Universal Multiple-Octet Coded Character Set
UCS

ISO/IEC JTC1/SC2/WG2 IRG N999

Date: 2003-11-16

Doc. Type:	Member body contribution
Title:	China's Old Hanzi Encoding Proposal
Source:	China
Status:	Input to IRG
Action:	
Distribution:	IRG Members and Ideographic Experts
Reference:	
No. of pages:	7
Medium:	Electronic

I. Content

The proposal mainly suggests the Old Hanzi Encoding into the ISO/IEC 10646 with independent codes.

The Old Hanzi refer to historic characters before Qin Dynasty excavated or edited, stored by ancient people without repeated transmitting, including Oracle-bone inscriptions, inscriptions on bronze object, the script of the Warring States and small seal. These characters of various fonts are taken from genuine ancient scripts, not those characters losing original shapes after repeated transmitting. After sorting out historical characters of various fonts, optimizing the shapes, they occupy zone bits in international standard character set common in the world, so that people all over the world may use these Old Hanzis for text publication and historical studies.

II. Importance

Part of Old Hanzi, e.g. Oracle-bone inscriptions, become world cultural relics. Others are also concerned by many countries. Chinese characters belong to ideographs, reasons for the shapes exist in large amount of Old Hanzis; so not only Chinese character education needs Old Hanzis for the explanation of contemporary characters, many bilingual education with Chinese as target language also needs Old Hanzi shapes to explain characters. Old Hanzis have become common optional fonts in publications in China, the construction of Old Hanzi font and feature database is needed by Chinese research experts all over the world. However, due to the lack of general coding in international standard character set, the network of Old Hanzi cannot be realized. The

success in developing Old Hanzi publications, teaching multimedia software, large Old Hanzi font and databases is great, but the lack of international standard codes, these products needed by everyone cannot be shared internationally. Thus, original Old Hanzi shapes' into ISO/IEC 10646 should be raised on the agenda quickly.

III. Feasibility

Inputting Old Hanzi into computers is important event for Chinese information processing at the end of 20th century with many pioneer fruits in this area. Since 1990s, some people tried to build a "Chinese character set", but due to shortage of strength and investment, it was not fully realized, but much experience is accumulated and some local fruits were achieved. In recent years, the fonts for Oracle-bone inscriptions, inscriptions on bronze object and small seal were developed successfully, the script of the Warring States inscriptions and bamboo inscription fonts are being constructed. The solution of character shapes serves as a font for constructing an international standard Old Hanzi.

Inter-character sorting is another key for international standard Old Hanzi. Now, the systematic theory of Chinese characters support the solution of this problem, among the character fonts built, most inter-character relation was sorted out, providing a necessary condition for the setting of the zone bits.

The satisfaction of the 2 conditions above makes this work feasible.

IV. Plan

1. Based on the present conditions, the 4 kinds of Old Hanzi are suggested:

a. Oracle-bones are more than 4000, and 1300 characters are understood.

b. Inscriptions on bronze object are over 4900.

c. The script of the Warring States are over 2300.

d. Small seals are 10560.

These 4 kinds of Old Hanzi have mature assortment, relatively more characters, therefore have conditions into ISO/IEC 10646, meanwhile they can serve as coordinates for Old Hanzi coming later.

2. Collecting and Positioning of Character Sets:

Due to the irregularities in various Old Hanzi in historical development and correspondence, what's more, the functions of Old Hanzi are not equivalent of another or other non-mainstream fonts of Song characters, therefore, Old Hanzi of various types are independently positioned, parallel with Song level, providing an independent, fixed zone for various Old Hanzi.

Various types of character sets adopt a descriptive mode, i.e., treating character shape as a unit, all character samples after assortment, any variant character occupies a zone bit.

The shape order of various character sets is based on ShuoWenJieZi's radicals, characters in the radicals correspondent with Qin standard scripts in ShuoWenJieZi are listed according to ShuoWenJieZi, inconsistent ones are listed below correspondent characters according to the radicals in ShuoWenJieZi.

