

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.