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 ⿰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 𫫇 to their traditional counterpart by IVS but leave characters like 𨍭 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.