JTC1/SC2/ISO/IEC 10646/IRG N2551

Date:2022-07-28

ISO/IEC 10646/JTC 1/SC 2/WG 2/IRG

Ideographic Research Group (IRG)

Title: **Rules to Restrict the Encoding of Modern "Self-created Characters"**

Author: Lu Qin (IRG Convenor)

Source : IRGN2521&Feedback

Date : 2022-07-28

Distribution: IRG Experts for discussion at IRG #59

1. Background:

In historic perspective, characters are always created by individuals. Thus, in general, almost all characters are self-created. Some of these self-created characters are eventually accepted and used commonly by at least a group of people in the public. That is why we often hear the terms: commonly used characters, general purpose characters, which are mostly coded already. IRG also code characters evidenced through historic documents in some permanent forms, such as reputable dictionaries, reputable collection of books, and other printed materials of common interest. In the computer age, encoded characters with font support can be used to produce printed documents to facilitate information exchange.

Computer systems also provide technology to allow individuals to create characters through private-use areas using self-defined glyphs. Consequently, some of these characters get to be introduced to the pubic in printed forms. Currently, IRG encoding principles requests evidence to show character use for encoding. Because of the ease to create characters on the go, IRG experts raised concerns on the encoding of self-defined characters which many not be of much public interest. Thus, there is a need to restrict the encoding of self-created characters that are mainly for individual use. This can ensure that CJK unified characters are kept at a manageable size. This can also avoid introducing noisy characters that should only remain private even if there are printed evidence of these characters. This document attends to summarize previous discussions to come up with definitions and restrictions.

2. Definition:

In general, ideograph characters do get created from time to time even in the modern age and those that have common interest for use should be encoded for public use. For encoding purpose, IRG needs to make a decision on whether to encode self-created characters with printed evidences. IRG only considers to encode self-create characters that have public interest, whereas others should be considered private and thus not suited for encoding. To avoid confusion, IRG refers to these newly created characters in computer systems that are not suited for encoding as **private characters**.

3. Factors to determine self-created characters for encoding:

IRG considers that the number of modern self-created characters need to be encoded small. Thus, the suitability of encoding should be determined on a case-by-case basis by IRG. To assist in determining whether self-created characters are appropriate for encoding or should remain private, the following guidelines are used by IRG.

3.1 Factors to argue for encoding:

- Created to fulfil a specific need (e.g. representation of spoken morphemes in dialect usage; transcription of Old Hanzi character forms; naming of chemical elements, animal species , or plant species; or any other use with certain degree of common interest either for the public or a group.) ;
- Created by an acknowledged expert in a specific field for which the characters are intended for use by at least a group of people;
- Published in print by a reputable publishing house with vigorous review editing process and the publishing house has reasonable size of readers;
- Occur as part of a larger corpus of related characters (e.g. in a dictionary of words or characters used in a particular dialect) ;
- Adopted in print by other users at least in a speciality group;
- Required for use in government databases.

3.2 Factors which argue against encoding:

- Ephemeral (e.g. characters created for special events, competitions, or puzzles) and it is generally treated as symbols;
- Personal use (e.g. personal name characters created to be different) or logos for personal use or identification;
- Only used on the creator's web site or in a self-published book
- Created for fun or amusement only
- Graphic variants or alternative forms for existing encoded character

The list of factors listed above are not meant to be exhaustive. Submitters can argue by other factors not listed here as long as they are accepted by IRG experts and the list can also be updated based on discussion result.
Acceptance of characters for encoding is an IRG decision determined by IRG expert review. The above list are consideration factors, not rules. The more supportive factors there are, the better chance for encoding is a submitted character. If there are good reasons not to encode a character even if it fit several supporting factors, the decision lies with IRG.

Note that a character deemed private by IRG does not prevent it from being encoded in the future if new evidences show that they are also used by other people with reasonable common interest or use.

4. Example Cases for Encoding Consideration:

Below gives the rationales for the consideration of four character submission cases with suggested outcomes to demonstrate how these factors should be applied in practive.

