X3L2/96-055

ISO
International Organization for Standardization
Organisation Internationale de Normalisation

ISO/IEC JTC 1/SC 2/WG 2
Universal Multiple-Octet Coded Character Set
(UCS)

ISO/IEC JTC 1/SC 2/WG 2 N 1383
April 25, 1996

Title:	Initial Comments on encoding Mongolian into ISO/IEC 10646
Source:	Ad-hoc group (China, Mongolia, U.K., Ireland, Unicode Consortium)
Status:	Ad hoc contribution
Action:	For the consideration of SC 2/WG 2
Distribution:	ISO/IEC JTC 1/SC 2/WG 2

1 Document Status and Present Situation

The document N1368 containing a very complete list of glyphs used for writing Mongol, Manju, Tod and Shibe (called Xibe in the document) was produced by Mongolia and China as a joint effort after the ISO meeting in Tokyo, November 1995.

Drafts of the document were available at earlier stages of its production through spring 1996 but a final version was only made available immediately before the meeting. This version, labelled *Zhong Meng Lianhe Cao'an.doc* "Chinese-Mongolian Joint Draft Proposal", is dated April 10, 1996 (printing date as shown in label is April 13, 1996).

Since certain issues relevant to the problem of encoding Mongol and its derived scripts had not yet been addressed by this draft, a separate discussion group was formed with the following participants of the main conference (in alphabetical order):

- O. Chilkhaasuren, MNISM, Mongolia representative
- Oliver Corff, UNU/IIST
- Myatavyn Erdenechimeg, UNU/IIST
- Michael Everson, Ireland representative
- Asmus Freytag, Unicode representative
- Mao Yong Gang, Chinese Electronics Standardization Institute
- Hugh Ross, UK representative.

2 Target Issues

The present draft is evidently a complete listing of all possible characters and glyphs found in the Mongol script as well as its derivatives (Tod, Manju, Shibe) and extensions (Aligali – mainly used for transliterating Sanskrit and Tibetan words). Two tables, 1. *Basic Mongolian Set* (423 used positions), and 2. *Mongolian Presentation Form Set* (149 used positions), hold a total number of 572 entries. Both tables contain eleements that ISO/IEC 10646 would consider presentation forms. A look at the the second table in the draft shows that many of the forms given there are ligatures, or more correctly called *logotypes*, as Mr. Ross pointed out.

2.1 Conflict Potential between the draft and existing Encoding Principles

Within the ISO/IEC 10646 and the Unicode Standard there is agreement that a character encoding covers canonical characters, not special forms used for presentation only. The latter are to be produced by a rendering mechanism as long as the rules inherent to the script of a given language permit this action without referring to a dictionary or similar mechanism. In cases where a decision cannot be completely based on rules it is up to the user to indicate explicitely the desired results. In the case of Arabic which has a similar shaping behaviour this is done by inserting special, zero-width characters (e.g. *joiners, non-joiners*) which provide a selection mechanism for the desired character form.

Any single shape for two or more characters that is produced by a purely rule-based mechanism is a so-called *mandatory ligature.* Mandatory ligatures are not encoded in the character set. There are, however, also style-dependant ligatures.

In Mongol writing, we can observe *free* variations of medial and final forms, free to the respect that the actual choice is based on the writer's linguistic competence and that this choice is not determined by the immediate character context. Unfortunately, as may be deducted from the present draft, the number of possible choices might be as high as five for some final forms.

ISO/IEC 10646 contains a generic zero-width joiner and non-joiner symbol eliminating the need for implementing these again in the Mongol encoding. There is, however, the special problem that Mongol vowels and certain consonants (notably [k, g, γ] and [x]) are sensitive to the vowel gender. Vowel harmony is responsible for combinatory constraints which are reflected in the writing system. In order to match these constraints, it may well be necessary to include additional, vowel gender specific joiner/non-joiner symbols in order to override the rule-based constraints at certain times.

2.2 Requirements of Documentation

Due to the particular nature of the Mongol script, one must define the logical constraints before actually selecting glyphs out of a huge repertoire as candidates for an encoding.

A final document must provide an explanation how the proposed encoding addresses the constraints mentioned above and answer questions like shape conventions used and how selective coding of explicite final forms is done. Based on the proposed enccoding, it also needs to address issues such as

- sorting of,
- searching through and
- indexing texts,

especially with *artificial* contexts in mind, i.e. context produced by the user with the help of joiners and non-joiners. Generally joiners are ignored in sorting and searching. Any exception to this for Mongolian needs to be motivated.

2.3 Input of words with unknown or unorthodox writing

Another issue which was not addressed so far concerns the question of keyboard input and catalogue search (e.g. in libraries) of words with unknown pronounciation and/or unorthodox spelling. It should be possible to have a non-semantic approach which accesses letters only according to their shapes, not to their semantic contents.

Since all glyphs which could perform this function are already contained in the encoding, the only task will be to define a canonical list of glyphs serving as *bellies, teeth* etc. (using Mongol conventional names once). No additional encoding is needed; the list of canonical glyphs will be informational part of the documentation accompanying the encoding.

2.4 Suggested Actions

The present draft, brought forth jointly by the Mongolian and Chinese national bodies, is an important achievement since for the first time it presents an exhaustive repertoire of Mongol and related characters, glyphs, and logotypes (ligatures). The completion of this draft as such is indeed the first out of four necessary steps to proceed towards a working character encoding. As a second step, the semantically relevant characters, i.e. the ones carrying the information have to be isolated. Third, the coding conventions of these characters have to be analyzed. As fourth and last step, the presentation conventions must be analyzed. Only then a feasible *character* encoding can be defined.

3 Preliminary Results

3.1 Present Status

The Mongolian and Chinese delegates presented information about document N1368, "Joint Proposal Draft on encoding Mongolian Character Set".

3.2 Suggested Meeting

The Mongolian and Chinese participants suggested to hold another working meeting for solving all pending issues in early August 1996, in Inner Mongolia, China.

The target of this meeting will be to produce an encoding working draft compliant with the requirements mentioned in the previous section.

3.3 Immediate Follow-up Action

The result of this meeting will be presented to the Unicode Consortium and other members of the ad-hoc Committee for an off-line review and feedback on the proposal.

3.4 Next WG 2 meeting

After review and feedback, a revised document can then be submitted to the next WG2 meeting for voting. That meeting will take place in Singapure in 1997.