
Item 6 (d) of the provisional agenda*

TOPONYMIC DATA FILES: COMPATIBILITY AND STRUCTURE OF SYSTEMS

The treatment of modified extended Roman alphabets and syllabics in Canadian toponymic databases

Paper submitted by Canada**

* E/CONF.85/1/Rev.1.

** Prepared by Helen Kerfoot, Executive Secretary, Canadian Permanent Committee on Geographical Names.

Recording geographical names used by native peoples

The Canadian Permanent Committee on Geographical Names is currently collecting and recording geographical names used by various language groups of native peoples in different parts of Canada. As an integral part of documenting this information for future generations, these native toponyms (together with attribute information) should be stored along with other Canadian geographical names. Today this implies that these names should be included as an integral part of digital geographical names data bases.

Many of the names used by native Canadians are in languages which have accepted writing systems in Roman orthography, but which contain additional characters non-standard to the English- and French-language geographical names now stored on digital data bases. For example, languages of the Athapaskan family, (of northwestern parts of Canada) may contain "hard-to-construct" characters, such as Ī, K̲, é. In addition, several native groups use syllabics, rather than the Roman alphabet, in their written languages. Names in Inuktitut, Cree, Oji-Cree, to mention just a few, may be recorded using syllabaries, which include characters such as ᓴᐱ, ᐋ, ᓚ, etc.

In order to accommodate the use of syllabics and hard-to-construct (or modified, extended Roman-alphabet) characters in digital toponymic data bases, the CPCGN wishes to develop a systematic way of encoding these geographical names. To suit the requirements of the CPCGN members, the approach should be operationally effective in the short term, while still being sensitive to evolving coding standards (national and international), and plans of CPCGN members for future computer systems.

To address these issues the CPCGN Secretariat recently contracted out a study[1] on the treatment of modified, extended Roman alphabets and syllabics in Canadian toponymic data bases. This work involved looking at existing and future needs of CPCGN members, reviewing existing standards, and presenting recommendations on a suitable functional approach to deal now with the addition of syllabics and modified, extended Roman alphabet (hard-to-construct) characters on Canadian toponymic data bases. Eventually, with as little work as possible, the systems of CPCGN members should be made compatible with final international standards.

Some of the findings and recommendations of the contractor's report are summarized here.

Character sets in computer records

It was established that to record adequately all geographical names in Canada, four different character sets must be discussed:

Level 1	ASCII (Latin) Alphabet Character set per ISO 646 (basic Roman)	e.g. e, I, m, #
Level 2 (includes Level 1)	Extended ASCII (Roman/Latin) Character set per ISO 8859 (extended Roman)	e.g. é, Ü, ø, ñ
Level 3	Modified Extended Latin Alphabet (modified extended Roman) - as required for many native Canadian languages	e.g. ę́, K̲, ł
Level 4	Syllabic Characters	e.g. ᑎ ᐃ ᓚᒃ

[1] Undertaken by IDON Corporation, Ottawa. December 1991 - March, 1992.

In effect, level 1 and 2 characters are already accommodated on CPCGN toponymic data bases that store French- and English-language names. Different options for handling level 3 and 4 characters were identified and explored, as suitable international code tables (CCITT or ISO) are not yet available. ISO 10646 (encompassing Unicode - one character, one code) will provide the capability to support virtually all world alphabets on standard commercial systems. However, initially this will not include most native North American languages, which must wait for a future version of the standards.

It is anticipated that standardized commercial solutions for encoding native language character sets could be available by the year 2000. In the meantime, a transition scenario must be selected to gather data so that conversion into a standardized form is possible when formal standards are adopted and commercial products based on the standards are available. The transition approach should be neutral and not in itself try to compete with, or hinder, the definition and adoption of final standards.

Options for storing level 3 and 4 characters

Three options were presented initially for storing modified, extended Roman alphabet and syllabic data:

(1) in an interim standard based on the philosophy of current standards developments

(2) as neutral imbedded symbol numbers, based on the inclusion of ASCII numerals as substitute characters

(3) in a displayable system font on a stand-alone or auxiliary data base on a PC or Macintosh.

Option (1) was ruled out because it would almost certainly be misaligned with the eventual future standard; later conversion might be difficult; it might create tensions with those working on standards; and it might inadvertently delay the adoption of a formal standard.

For Option (3), IBM PC compatible personal computers were discarded at this time, as operationally, costs and difficulties would currently be too great, as special fonts would have to be developed. The Macintosh, however, provides an alternative, with

a) one stand-alone Macintosh data base

b) an auxiliary data-base linked to the primary data base stored on different hardware

c) a Macintosh acting as a terminal interface to a primary data base (i.e. a subsystem approach allowing correct display of imbedded symbols/characters as proposed in Option 2)

The maintenance of primary and auxiliary data bases was considered unduly complex and expensive. Except where data bases already existed on Macintoshes, it was recommended that the Macintosh be used as a possible terminal interface to make a more user-friendly environment for search, read, input and edit purposes.

Option (2) is considered to be the most useful option for the Canadian Permanent Committee on Geographical names, particularly on the national Canadian Geographical Names Data Base (using SUN/UNIX/ORACLE hardware and software).

/...

Two sets of special brackets and a sub-delimiter character, { }, [], | would be selected to identify level 3 and level 4 imbedded neutral symbol numbers, which would themselves be ASCII numerals.

e.g. K' ałgwách Gèwú Ìti
would be coded as: {34}a{28}gwách {30}èwú Ìti

where 28, 30 and 34 identify the 28th, 30th and 34th characters in a list of level 3 modified extended Latin characters

Similarly:
ᐃᕐᒐ ᓗ ᐃᑦ might be coded as [4|13|47|63]

where 4, 13, 47, 63 identify the 4th, 13th, 47th, and 63rd characters in the list of level 4 syllabic characters.

The list of level 3 and level 4 characters would grow as new characters are encountered in geographical names records. A paper copy registry of the coding would be centrally maintained. At this time, there is also a project in hand in Canada (outside the realm of geographical names) to define character sets for computer coding of syllabics and hard-to-construct characters from some 19 native languages. By keeping in touch with this group the CPCGN can learn of their advances in terms of the Canadian Standards Association and also possible extensions to ISO 10646 to meet the needs of Canadian native languages.

All geographical name records which include level 3 or level 4 modified, extended Roman alphabet characters entered into the CGNDB will need to be identified by an orthographic flag at the record level. Each level 3 or level 4 character will be represented by a unique number. Software will need to be written to allow the laser printer to print these characters in an appropriate font. In addition, with appropriate software, it would be possible for X, which might be represented by {35}, to be shown by a rather more user-friendly alias, such as {X_}. If a Macintosh computer sub-system can be established, this would facilitate large quantities of data entry, editing or searching, as the imbedded symbol numbers would be displayed correctly in a user-friendly way. (The main problem here is the choice of fonts available on the Macintosh. No one font is the standard; usage varies from region to region, based on various option key approaches.)

Conclusion

In summary, it is hoped that the selection of the neutral imbedded symbol numbers approach to data entry of native geographical names will prove useful in Canada over the next few years. Such an approach will allow CPCGN members in different provinces and territories of Canada to work to the same standards, which will permit exchange of data with translation software kept to a minimum. To modify current systems to accommodate input of imbedded symbol numbers for level 3 and 4 characters is quite inexpensive; additional printing and viewing/editing capabilities can be developed as need arises and funds become available. This approach also prevents the adoption of partially developed standards, which will in future years likely prove to be incorrect and costly to correct. At the same time this approach facilitates changing to ISO (Unicode) standards when they are developed and commercially available.
