ISO/IEC JTC1/SC2/WG2 IRG N2274

Date: 2017-10-19

Source:	Lu Qin with feedback by IRG experts in IRG#49
Meeting:	IRG#49, San Jose, USA
Title:	Proposal to encode derived simplified Chinese schematically
Status:	Individual
Actions required:	To be considered by IRG
Distribution:	IRG
Medium:	Electronic
Pages:	11
Appendixes:	4

1. Background

The simplification of Chinese characters has been an ongoing process in different places where ideographs are used. However, the process is officiated in China through a formal method with specific reference to the document <简体字总表>（Simplified Characters List）(referred here as the document). The characters listed in that table are separately coded from their traditional forms in the CJK. However, the document also gives a set of rules either explicitly or implicitly which can be used to produce derived simplified characters(DSC). Even though the Chinese delegate to IRG has expressed a number of times that China has no intention to push for the use of DSCs, characters that fit the DSC definition are proposed to IRG for encoding with evidence of actual use, which according to previously established IRG rules should be accepted. Many IRG members are concerned about the amount of the potential size of DSCs to be separately coded using the current encoding method. Since DSCs do not provide additional lexical information, it may be more appropriate to use an encoding method that can establish their relationships with the corresponding traditional form characters.

New technology developed in recent years, more specifically ideograph variation sequence with the support of ideograph variation selectors (IVS), are used to support unifiable characters. However, IRG has a common understanding that simplified characters and their traditional counter parts are generally not considered unifiable and thus the derived simplified form cannot use IVS currently. A number of IRG members considers it important to develop a systematic encoding scheme to establish the link between a DSC to its traditional form character (TFC). Firstly, the scheme is aimed at simplifying the encoding process. Secondly, the establishment of the link provides additional information to better facilitate searching and indexing of related characters.

2. Proposed Solution

2.1. The principle

The proposed solution is to select a designated IVS to be used as the Derived Simplified Character Designator(DSCD). The encoding of a derived simplified character will use its corresponding traditional form character TFC followed by the DSCD as a sequence <TFC><DSCD>. There are a number of rules when using DSCD:

1. <TFC><DSCD> represents the fully simplified form by taking all the applicable rules. A character which does not apply all the simplification rules in the List cannot use this this scheme.
2. If the corresponding TFC is not coded, the TFC needs to be coded first before the derived simplified character can be coded.

2.2. The handling body of Encoding

The scheme is not intended to be administered by Unicode using IVD. In other works, the IRG should still be the reviewing body for approval of the DSCs and the approved characters are still a part of the CJK repertoire. Thus, proposed characters for encoding under this scheme should still demonstrate its actual use. In other words, this proposal suggests that IRG adheres to its rule that no DSC will be encoded if actual use evidence cannot be established.

2.3 Suggested code position of DSCD

This proposal suggest to use the last IVS in the collection of 240 reserved IVSs. That is, use the code point of U+E01EF as the designator. This suggestion is only based on the fact that the proposed scheme is extended from the basic idea from IVS. If there are other code position available in BMP, it would even be better as it is more convenient for use.