Document Type	Working Group Document
Title	Proposed additions to WG2 principles and procedures and Proposal Summary Form (Replaces L2/10-397 / SC2/WG2/N3944 of 2010-10-07)
Source:	V.S. Umamaheswaran, IBM Canada
Status	Expert contribution
Action	For consideration by WG2
Distribution	UTC

This document proposes four additions to WG 2's P&P document (N3902) and the Proposal Summary Form.

A. Contiguous encoding of decimal digits (to be added as new section 13 in P&P)

When script-specific decimal digits are encoded in UCS, the decimal digits will be encoded contiguously and in order, with room left for missing digits so that, if digits are later used as part of a place-value notation (i.e. a decimal radix notation) they can be used in that manner. Exceptions may be made only where (like numeric ideographs) the digits also serve as letters, or otherwise their use in decimal-radix notation can be safely excluded.

B. Information about use of standardized characters as part of a script's repertoire.
Add the following text as item 12 in "Submitter's responsibilities" attached to the proposal summary form:

> If the proposal is for a new script, identify all the standardized characters that are used directly in the script, or proposed to be unified with the characters of the script, in particular standardized characters allocated in different blocks. Examples include punctuation marks and combining marks. Such information will assist in assigning properties for characters shared across multiple scripts or in identifying character repertoires needed to support particular languages.

Note: This is not proposed to be added as a separate question in the form itself.

C. Information about confusable characters

i. Insert 'or could be confused with' in Q 10 in Section C – Technical Justification, to read:

> 10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to, or could be confused with, an existing character?

and

ii. Add the following text as item 12 in "Submitter's responsibilities" attached to the proposal summary form:

> If you are aware of already standardized characters that are visually close to any of the proposed characters, you are invited to list them in the proposal. This will assist in the analysis of the script for 'visually confusables', towards providing additional guidance on use of the standard from a security perspective (see *UTR#36 - Unicode Security Considerations* – at http://www.unicode.org/reports/tr36/).

Rationale:

UCS contains such a large number of characters and incorporates the varied writing systems of the world. Incorrect usage of characters can expose programs or systems to possible security attacks. In order to address this problem the Unicode Consortium has prepared two technical reports:

> - UTR #36: Unicode Security Considerations (http://www.unicode.org/reports/tr36/) describes some of the security considerations that should be take into account, and provides specific recommendations to reduce the risk of problems.
> - UTS #39 - Unicode Security Mechanisms (http://www.unicode.org/reports/tr39/ specifies mechanisms that can be used in detecting possible security problems.

Visually confusable characters lead to non-uniqueness in identifier strings and leads to problems such as spoofing on the internet – see section 2 on Visual Security Issues in UTR #36. UTS#39 describes the different kinds of confusables. It also includes data files on characters that can be confused with each other that was prepared by examining UCS for visually confusable characters within and across scripts.

As new proposals come on board, it would be useful to have some information about confusability of newly proposed characters with standardized ones. The proposers already have to ensure that a proposed character does not already exist, or can be unified with already encoded characters, while answering the set of questions that are in the current proposed summary form.

As an extension of that exercise, the proposers are invited to add any information they can add in the proposal, towards analyzing the script to enhance the data tables for visually confusable characters. The modification of Q. 10 and the proposed text in the submitter's responsibilities are to facilitate gathering such information. The information could be as simple as "similar to characters in script xxx or block yyy'. It could also be a listing of one or more encoded characters with which a proposed character could be confused with. This information would be optional and the level of detail is at the discretion of the proposer.

D. Stroke Counts for Ideographs

Add the following new question between current questions 12 and 13 in section C of the proposal summary form:

1x. Does the proposal contain any Han Ideographs? _____
If YES, is the total number of strokes (including the radical) for each glyph to be shown in the UCS charts specified? _____

Rationale:

Over the course of checking data in the Unihan database, a number of errors in the *kTotalStrokes* field were identified. This field contains the total number of strokes in the character (including the radical). Its value is for the character as drawn in the UCS charts. The total number of strokes is frequently used for sorting ideographs, and so getting these values correct is important. In order to expedite keeping this data current, the UTC has requested that the IRG have total stroke counts included in future submissions. Since some of the ideograph submissions are dealt with directly in WG2, a similar request is being made here to submitters to WG2 also.