ISO/IEC JTC1/SC2/WG2 N4925
L2/18-005
2018-01-11

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **Working Group Document**
Title: **Naming conventions for larger character sets**
Source: **Michael Everson and Andrew West**
Status: **Individual Contribution**
Action: **For consideration by JTC1/SC2/WG2 and UTC**
Date: **2018-01-11**

0. Preliminaries. A recent comment in L2/17-353 "WG2 Consent Docket" suggests that Shuishu character names have "significant issues". Such issues have not been put forward in a discussion document by anyone from the UTC, and at the WG2 meeting in Hohhot when the editor suggested that the names were problematic, it came as a surprise to everyone who had worked on Shuishu. Pressure to suddenly change character names which have been established throughout the encoding process is unwelcome, particularly when no evidence is given to support a suggestion that the names are in fact "problematic".

Character names in the UCS are intended to identify the meaning of characters. There are three ways of doing this in the UCS: with actual names, with catalogue numbers, and with algorithmic names. This document sets out principles by which these options can be chosen, and recommends which sort of names should be used for the unencoded scripts on the Roadmap.

1. Meaningful names. Ordinary character names are the most convenient to the end user of the standard, and the bulk of non-CJK character names in the UCS use these. Names for Old Bamum, for instance, give the logographic readings of all of the characters from Phase A through Phase F. The modern Bamum syllabary uses the syllabic names. Cuneiform and Early-Dynastic Cuneiform signs are also given meaningful readings, and Cuneiform is a large character set. Where a script is an alphabet, a syllabary, or a logography, most characters have only one reading. The best practice is to use such names unless there are very good reasons for using something else. (Programmer preference and dislike of "funny-looking" names are not good reasons.) We recommend that the following scripts on the Roadmap for the SMP be encoded using meaningful names:

Afaka, Baburi, Balti, Blissymbols, Book Pahlavi, Cirth, Dhives Akuru, Eebee Hmong, Elymaic, Eskaya, Garay, Gurung Khema, Jenticha, Kawi, Kerinci, Khambu Rai, Khimhun Tangsa, Khotanese, Khwarezmian, Kirat Rai, Kpelle, Kulitan, Lampung, Landa, Leke, Loma, Mandombe, Micmac Hieroglyphs (perhaps), Moon, Mossang Tangsa, Mwangwego, Palaeohispanic, Pau Cin Hau syllabary, Pitman Shorthand, Proto-Cuneiform, Pyu, Ranjana, Shavian Quikscript, Shuishu, Tengwar, Tikamuli, Tocharian, Todhri, Tolong Siki, Uighur, Vatteluttu, Vithkuqi, Wancho, Western Cham, Woleai, Zou.

The majority of these scripts are alphabets, abugidas, or syllabaries. Naxi Geba is a syllabary, Shuishu (486 characters) is a logography just as Old Bamum (569 characters) is. Naxi Dongba as intended for use by the modern user community as a syllabary proper.

2. Catalogue names. Some scripts are by their own users chiefly or partly identified by catalogue number. The earliest such script encoded was Linear B, where decipherment has been successful for the syllables and many units of measure, but there remain some characters whose identity is only described by the catalogue number. Egyptian Hieroglyphs (deciphered) and Linear A (undeciphered) use catalogue names, as do Anatolian Hieroglyphs (partially deciphered). We recommend that the following scripts on the Roadmap for the SMP be encoded using catalogue names, since this enables users to relate them to the existing taxonomies and reference materials:

Cypro-Minoan, Egyptian Hieroglyph Extensions, Indus Valley, Maya Hieroglyphs. Bagam, Micmac Hieroglyphs (perhaps), Rongorongo, Voynich

3. Algorithmic names. Algorithmic names are not particularly convenient for users of the code charts or names lists, but there are a few good reasons for using them. CJK is the paradigmatic example. So many characters have multiple readings even within a single language, and there are so many of these characters in all, that a suite of database entries are the best way of providing additional information on each character. Nushu (396 characters) is a good example of a script where many individual characters have multiple meanings. Tangut (characters) is only partially deciphered and has issues with transcription and polyvalence; algorithmic names are useful for Tangut too. We recommend that the following scripts on the Roadmap for the SMP be encoded using algorithmic names:

Jurchen, Khitan Large Script, Khitan Small Script

All three of these are large and partially deciphered, with multiple readings for a reasonably large number of characters.

We believe that accepting these guidelines will enable SC2 and UTC to work together with less stress and fewer misunderstandings in future.