ISO

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION

ISO-IEC JTC1/SC2/WG2

MULTI-OCTET CODED CHARACTER SET

ISO-IEC JTC1/SC2/WG2
N 513
October 16, 1989

Title	:	Further Explanation on HCC
Source	:	China
Requested Action	:	For Consideration at the Ad-hoc Meeting in Beijing For Consideration at the WG2 15th Meeting in Amman

对 HCC 的进一步说明

致：ISO-IEC JTC1/SC2/WG2 (1) Ad-hoc 会议 (北京)
(2) 第 15 次会议 (安曼)

送：(1) ISO-IEC JTC1/SC2
(2) ISO-IEC JTC1/SC18
(3) ISO TC46/SC4

自：中国

目　　录

引言

1. 基本事实
1.1 CJK 汉字属同一种文字体系
1.2 文字的比较分析
1.3 CJK 合并≠ CLG 合并

2. 基本需求
2.1 各国家标准进入最惠平面
2.2 需求的增长

3. 求大同存小异

4. SC2 的编码之道
4.1 字形是编码时考虑的首要因素
4.2 广义字形的概念是有益的
4.3 按文字的种类而不是按国家／民族分类编码
4.4 表意文字编码规则尚待研究
4.5 严格控制冗余编码

5. 兼容性问题
5.1 字符集的兼容性
5.2 兼容性的损失是发展的代价

6. 结束语－中国的建议

引　言

HCC 的提出，乃是基于下面的一个最基本的事实和一个最基本的需求：

- 基本事实 － 中国汉字、日本汉字和南朝鲜汉字属于同一种文字体系 (script)，尽管它们之间存在着这样那样的差异。
- 基本需求 － 包含简繁两种形式的中国汉字、日本汉字、南朝鲜汉字，应尽可能多地与其它非表意文字共存于同一平面(B.M.P)，以便在大多数情况下，汉字都可以用定长的、双八位的形式表示、存贮、处理和交换。这将大大提高 B.M.P 乃至整个 ISO 10646 的可用性和使用价值。

本文将对上述两点详加解释，并阐述与之相关的观点。

1. 基本事实

1.1 CJK 属同一种文字体系

无人否认，汉语、日语和朝鲜语是三种不同的语言。然而，众所周知，由于地理的和历史的原因，中文汉字、日文汉字和朝鲜汉字(以下简称 CJK) 同属一种文字，这是一个不依人的意志而转移的客观事实。

这是因为：

- CJK 汉字都是表意文字。
- CJK 汉字具有相同的渊源，这已为大量的史料所证实。
- CJK 汉字在三个层次上：

> 笔画相同：　一　丨　丿　丶　㇀

> CJK 汉字的基本构件（部首，偏旁）绝大部分是相同的、近似的、可以认同的。即使 C／J／K 的独特字，也是遵循相同的构字方式（如畑－会意结构方式）构成的［参见朱岩、康加深文章］

说到汉字的基本信息单元，现代的汉字输入技术与理论已经证实，在一定程度上这些部首就是基本的信息单元；但也有一种观点认为，基本的信息单元仍然是字。

> 整字：CJK 信息交换用字符集大面积交叉覆盖。下一节的数据可以表明：

* GB 2312－80	的独特字	只占自身的 32%
* J1S x 0208	的独特字	只占自身的 11%
* KS C 5601	的独特字	只占自身的 2%
* CNS 11643	常用字的独特字	只占自身的 7%

1.2 文字比较分析

由于种种原因，汉字在各个国家和地区都有某些规定的字表或字集。下面的比较数据，从多个侧面说明了这些字表/字集间的偶合关系。毫不奇怪，由于认同规则和比较内容、比较方式的不同，这些数据之间会略有出入。但这并不影响我们看出CJK 汉字之间的区别与联系的总轮廓。

1.2.1 根据我方搜集到的日本专家所作的“日中国家标准汉字比较表”，我们作了一些统计分析。

- 它说明，除去中简/日繁和中简/日异对应的情况，中日汉字完全相同的和具有微小差别的达 3,393 个，占 GB 2312 的 50 %以上。
- 那些具有微小差别的中日汉字，几乎全部可以按 JIS x 0208 和中国最近的建议加以认同。其中类似于“文”／“文”之别的，竟达577 对之多！
- 如果拿 GB 2312 非简化形式与 JIS 比较，相同部分将会大幅度上升，达 80%。
- 将 JIS 中的 645个独特字进一步与中国的繁体字集辅 3、辅 5 加以比较，可以发现只有 271 个字是真正独特的，只占 JIS x 0208 的 4%。

