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.

Some Brief Comments on IRGN2274

(John Knightley 19 October 2017)

China is a multilingual country that currently has one fifth of the world's population, and has used CJK ideographs for thousands of years, hence even though other IRG members may have reached the end of, or be near to the end of CJK ideographs they wish to encode China has not. However whilst there are a large number of unencoded CJK ideographs in China, the vast majority of these are not derived simplified characters. The schema suggested would certainly not decease the workload of the IRG, the existing process is more than adequate for dealing with derived simplified characters, and reflects the wishes of the largest user community.

For many the biggest objection would be that this is a option that to be useful should have been implemented decades ago. If at that time implemented many other things would probably have been different. If the characters with Chinese simplified wind 风 in them were unified to traditional counterparts then characters with the Vietnamese simplified wind 𠃌几二, of which there are 17 characters in ws2017, would also be unified by the same mechanism. Thousands, maybe tens of thousands depending on your definition, of simplified characters have already been encoded and a precedent has been set. The chance to benefit from such a schema has been missed.

For over 20 years difference in abstract shape, be it difference in number of components, position of components or different of components has been the primary model for separate encoding of CJK ideographs. Abbreviated , or simplified, CJK ideographs and their traditional counterparts in general have a difference in abstract shape. Furthermore it is a difference that is often bigger to the traditional counterparts than many of its other variants. Consider the difference between 發 and 发 compared to that between 發 and either 発 or 𤼵 . To apply the schema as suggested, one that applies a different criteria to 《简化字总表》type derived simplified characters would unify characters like 啬 to their traditional counterpart by IVS but leave characters like 轱 separately encoded.

Simplified characters, characters designated as simplified in IRG character attributes, can contain certainly new lexical information, this is why there are simplified characters that exist with no corresponding traditional form. This is easily illustrated by extreme example of U+96D9 雙 vs U+53CC 双. In Unicode there are well over twice as many characters than contain the simplified form 双 as contain the traditional form 雙. This is not just a thing of the past, at least 12 characters containing 双 are in ws2107 submissions, but none containing 雙. Whilst for many simplified components there maybe fewer simplified characters than their traditional counterparts in real life there will always be some simplified character for which no traditional counter part exists. Only characters that actually exist should be encoded.

The reason there is not always a traditional counterpart is because simplified characters are not all by any means the product of the Chinese government and the《简化字总表》, nor for that matter regardless of components used not always derived from a traditional counterpart. Most if not all of the simplified components mentioned in 《简化字总表》were widely used centuries by the Han Chinese for their languages before it was published. Many of the simplified components have also been used by other peoples such as Japanese, Koreans, Vietnamese and Zhuang to form new simplified characters for their own languages. Simplified come from many sources not just one.

It is those simplified characters with proven existence that are submitted to IRG the number of which is considerably less than some might imagine, and they should be dealt with using the same unification criteria and encoding method as other characters submitted.