关于中国古汉字进入国际标准字符集的提案

一、内容

本提案的中心内容是建议把中国古汉字放入国际标准字符集，单独编码，占有全世界通用的区位。

中国古汉字，指的是地下出土或古人专门编排、贮存未经反复转写的秦代以前的历史文字，包括甲骨文、金文、战国文字和小篆的原形字。这些不同字体的汉字，是从古代实际文本上摘取下来的，而不是字书反复转抄后失去原形的。把这些不同字体的历史汉字经过整理，优选字形，在国际标准字符集里占有全世界通用的区位，便于全世界都能使用这些古汉字进行文本印刷和学术研究。

二、意义

汉字属于表意文字，它的形体构造的理据大量贮存在早期的古汉字中，因此，不仅中国的汉字基础教育需要借助古汉字来讲解现行汉字，很多以汉语为目标语的双语教学，也需要借助古汉字的字形来帮助汉字的讲解。与语言文字学有关的现代印刷出版物中经常使用古汉字字形。作为世界上自产生至今未曾发生断裂的唯一现存表意文字，现行汉字的早期形体——古汉字，既是中华文化的瑰宝，又是世界文化遗产的重要组成部分。古汉字的研究，不但深受中国学者的高度重视，同时也已成为世界各国汉学专家共同关注的焦点。然而，由于没有国际标准字符集的统一编码，无法实现古汉字的信息处理与交换。带有古汉字的印刷品、教学多媒体软件、大型古汉字字库与数据库已经有相当数量的产

品研制成功，但是由于缺乏国际通用编码的支持，这些大家共同关心和需要的成品，无法实现国际共享。为此，古汉字原形字进入国际标准字符集，应当尽快提到议事日程。

三、可行性

使古汉字进入计算机，是上世纪末汉字信息处理的一件大事，在这方面已经有了不少先期成果。从 90 年代起，就有人尝试创建“全汉字字库”，由于力量和投入不足，未能全部实现，但是也积累了很多经验，留下了一些局部成果。近年来，甲骨文字库、金文字库、小篆字库，都已经初步研制成功，战国文字字库和简帛文字字库正在创建中。字形问题的解决为国际标准古汉字字符集的创建打下了基础。

创建国际标准古汉字字符集的另一个需要关注的问题是古汉字字际关系的整理。目前，汉字系统性的理论给这个问题的解决提供了依据，在已经创建的字库中，大部分古汉字的字际关系已经经过整理，为古汉字字符集的编码方案提供了一种可行的方案。

以上两个条件的具备，使这一工作具有了可行性。

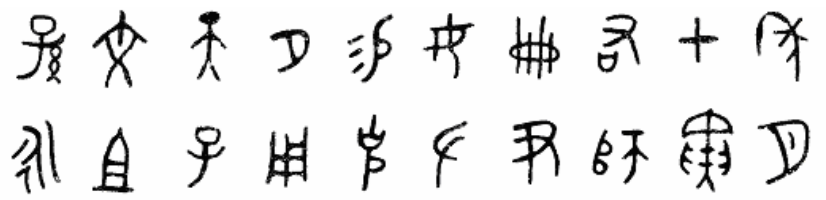
四、方案

一、第一批列入国际编码的古汉字的类别：

根据目前的条件，建议将以下四类古汉字首先列入国际编码：

1. 甲骨文字符集(已释字 1300 个左右，总字数 4000 多个)

2. 商周金文字符集（4900 多个）



3. 战国楚简文字字符集（2300 多个）

4. 《说文》小篆(包括收入《说文》的籀文和古文重文)字符集（10560 个）

这四类古汉字目前的整理比较成熟，字量相对较多，具备进入国际标准的条件，同时可以起到创建下一批古汉字字符集的坐标作用。

二、字符集的收字和放置：

鉴于各类古汉字的传承对应关系是不整齐的，加之古汉字字符集的功能并不相当于宋体字的另一种或几种非主用字体，因此，各类古汉字字符集全部独立放置，与宋体层面平行，也就是给各类古汉字单独开辟固定的区域。

各类字符集均采用描述型，即，以字样为单位，收录整理

后的全部字样，所有异体字均独立占有一个区位。

各字符集内字形的排序原则上依照《说文解字》部首，部首内的字与《说文解字》小篆能对应的，按《说文解字》排列，不能对应的，参考《说文解字》归部，置于能对应的字后。