1.2.2 JIS Ⅲ 级字与中国 GB 1,3,5 集的比较也说明，中日汉字之间有很高的重叠率，达 4,094 字，占 JIS Ⅲ 的 70%。

1.2.3 多种分析都表明，南朝鲜汉字 与 中日汉字有很大的交集。

- 从 Unicode 可以看出，拿 KS 汉字与 GB，JIS 的汉字相比，只有 347个字无交叉，占 KS 所有汉字的 7%。
- 另一项研究是将南朝鲜文教部规定的大中学用汉字（1800字）与中国 GB 2312－80 非简化形式（GB 1）加以比较。结果发现：

  > 1,791 个字（占 99.5 %）均可在 GB1 中找到。
  > 只有 5 个字（占 0.3 %）是独特字，在 GB 中找不到对应体。

1.2.4 繁体汉字

在 JIS 和 KS 中，包含大约 2515 多个字与 CNS 中的字(常用字居多)相同。经过分析，这些字中，大约有 63% 就是 GB 2312－80 中某些汉字的繁体形式。中国已在 辅 I 集中收入了这些汉字。

CNS 中的次常用字与 GB 3 相比，相同+相似的汉字有 2382+221 之多。

同时，北京图书馆将CCCII(21882 字) 与 GB 辅1,3,5 (23263 字) 进行了比较，比较结果是：

- 双方对应：16,394 字
- CCCII独有：5,488 字
- GB 独有：7,070 字

1.3 CJK 合并 ≠ CLG 合并

任何一种文字的内部，都会由于时间和地域的不同，派生出几种有差异的集合。差别是普遍存在的，问题不在于是否有差别，而在于这些差别的大小、性质和处理的方法。

中国汉字、日文汉字与南朝鲜汉字（CJK）和西里尔文字、拉丁文字与希腊文（CLG）这两大类文字之重叠部分有巨大的差别：

●相对数量（重复字符占各字符集的比例）

	CJK	:	CLG
=	68～80%	:	33～44%

●绝对数量（混合编码时可节省的码位）

	CJK	:	CLG
=	6000～10,000	:	50

所以二者之间有质的不同。

CJK 字符的归并与 CLG 字符的归并是完全不能类比的事物。从根本上来说，由于 Cyrillic, Latin 与 Greek 具有的共同字符太少，即使加上与发音符组合形成的字符，也为数甚微，因此，很难认定它们是一种文字。所以，由 CJK 归并＝CLG 归并的假设前提推导出的结论都是没有意义的。对CLG 归并的否定，并不能推导出对 CJK 归并的否定。

可以推算，分别编码和混合编码，对 CLG 来说，节约的码位大约是五十个以下；而对于 CJK 来说，则可节约上万个码位。

2. 基本需求

2.1 各国家标准进入最惠平面

不必讳言，在多八位体系的诸平面中，基本多文种平面（032 面）对于信息交换和信息处理，都是“最惠平面”。

对于中、日、南朝鲜等国，基本多文种平面可以

- 保持其字符代码的双字节、等长模式，而不受 ESC／ HOP序列或其它伪码处理的困扰。
- 与世界上最主要的各文种并存共用，而无需在平面间进行切换。

因此，各国标准化组织都希望自己的编码字符集进入这个平面，同时，保持着其

●字数 ●字形 ●字序 ●字位

的一致性。

这些要求是可以理解的； 同时，各国的需求又是在发展着的、并且相互之间是有矛盾的。

1986年在撰写 N146 时，囿于当时的条件，人们没有，也不可能照顾到这些发展变化着的情况。 但是当今我们在制订一个牵动全局的国际标准时，应当而且必须实事求是地处理这些矛盾。

2.2 需求的增长

1986 年以来，信息世界有哪些重要的发展需要我们认真对待呢？

- 南朝鲜正式颁布 KS C 5601-87
- 日本积极制订 JIS Ⅲ 级汉字表的（5000字）
- 中国制定了第 2 、第 4 辅助集
  GB 7589 － 87
  GB 7590 － 87

