Feedback on IRGN2582 IRG PnP Version 16 Draft

by John Knightley

for discussion by IRG.

(2023-03-15)

Introduction into PnP of the existing linguistic term “nonce”, which is used in among other places in tr45 (https://www.unicode.org/reports/tr45/) and in the Unicode standard section on CJK ideographs, would be helpful to identify a category of characters that is found in literature but in general not suitable for encoding and so should not be submitted to IRG unless there is evidence of wider usage. A nonce word is a word created specifically for use a literary work, or series of works, and not intended for general usage. The lack of intent for wider usage is such that these words may even be copyrighted or registered in some way. Such words are common in literature like science fiction as the name of some imaginary gadget or alien race. When new ideographs are created for such words they are “nonce characters” or “nonce ideographs”. Nonce words in any language are notoriously short lived and so are not usually included in ordinary dictionaries unless over many years they have gained wider use.

Earlier discussion on IRGN2551 and it’s predecessors centered around among other things the differences between characters in an ordinary Cantonese dictionary, submitted to IRG, written by a single author and characters in a book, never submitted to IRG, of articles from from a newspaper column where the author starts each article by making up a new word and the characters to go with it, and continues by writing about the word. The conclusion was that evidence from the Cantonese dictionary was acceptable as evidence to the IRG, but that evidence from the newspaper column book was not sufficient. The reason that the book of articles from a newspaper column was clearly not sufficient as evidence was expressed in many ways, but in short it is because they are nonce characters that have not gained wider use.

The two other cases discussed at length, of characters not submitted to date to IRG, are also in the widest sense nonce characters for nonce words. The first for a word made up for a series of shampoo commercials, and the second nonce characters used for the nonce words in a hundred year old rendering of Jabberwocky. The evidence for the former was considered insufficient but that for the latter sufficient. The Jabberwocky characters an exception for nonce characters in that they have lasted over a century, been the subject of academic papers and are included in school textbooks.

It should be noted that in most cases nonce characters become encoded not because they increase over the years, but because they are used elsewhere and such usage often predates the nonce usage.

Two changes are currently proposed for section 2.2.1 of PnP namely:-

d) Context (上下文信息): Sufficient context in text to decipher the semantic meaning of the character. IRG has the right to reject characters that do not have sufficient evidence for IRG to decipher its semantics.

e) Usage (需求限制): The use of characters must be for justifiable public interest. Examples of public use include evidence of: governmental needs; scientific use; digitization projects for public use; and working systems of significance as accepted by IRG. IRG has the right to reject characters that are created by individuals mostly for personal use with little public use value.

The former change is clear and concise so no change is suggested. It is suggest the the second change follow the same format as the first and a statement about not submitting nonce characters without evidence be added later in the section below that:-

e) Usage (需求限制): The use of characters must be for justifiable public interest. Examples of public use include evidence of: governmental needs; scientific use; digitization projects for public use; and working systems of significance as accepted by IRG. IRG has the right to reject characters that do not have sufficient evidence for IRG of justifiable public interest**.**

The section continues:-

> Submitters must make sure that the ideographs they submit do not fall into any of the following categories:
> a) Ideographs already standardized in the ISO/IEC 10646 standard (including its amendments).
> b) Ideographs currently in WG2's working drafts.
> c) Ideographs currently in IRG working sets including both M-set and D-set[1].
> d) Ideographs mis-unified or over-unified with ideographs in the current standard based on the lists maintained by IRG in its working document series, namely IWDS_MUI and IWDS_NUC.
> e) Ideographs from ancient documents that are rare and not in general use, along with variants from tombstone carvings that are not in circulation nor used in printed form, should have an appropriate base character

[1] See Section 2.3.4 for the purposes of M-set and D-set.

identified through the use of authoritative dictionaries and other references, then be submitted as IVSes to be registered in a new or existing IVD collection. See Section 2.2.1g.

f) Nonce characters are not in general considered suitable for submission to IRG and evidence from the original publications alone of such characters is insufficient. Nonce characters should only be submitted if there is also evidence of significant wider usage.

If required a definition of nonce characters can also be provided for the glossary of PnP.