

Universal Multiple-Octet Coded Character Set UCS

ISO/IEC JTC1/SC2/WG2 IRG N 998

Date: 2003-11-14

Doc. Type:	Member body contribution
Title:	Activity Report form TCA
Source:	TCA
Status:	Input to IRG
Action:	For Your Information
Distribution:	IRG Members and Ideographic Experts
Reference:	
No. of pages:	2
Medium:	Electronic

1. The extension of CSIC (CNS 11643):

The extension of CSIC (Chinese Standard Interchange Code, CNS 11643) is in progression. The last version of CSIC was published on 1992, more than 10 years ago. During the decade, there were ten-thousands new Chinese characters found and there are out of CSIC. BSMI (Bureau of Standard, Metrology and Inspection), the organization for standardization in Taiwan, therefore offered a contract to CMEX (Chinese Foundation for Digitization Technology) for extending CSIC. The new version of CSIC will be published on 2004.

CSIC-1992 has 16 planes (with 8,836 code points per plane) for encoding characters. But there are actually 7 planes used for encoding Chinese characters and symbols. In contrast, CSIC-2004 has 80 planes for encoding characters. In current stage, we encoded Chinese characters and symbols in 15 planes. The additional 8 planes are listed below:

Planes 8 & 9: are used to encode the alphabets and symbols from BMP with there are out of CSIC-1992;

Planes 10 & 11: are used to encode the ideographs from CJK extension B with there are out of CSIC-1992;

Planes 12 - 15: are used to encode the new Chinese characters from the household registration system and the ideographs from CJK extension A with there are out of CSIC-1992.

2. The issue of IBS (International Basic Subset):

The IBS for Taiwan (named IBS-TW) was already done and submitted to IRG. There are 6,608 Chinese characters in the IBS-TW. These Chinese characters came from the 1st plane of CSIC, text books for primary schools and junior high schools, the names of places and very frequently used Chinese characters in newspapers.

Besides, we compared the IBS from Taiwan, China, Japan and DPR Korea. We found there are 10,576 characters in the union of the 4 IBS and 2,285 characters in the intersection of the 4 IBS.

3. The audit for new Chinese characters submission:

In order to resolve the problem of interchanging new Chinese characters (which are out of CSIC) for computer processing among government agencies, a registration and assessment system for new Chinese characters was established by our government. If anyone found a new Chinese character, he can register the character by filling an application form on the website:

<http://www.cns11643.gov.tw/web/newword.jsp>.

Each submitted character should be first checked from CSIC database to make sure that it's actually a new character. Then the new character should be assessed by an authority department depending upon the usage of the new character. After the new character is conformed, it'll be encoded into CSIC with a propoated CSIC code.

4. Renew the CSIC/UCS mapping table:

According to the extension of CSIC, we also need to renew the mapping table among CSIC, UCS and Big5. This work will be done on 2004.

In the renewing work, we found some errors of mapping CSIC codes to UCS codes. For instance, we found that about one thousand CSIC-1992 characters have no mapping UCS codes. We now try to find why and hope most of them are simply "T-source missing" in UCS.