建立统一的汉字字符集的提案

TITLE : PROPOSAL FOR CREATING UNIFIED CJK HAN CHARACTER COLLECTION (HCC)

TO : MEMBERS OF ISO/IEC JTC1/SC2/WG2 AND OTHERS WHOM IT MAY CONCERN

SOURCE: CCIPS
CHINA COMPUTER & INFORMATION PROCESSING STANDARDIZATION TECHNICAL COMMITTEE

FAX : (+861) 841-1363

DATE : 22 JULY, 1989

目　　录

1. 引言

2. 建立统一的汉字字符集(HCC)

2.1 定义(HCC)

2.2 字料

2.3 基本集
2.3.1 优先选入字
2.3.2 字量与分布
2.3.3 选字原则
2.3.4 分级
2.3.5 字序
2.3.6 字位的安排

2.4 辅助集

3. 实施办法

4. 结束语

附录A　几种可能采用的方案的评述

A.1 方案 I
A.2 方案 II
A.3 方案 III

附录B　HCC参考数据

1. 引言

中文信息处理的基本矛盾之一便是大字符集与小编码空间的矛盾。尽管 ISO/DP 10646 提供了广阔的四维空间，但是，在基本多文种平面这个“最惠平面”，为汉字编码所剩的空间却十分狭小有限。

为了公正、妥善、科学地解决 DP 10646 中汉字编码的问题，半年来，中国的有关标准化工作者、计算机专家、文字专家，曾反复地对多种解决方案进行了多方面的分析、比较。(参见附录A)

现向 ISO/IEC JTC1/SC2/WG2 提议建立统一的C/J/K 汉字字符集。该方案是在以下三个方案分析比较的基础上提出的：

- 方案Ⅰ，B.M.P.上 C/J/K 格局基本不变。但为 C/K 分配的两个区可能重新安排。
- 方案Ⅱ，C/J/K 退出B.M.P.建立各自的基本民族平面。
- 方案Ⅲ，建立广义汉字字符集。

从各方面的指标综合分析，唯有方案Ⅲ，才可以比较全面地满足整个汉字文化圈和世界规模的信息技术发展的需求，并为未来构筑发展的平台。

由于方案Ⅲ，广义汉字，并非只为中国人使用，也并非只含中文汉字，故改称“统一的汉字字符集”，简称汉字字符集 (HCC)。

2. 建立统一的C/J/K 汉字字符集

2.1 定义

Unified C/J/K Han Character Collection

统一的C/J/K 汉字字符集，简称HCC。

它是中文汉字、日文汉字、朝文汉字的超集。

在ISO/DP 10646 中，HCC 特指在中(含台湾)、日、南朝鲜三国现有信息交换用字符集的基础上，经过认同与甄别、删除与增补、重新排列而生成的新字符集。

2.2 字料(初始材料)

- GB2312-80
  GB7589-87　　及相应的繁体字集
  GB7590-87

  现代汉语通用字表 (1988 年 3 月，中国国家语委)
- JIS x 0208
- JIS Level Ⅲ
- KS C 5601-1989

并参考 CNS-11643（中国台湾标准）

2.3 HCC 基本集

2.3.1 优先选入字

- GB2312-80 6,763 及对应的繁体字
- JIS x 0208
- KS C 5601-1987 4,888(HANJA)
- CNS 116435,401

2.3.2 字量与分布

≈ 2.5 ZONES 约 2 万字，分布在

- 原 I00 下半部
- 原 I10
- 原 I11

2.3.3 选字原则

- 应收入C/J/K 各标准中所有的常用汉字
- 以形为主，兼顾意、源

● 字形完全相同的字，一律入选
● 字形结构完全不同而意义相同的字，一律入选，并给予不同编码
● 字源相同、 字义相同、 字形相近的字，按照一套C/J/K共同认定的规则，实施认同与甄别，有所取舍。

