JTC1/SC2/WG2/IRG N2533
2022-03-03

Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation Internationale de Normalisation
Международная организация по стандартизации

Doc Type: **ISO/IEC JTC1/SC2/WG2/IRG**
Title: **UK Activity Report for IRG #58**
Source: **Andrew West**
Status: **Member Body Contribution**
Action: **For consideration by IRG**
Date: **2022-03-03**

1. Review of WS2021

UK experts provided comments and feedback on WS2021 using the online review tool. We note the following issues which we encountered while doing our review:

A. Clipped evidence images

We note that the evidence images for many characters submitted by China have been clipped to only show the proposed character and a few immediately surrounding characters. This makes it very hard to understand the meaning and usage of the proposed character, and to verify whether the glyph shown is correct or whether it may be an error form for an existing character. We therefore request that in future IRG submitters provide images of the complete evidence (complete entry or full page). We additionally ask China to provide more complete evidence images for WS2021 characters where the current images have been clipped.

B. Incomplete references for evidence images

We note that some references for evidence images provided by China are incomplete, and are not sufficient to easily locate and verify the source text. In particular, some of the references given by China are to large collections of books such as 粵雅堂叢書 (185 books), 平津館叢書 (43 books), and 鄦齋叢書 (21 books), but the actual book title and volume/page references for the evidence are not provided. We request that in future IRG submitters provide full references for all source evidence. We additionally ask China to provide complete references for WS2021 characters where the current references are incomplete or ambiguous.

C. Error forms in modern sources

We note that some of the modern sources used as evidence for characters submitted by China give error forms for characters which are already encoded, and the evidence text is often a misquotation from a well-known classical text such as 爾雅 *Ěryā* or 山海經 *Shānhǎijīng*. We consider that error forms given in modern sources are not normally appropriate for encoding.

D. Personal name characters submitted by TCA

We consider that there are several problems with the personal name characters submitted by TCA.

Firstly, there is no corroborating evidence for the characters to indicate that the provided glyph form is correct. Without additional evidence it is very difficult to determine whether a proposed character may be a variant of an encoded character or may even be an error form of an encoded character. It is very possible that in some cases a submitted personal name character could be a clerical error resulting from the misreading of the handwritten form of a character.

Secondly, we believe that TCA should normalize glyph forms to be consistent with TCA conventions (e.g. use the standard form of the grass radical ⺿ instead of 艹). We very much doubt that there really is a technical reason to encode a non-normalized glyph form for use on an ID card, as there must be many citizens who prefer to write their name with a non-standard glyph form of an encoded character (e.g. they might write 芳 with 艹 instead of ⺿), but either this difference is not reflected on their ID card or the glyph difference is dealt with at the font level. If a citizen who prefers to write their name as 艹方 can have 芳 in its place on their ID card, then a citizen who prefers to write their name as 艹閣 can certainly have ⺿閣 in its place on their ID card. If TCA really believes that it is necessary to use a non-normalized glyph form on a person's ID card then the normalized form should be shown in the code charts, and the non-normalized form should be registered as an IVS.

Thirdly, we are not convinced that it is appropriate to encode characters that are only used on ID cards of individual citizens and in associated government databases. Such usage is inherently private use, and encoding characters which are not attested in published documents seems to go against the principles of character encoding. Furthermore, once the citizen with a unique name character dies, there is no longer any need for the encoded character (the ID card is destroyed, and the citizen's name is removed from government databases). Is it appropriate to burden the standard and font developers with hundreds or thousands of ephemeral personal-use characters?

2. On the encoding of modern self-created characters

In response to Recommendation IRG M57.5, we submitted the document IRGN2521 "On the encoding of modern self-created characters", which discusses the factors for and against encoding modern self-created characters, and provides several case studies.