4.1 Case Study 1

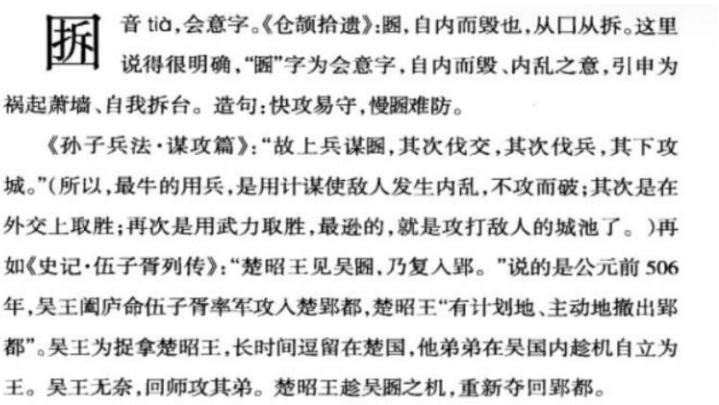
The character 龘 / 𪚥 anonymously created to represent the neologism *duāng* derived from a 2004 advertisement for shampoo spoken by Jackie Chan (stage name 成龍).

- Ephemeral (widely used internet meme from 2015, but no indication that the character has long-term significance)
- Created for fun only
- Few or no attestations in printed usage

Decision: Not suitable for encoding at the present time.

4.2 Case Study 2

Characters created by Yú Shǎoléi 余少镭 (see IRGN2482 pp. 2–3) such as:



囧 音 tiǎ，会意字。《仓颉拾遗》：囧，自内而毁也，从口从拆。这里说得很明确，“囧”字为会意字，自内而毁、内乱之意，引申为祸起萧墙、自我拆台。造句：快攻易守，慢囧难防。

《孙子兵法·谋攻篇》：“故上兵谋囧，其次伐交，其次伐兵，其下攻城。”（所以，最牛的用兵，是用计谋使敌人发生内乱，不攻而破；其次是在外交上取胜；再次是用武力取胜，最逊的，就是攻打敌人的城池了。）再如《史记·伍子胥列传》：“楚昭王见吴囧，乃复入郢。”说的是公元前 506 年，吴王阖庐命伍子胥率军攻入楚郢都，楚昭王“有计划地、主动地撤出郢都”。吴王为捉拿楚昭王，长时间逗留在楚国，他弟弟在吴国内趁机自立为王。吴王无奈，回师攻其弟。楚昭王趁吴囧之机，重新夺回郢都。

- Created for self-serving purpose (do not fulfil any specific need)
- Not adopted by other users or in other printed publications

Decision: Considered it private, not suitable for encoding now.

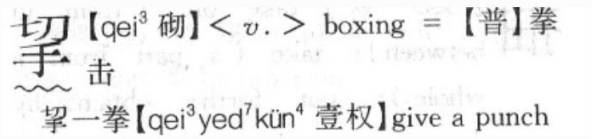
4.3 Case Study 3

Cantonese-usage characters in the *Concise Cantonese-English Dictionary* compiled by Yáng Míngxīn 杨明新.

- Created for the specific purpose of representing Cantonese morphemes for which no written character previously existed
- Published by a reputable academic publishing house (广东高等教育出版社)
- Characters adopted (and corresponding traditional forms created) by Cheung Kwan-hin and Robert S. Bauer in “The Representation of Cantonese with Chinese Characters” (*Journal of Chinese Linguistics*, Monograph Series Number 18, 2002)
- These characters are commonly understood by Cantonese speakers if used in running text. The following character is a very typical example of a character with specific meaning, spoken and understood by most Cantonese speakers.

Source 1: 杨明新：《简明粵英词典》（广东高等教育出版社, 1999年） p. 353
Note: Cantonese: qei³

Evidence 1



掔【qei³ 砌】< *v.* > boxing = 【普】拳击
掔一拳【qei³ yed⁷ kün⁴ 壹权】give a punch

The Representation of Cantonese with Chinese Characters, Journal of Chinese Linguistics, Monograph Series Number 18, 2002: p406, pos09

Evidence 1

Decision: Suitable for encoding.

4.4 Case Study 4

Characters created by Yuen Ren Chao (趙元任) for the translation of Lewis Carroll's poem *Jabberwocky*,

- Created for the specific purpose of representing nonsense words created by Lewis Carroll (e.g. 炸脖𪚥 *zhábówò* for *Jabberwock*; 㶽裡 *béilǐ* for *brillig*; 鵓𪇰鴝子 *bólōgōuzi* for *borogoves*; 𡚒妙 *biǎmiào* for *beamish*)
- Chao was a renowned linguist and translator
- Quoted in other works, e.g. in the 1996 translation of Douglas Hofstadter's *Gödel, Escher, Bach: an Eternal Golden Braid* (哥德尔・艾舍尔・巴赫：集异璧之大成)

Decision: Suitable for encoding.