

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N 1199
Date: 2006-05-31

Source:	China
Title:	Chinese Preliminary Proposal for CJK_C2
Status :	
Actions required	To be reviewed by IRG members
Distribution:	IRG#26, Hue, Vietnam, 2006-06
Medium :	Electronic

This document collects 4,952 ideographs encoded in neither ISO/IEC 10646: 2003 nor the CJK_C1. All the characters will be proposed to the future CJK extension project

This document is not a formal proposal, the total number of characters and characters' attributes and sources may be changed in the formal proposal. This document is for IRG preview only.

1. In 2004, the National Library of China (中國國家圖書館) started the “digitalization project of local history books”, its stage 1 (500,000 pages) and stage 2 (400,000 pages) were completed in 2006. The mentioned 4,952 ideographs are all used for the project.

Statistics of stage 1 show: In the used 204,808,490 hanzi, 203,781,248 are covered by the CJK Unified Ideograph, 274,847 are covered by CJK_A and 732,675 are covered by CJK_B. There are 19,720 hanzi not covered by the current standard, thus, 4,866 user defined characters are needed. Besides, there 19,000 hanzi are unable to be recognized.

Statistics of stage 2 show: In the used 161,222,531 hanzi, 160,135,692 are covered by the CJK Unified Ideograph, 279,401 are covered by CJK_A and 759,774 are covered by CJK_B. There are 47,622 hanzi not covered by the current standard, thus, 3,009 user defined characters are needed.

According to the comparison between the CJK_C1_v41 and the user

defined characters for stage 1 and 2, there are 4,952 characters selected.

2. The ongoing “press and publication used Chinese character set project” of The General Administration of the Press and Publication selected about 15,000 Chinese characters. The characters and their information are being checked and may be proposed or provided to IRG later.

The ongoing project specified and selected a batch of Chinese characters used for about 500 Chinese ancient classical books and their annotation books. There are about 5,000 characters are not encoded in the current international standard. Their glyphs are special and must be kept in press and digitalization.

The project also selected more than 10,000 Chinese characters used in about 50 ancient dictionaries. Their glyphs are not intended to be unified to the encoded CJK ideographs for the reason of their special uses of explaining script meanings. The project is not finished.

Attachments:

1. 4,952 glyphs and their attributes such as Kangxi index, etc.
2. sources of 4,952 characters