We note that in the review comments for WS2021 v. 2.0, no experts representing IRG member bodies have raised any objections to the use of *Jiǎnmíng Yuè-Yīng Cídiǎn* 简明粵英词典 [A Concise Cantonese-English Dictionary] (Guangzhou: Guangdong Higher Education Publishing House, 1999) as evidence for characters submitted by UK. Therefore we consider that there is no need for any further discussion by IRG of the appropriateness of encoding characters from this source.

3. On the encoding of Daoist-usage characters

In response to Recommendation IRG M57.9, we submitted the document IRGN2522 “On the encoding of Daoist-usage characters”, which discusses the appropriateness of encoding Daoist-usage characters. The document concludes that the Daoist-usage characters proposed by UK are not intrinsically different from any other CJK unified ideograph, and should be encoded on the basis of the evidence of usage we have supplied.

We note that in the review comments for WS2021 v. 2.0, no experts representing IRG member bodies have raised any objections to the encoding of Daoist-usage characters submitted by UK, or questioned the validity of the evidence provided. Therefore we consider that there is no need for any further discussion by IRG of the appropriateness of encoding Daoist-usage characters as CJK unified ideographs.

4. Request to move the source reference for UK-02830

We submitted document IRGN2520 “Request to move the source reference for UK-02830”. This document requests that the source reference UK-02830 be removed from U+238A7 𣪧, and added to U+4DBE 敗.

See IRGN2534 2.b for the UTC resolution for this issue.

5. Request to correct the radical and residual stroke count for UK-10989

We made an urgent request to the UTC to correct the radical and residual stroke count for UK-10989 in CJK Ext. H, from 40.1 to 25.7.

See IRGN2534 2.a for the UTC resolution for this issue.

Universal Multiple‐Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2/IRG N2533
Date: 2022-3-17

Source:	China
Author:	TAO Yang
Title:	IRGN2533 Feedback by China
Meeting:	IRG #58
Status:	Member's submission
Actions required:	To be considered by IRG
Distribution:	IRG
Medium:	Electronic
Page:	2
Appendix:	0

I noticed that Andrew made three comments on the data submitted by China, which are valuable opinions for improving the data quality.
Taking into account the balance between the quality of the proposal and the actual conditions,

A. Clipped evidence images

Previously, China has always provided complete pages, but the recognition effect is not clear due to the large format and small font size. IRG suggests that only the parts around the glyphs that can explain the meaning of the text should be intercepted. Therefore, China has only provided the cut evidence images since ws2017. If necessary, China will provide a full page for IRG members to read. On the other hand, considering that many websites will directly obtain the proposal data submitted to the IRG website, China does not want to show all complete pages to provide these websites with opportunities to get something for nothing.

B. Incomplete references for evidence images
Adding complete bibliographic data for font attributes is a good suggestion.
The reference data China is automatically generated from the database. The original data only describes the book name according to the highest level directory, so the reference data can only automatically obtain the name of the series.
Strictly speaking, the use of series names is not a mistake, but visually imperfect. At present, the method of retrieving other versions mainly depends on context retrieval, so the existence or absence of accurate book titles does not affect the collection of other versions.
In addition, because I am the only person in China who is engaged in the collection, sorting, textual research and preparation of proposals of thousands of words, I need to use my spare time to complete all the work, and there is no corresponding financial support, so the cost of manually labeling data and information is too high. If time permits, I will try to complete the relevant data. However, the proposal submitted in its current form does not violate the requirements of the IRG proposal.

C. Error forms in modern sources

Two possibilities should be considered for the so-called text errors in the modern sources. One is that the collation of the book selects a base copy different from the reviewer's, and the other is that the editors artificially make new errors in the collation process.

According to the current review results of modern source ancient books, the above two situations both exist.

First of all, we should not rashly assume that only one version of the text is valuable because there are differences in glyphs between the two versions. Secondly, I am really shocked by the editing quality of some national ancient books sorting projects.

In the future, if the resources submitted are ancient books sorted out in modern times, it is necessary to find as many original ancient books as possible and correct the text.