UNITED NATIONS
GROUP OF EXPERTS
ON GEOGRAPHICAL NAMES

WORKING PAPER
NO. 43

Twenty-sixth session
Vienna, 2-6 May 2011

Item 9 of the Provisional Agenda

Activities relating to the Working Group on Toponymic Data Files and Gazetteers

Text-encoding issues in geographical names*

* Prepared by the Working Group on Toponymic Data Files and Gazetteers

Text-encoding issues in geographical names

Document prepared by the Working Group on Toponymic Data Files and Gazetteers

In today's digital world, the question of text encoding is impossible to ignore. Geographical names written in non-Roman scripts or with their inherent diacritical marks or special characters can become scrambled when used in a digital environment. Without language support and compatible fonts, sharing and viewing such data can be problematic. Difficulties emerged as digital systems and software developed independently and different encoding systems[1] were being applied to the letters and characters used in geographical names, resulting in this data incompatibility and incorrect character display.

UNGEGN realised a number of years ago that this lack of conformity could impede the successful processing, international exchange and promotion of nationally standardized geographical names. A Working Group on Toponymic Data Exchange Formats and Standards was therefore formed during the 18th UNGEGN in 1996 to establish if an international standard existed that would fulfil requirements for the encoding of characters used in Roman-script geographical names.

Over a period of two years the Working Group conducted a worldwide survey of the main Roman-script alphabets and romanisation systems required for digital geographical names processing, drew up character inventories, and studied existing encoding standards, including 8-bit and 16-bit. It concluded that the existing ISO 10646 standard as reflected in the Unicode Standard[2] fulfilled most requirements. The Unicode Standard contains tables of characters, each allocated a unique number, or 16-bit encoding. It works in parallel with ISO/IEC 10646, and is designed to be independent of platform, program or language. The Working Group's findings were recorded in a comprehensive report which was presented to Seventh UN Conference on the Standardization of Geographical Names, New York, 1998, as E/CONF.91/CRP.11.

At that point the Working Group was absorbed into the Working Group on Toponymic Data Files and Gazetteers as its project had been completed. When UNGEGN published its *Technical reference manual for the standardization of geographical names*[3] in 2007, the contents of the Working Group's report were updated to reflect changes to ISO 10646/Unicode, and to make the report itself Unicode compliant.

Having studied Roman-script names, the Working Group raised the question as to how far Unicode catered for non-Roman scripts used in geographical names and how the Unicode Consortium[4] could be approached with any gaps identified. UNGEGN was subsequently granted Liaison Membership of Unicode in 2002, with Mrs Caroline Burgess of the United Kingdom as the principal point of contact. This liaison status gives UNGEGN full access to documentation, email lists and technical committee meetings and allows input into the ongoing development of the Unicode Standard.

During a meeting of the Working Group on Toponymic Data Files and Gazetteers in Zagreb in February 2011, it was decided to initiate a project to collect data from UNGEGN members highlighting gaps in the Unicode Standard encountered in the processing of digital geographical names data, not already covered by the Working Group on Toponymic Data Exchange Formats and Standards' earlier study. Information received will be collated into a submission to Unicode on behalf of UNGEGN. Attached is a sample form which we hope UNGEGN members will use to submit information. Forms may need to be completed in manuscript and can be sent to Mrs Burgess, at the address provided on the form, as the Working Group on Toponymic Data Files and Gazetteers' point of contact for Unicode Liaison.

[1] Where numbers are assigned to characters to enable a computer to display the text

[2] See www.unicode.org; the latest edition is version 6.0.0

[3] Available to download in pdf format at http://unstats.un.org/unsd/geoinfo/UNGEGN/docs/pubs/UNGEGN%20tech%20ref%20manual_m87_combined.pdf

[4] The Unicode Consortium develops and maintains the standard. It is a non-profit making organisation whose members are drawn from the computer and information-processing industry

Contributors should be aware that individual encoding problems may be due to font deficiencies, rather than omissions in Unicode, as the appearance of a character in the Unicode Standard does not guarantee its inclusion in any given font package. It is entirely the responsibility of the font producers to select which characters to include in their font packages. Also, despite its extensive scope, the Unicode Standard is being continually updated and there are some characters which are in the process of receiving codes and therefore may not yet be included in the latest version of the Standard. All existing and planned Unicode character encodings can be found at http://www.unicode.org/standard/where/ Information on Unicode-compliant fonts is available at http://www.unicode.org/resources/fonts.html

Thank you for your cooperation.

Working Group on Toponymic Data Files and Gazetteers
Text-encoding issues in geographical names

1. UNGEGN Division or Working Group:

2. Point of contact (in the event of any queries on the content of this form):

3. Character/s used in geographical names not identified in ISO/IEC 10646/Unicode (please draw the character and provide its name and any other relevant information, including context of its use and any associated glyphs)

4. Language and script in which above characters appear:

5. Non-Unicode fonts available (if any) to display character:

6. Any other information:

7. Submitted by: (Name and affiliation. Please also provide contact details if different from 2 above)

Please kindly return this form to:

Mrs Caroline Burgess, PCGN, c/o Royal Geographical Society, 1 Kensington Gore, London, SW7 2AR, United Kingdom, or email to cburgess@pcgn.org.uk