- 同形异音字只赋一个码值
- 严格控制异体字入选，但必须吸收那些实际上经常使用的异体字

2.3.4 分级

按照使用频率，兼顾构词能力、学科分布、口语应用， 将入选的汉字粗分为叁级。

I 级：常用汉字	约 7,500 字
II 级：次常用汉字	约 7,500 字
III 级：非常用汉字	约 5,000 字

应规定，对于 HCC ，其最小子集是 I 级 HCC。

2.3.5 字序

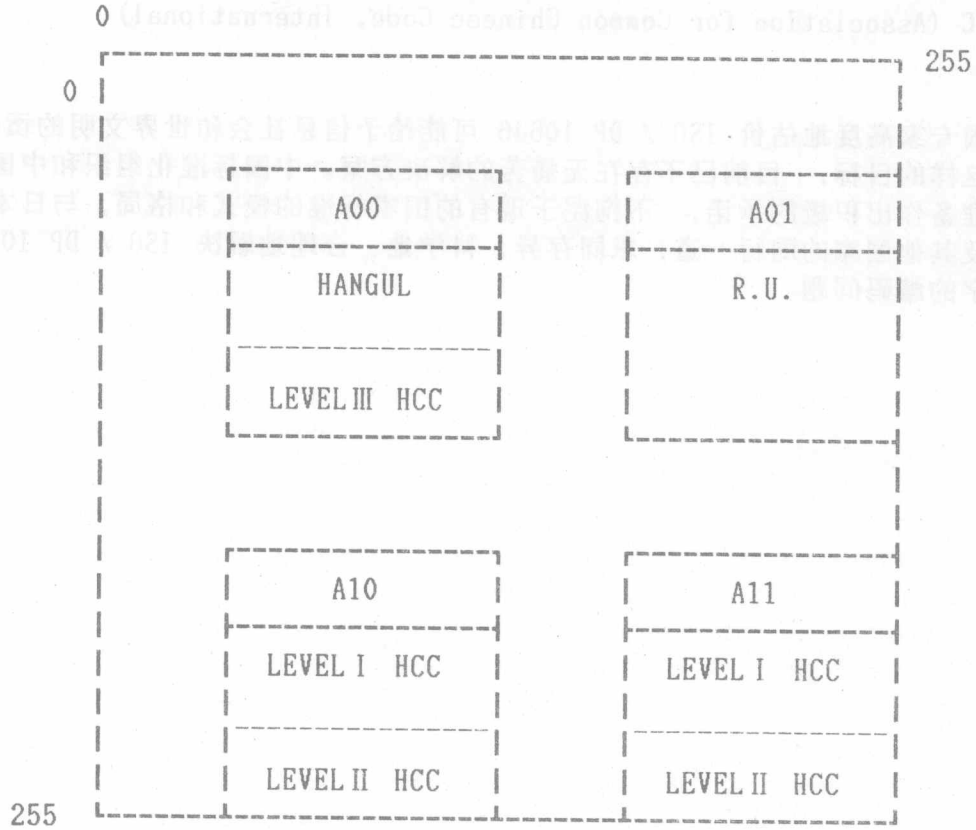
- 由于同一汉字在不同地域发音不同，采用音序已不可能。
- 形序是比较现实的处理方法。可推荐采用，但，

● 若采用部首序，必须首先产生一个 CJK 共同确认的偏旁部首的集合。
● 若采用笔画法，亦需确定笔画分解规则。

- 频序也是可以考虑的方案。

2.3.6 字位的安排

- 从左半行到右半行致密排列，然后转至第二行。余类推。(下半平面)上半面，原 I00 区仍按左半行、右半行地排列。
- I 级 HCC 占 176-216 行，其中 215-216 行留空，以备增补。
- II 级 HCC 占 217-255 行，其中 245-255 行留空，以备增补。



B.M.P

基本多文种平面(032平面)

2.4 HCC 辅助集

- 安排在 033 平面，或
- 不建立这一集合。若有需要，由 C/J/K 民族平面给予补充。

3. 实施办法

- 由 SC2/WG2 确定建立 HCC 的原则，划定 B.M.P 上 HCC 的区域。
- 组成由 C/J/K 专家参加的特别工作组

- 实施认同甄别，决定取舍。
- 研究排序规则。
- 提出字形规范建议与样本。

– 所有这些工作，都要与相关的组织协作：

- ISO/TC46/SC4/WG1
- ISO/IEC JTC1/SC2/WG2
- ISO/IEC JTC1/SC18/WG8
- AFII (Association for Font Information Interchange)
- ACCC (Association for Common Chinese Code, International)

4. 结束语

中国专家高度地估价 ISO / DP 10646 可能给予信息社会和世界文明的贡献。为了达成这样的目标， 目前已不存在无痛苦的解决方案。中国标准化组织和中国信息产业界准备作出积极的承诺， 不拘泥于现有的国家标准的模式和格局，与日本、南朝鲜以及其他国家的同行一道，求同存异，科学地、合理地解决 ISO / DP 10646 中表意文字的编码问题。

