Universal Multiple-Octet Coded Character Set UCS	
	ISO/IEC JTC1/SC2/WG2 IRG N 1327
	Date: 2007-06-05
Source:	Lu Qin, IRG Rapporteur
Title:	Proposal to Code Proper Names in a Separate Area
Status :	
Actions required	For discussion at IRG Meeting #28
Distribution:	
Medium :	Electronic

Introduction

The IRG' work on Unified CJK ideograph work has proceeded for over 17 years. In proceeding with the Unified CJK ideographs, the analysis is heavily based on source standards, the origin and meaning(cognate vs non-cognate) as well as glyph shapes with small differences in one to two strokes. Unification made it possible to eliminate duplicate characters as well character variants are essentially the same with slight differences due to difference in glyph preference in different geographic locations which can be made an analogy of "labor" vs. "labour" used in U.S. and the U.K.[1].

As the submissions from member bodies have changed from coded character sets to different sources with historic values, we have realized the importance of extending the unification examples and even reconfirming certain commonly practiced rules which are not written in the current Annex S. This revision is to reinforce the unification rules and made them more explicit in the unification process.

However, the unification of CJK ideograph characters poses a serious issue to a large subset of character submissions that are characterized as personal names[2]. These personal names often have glyph shapes that are slightly different from

---

[1] It should be pointed out, however that unification is not intended to eliminate variants in general as variants are an essential characteristics in Chinese writing system similar to the difference in perhaps "toilet" to "luau" in English.

[2] Place names and organization names have similar properties, these as a collection is normally called Proper Names. Here in this document, the personal name is used as a representative term only.

some coded characters, and thus falls into the unfiable characters and cannot be coded in the Unified CJK blocks. These characters are mostly collected by member bodies through authoritative bodies on personal name registration and there are similar on going projects by member bodies even if some members have yet to submit them for consideration. It is important that we realize the nature of personal names and the reasons that they are not unified.

Characteristics of Personal names
1. Uniqueness
characters used for personal names are by nature unique. Just like in English, some people prefer to call themselves, "Alan", others prefer to use "Allan', "Allaan", "Alaan", etc.. ideograph users also wants to put some unique characteristics in their names. Due to the writing restrictions of ideographs, people can only add/remove strokes from some existing characters, or constructing a characters with components having certain good meanings to individuals. Evan though this way of creating new characters should not be encouraged, people whose names have already accepted for use by authorities, cannot be refused to use their unique characters and computers should not bar them from using their existing names.

2. Arbitrary Selection of Strokes and variations
Again, because these personal names are constructed either using existing characters or some combination of existing components, some selection of strokes and components can be quite arbitrary and sometimes seems to be "wrong" characters to others. Yet, in the ideograph writing, new characters are created for one reason or the other, and we are not the body to judge right and wrong.

2. No direct Link to some Existing characters
Even though many personal names are very close in shape to certain coded characters and some times if pronunciations, yet, there is no reliable sources to prove its relation to the coded characters. It should also be pointed out that some personal names may have similar shapes to a number of coded characters, thus making the link to a particular character arbitrary because of the lack of solid information on the link to any particular coded character.

it is difficult and sometimes impossible to trace their origin. Even if pronunciation is provided, pronunciation alone does not provide enough clue to determine the "intended" characters. Also, member bodies are from different countries and regions where ideographs are used, yet the same character are

pronounced differently in different languages/dialects.

Thus it is difficult to use IVD to represent these characters.

Proposed Solution/Approach
Due to the unique nature of proper names and their problem by using IVD, the need to support this class of characters for electronic processing personal names, this document propose to request for a separate block to encode existing proper names including personal names and place names, organization names.

1. Separate Unification rules
This block of characters will not be subjected to the Unification rules in Annex S. Yet, it does not mean any character can be included in this block. A separate set of Unification/consolidation rules must be developed for this block.

2. Characters which has solid origins should be excluded from this block.