附 录 A 几种可能采用的方案的评述

A.1 方案 I

与现行方案冲突较小，但它面临两个困难：

- 现行 ISO/DP 10646 版本分配给中文的I10这个区域，既与中文交换码不兼容，又与广泛实际使用的内码(处理码)格局不兼容。
- 一个区域(ZONE)，远远不敷使用。特别是，与二千多个简体字相对应的繁体字，以及一批经实践证明应补充的汉字，已经没有足够的空间容纳。事实上，根据中国文字学家和信息处理专家的保守估计，若把繁体字也考虑在内，中文信息处理用的汉字基本集的字量应在 10,000 以上。

在一个ZONE(7500字)的前提下，曾考虑两种解决办法：

- 简繁同码
- 削减本应进入 B.M.P 的中文汉字字量。

本质上，这都是“削足适履”式的处理方法。

- 简繁同码的优点是使得在某些 1:1 的场合，转换简便易行，同时又虚拟地扩充了一个 ZONE 的空间，缓和了空间紧张的矛盾。
- 简繁同码的缺点是

  - 必须要有附加的标识信息(或HOP序列，或ESC序列，或其它控制字符) 来指明简繁的差别，否则不可共用。
  - 带来了一码两字的不确定性，有违 DP 10646 UNQIUE/ABSOLUTE 编码的初衷。

- 削减中文汉字在 B.M.P 的字量，将迫使编码方式改为3-Octet方式，或者频繁地使用HOP序列实现平面间的切换，从而大大地降低了系统的处理效率。其缺点也是显而易见的。

A.2 方案 II

倘若 C/J/K 退出 B.M.P，将会产生一个突出的问题：怎样有效地 (以双八位形式) 实现拼音文字与 C/J/K 的共用？出路无非两条：

- C/J/K 的汉字向 B.M.P 的 I00,I10,I11 区“动态映射”。
- B.M.P 的 A××区的内容向 C/J/K 平面“静态映射”。

动态映射表面上提高了 B.M.P 的利用率，但带来了一个如何在 B.M.P区分 C/J/K 的严重问题。至少，必须定义一个控制功能来实现动态映射。一旦我们定义了这一功能，我们或许会发现，新标准似乎蒙上了 ISO 2022 “切来换去”的阴影。况且，在这种双八位环境下，C/J/K 的共存共用亦非易举。

静态映射的本质是冗余编码，这也是有悖本标准的基本精神的。

A.3 方案III

－ 优点：

- 最大限度地消除了表形文字的冗余编码，充分地利用了B.M.P 的宝贵空间。
- 支持最大多数用户实现拼音文字与表形文字（ C/J/K ）在双八位环境下的共存共用。 同时也解决了简繁体汉字并存的问题。
- 与ISO/IEC JTC1/SC18/WG8 和ISO/TC46/SC4/WG1的发展趋势合拍 （统一C/J/K）：按文字种类，而不是按国家/民族编码。

－ 缺点

- 与 C/J/K 各国现行的标准不兼容，特别是字序不一致。
- C/J/K 汉字的认同 甄别、字形规范化、研订排序规则等，需要大量的协调与协商的工作，需要一个制订的过程。
- 达成协议最为困难。

附录B HCC 参考数据

B.0 引言

C/J/K 汉字归并为一个字符集后，容量究竟多大？这在很大程度上取决于归并时采取的策略，即认同/甄别的原则。

B.1 日本国文学研究所田岛一夫的对照工作是将 GB 2312 与 JIS C 6226/1983 进行了比较。JIS 汉字 6353 个，GB 汉字 6763 个，相同 + 相近 = 3070 + 1648 = 4718 个。从 GB方来看，69.8 % 的一致；从 JIS 方来看有 74.3 % 的一致。百分比相当高。但是应当注意，田岛一夫先生在比较时，没有考虑中国简体字与繁体字的差别，并在实际上往往采用了简体字的繁体形式去做比较。

B.2 中国国家语委所作的分析比较，是属于较为严格的形式，因此，GB汉字与 JIS 汉字： 相同 + 相近 = 1851 + 1393 = 3244个。这样，从GB方来看，只有48% 的吻合率；从 JIS 方来看有 51% 的吻合率。吻合率低的重要原因之一是 GB 中严格排除了繁体字和异体字。

B.3 美国图书馆系统 REACC（东亚字符集）

共用部分	SINOTERM	10,934
独有	GB	2,337
独有	JIS	785
独有	KIPS（南朝鲜）	7
独有	CCCII	537
假名、韩文符号等		1,600
总计		16,000

B.4 C/J/K 的交集

根据中国国家语言文字工作委员会的严式比较，C(含台湾)/J/K 现有标准字符集的交集共有 1,029 个汉字。

这就意味着，C/J/K 字符集的归并，按最保守的估计，至少可以节省1,029×2=2,058 个汉字的空间。