相应的辅 1 、辅 3 、辅 5 也在制订之中（繁体字）

- 中国制定了《现代汉语通用字表》（7000字），删去 GB 2312中 170 字，增收 407 字。
- 台湾制订了繁体字的 CNS 11643-86 (13,051字)

这些发展，表明了表意字符集无定界的特征和扩充的趋势。

基本多文种平面不可能收容如此之多的字符集， 但是，对于其中最重要的两项则应予以考虑：

- 朝鲜谚文。
- 中国繁体字， 常用的约 2,235字，与简体字相对。作为国际标准，绝不可缺。 因为，在港、澳、台地区， 繁体汉字仍然是主要的应用形式；即使在中国大陆，在古籍、古汉语读物中及其它有特殊需要的场合，也均使用繁体汉字。

此外， CJK 的汉字使用情况不同：南朝鲜以谚文为主，日本有假名相辅，中国则是全用汉字，所以，汉字用量自然要比日本、南朝鲜为多。正如 N480 所指出的，为满足一般中文信息处理，总共需要 10,000字左右。

3. 求大同存小异

无可否认，CJK 字符集之间，确实存在着差异。

这些差异，可分为两类，一类是结构性的， 如 図/图，広/广，気/气等，毫无疑问， 对这些字应给予不同的代码；另一类是笔形上的，笔划数量上最多是一划之差。对这一类，也还可加以分析：

- 有的差别，是本质性的，表征着某种国家的或民族的文化习惯，如团 / 团，对/对，单/单等，中国提交的“认同规则”已建议对这些字作单独处理，分别编码。
- 有的差别， 是非本质性的，纯粹是笔形新老之差，不是国别所致。如：文/文，户/戶等，对这种情形，JIS x 0208 已经将它们认同，中国提出的认同规则也主张把这类字予以认同。

应当强调指出， 所谓认同，只是赋予同一代码。而在 rendering 时，不同国家与地区可以有不同的 presentation，即 glyph 或 typeface。如文/文，一切面向日本的产品都为“文”，一切面向中国的产品都为“文”。在国际多文种环境，文和文都应视为合乎标准，不存在谁比谁更规范的问题。

我们希望这是一条求大同存小异的正确途径， 它既未修改也未消灭任何文字，因而也未触动任何国家的文字政策， 不涉及文化与法律的问题。 从 Glyph 的观点看，也是科学及现实的。

HCC 将犹如一个共享的（只读）数据库，每一个国家都有自己的视图（ VIEW ）和子集，但不排除在必要时可以方便地访问和使用其它的内容。

总之， HCC 体现求大同存小异； HCC 可能为未来的某种文字统一奠定一些基础，但它本身并不是文字改革，更不会抹杀任何国家和民族的文化。

4. SC2的编码之道（The way SC2 has coded characters）

纵观 SC2 编码的历史和现状，我们可以清楚地看出几条规律性的东西：

4.1 字形是编码时考虑的首要因素

文字的四属性： 形、音、义、源，都是编码时可考虑的因素。但文字是语言的物质外壳， 作为书面语的字符， 编码时，字形永远是第一位的、主要的因素；但不是唯一的因素， 有时还要兼顾音、义、源。应当承认，这种情况是个别的。不应把个别的情况夸大为一般的原则，如同形异音字
编码，同形异义字编码在国际标准中都是屈指可数的个例。

4.2 广义字形的概念是有益的

尽管 SC2 和 SC18 之间尚未就字符与图符（glyph）的区别与联系达成正式的共识，我们仍可以看出， SC2 从来是按照形和义编码的，而且，此处之“形”，乃是一种抽象的、内在的、逻辑的概念，它是

- 独立于字体
- 独立于字号
- 独立于设计

的一种广义字型。 笔画之粗细、曲率，字形的长宽比、斜度、走势等具体的细节，从来不是，也不应是 SC2 关心的范畴。

这一概念，引伸到表意文字无疑是有益的，因为它可以控制本来就很巨大的字符集的膨胀，使众多的书法的、字体的变异归为一体，大大地节省编码空间而不致混淆字符的本征。

4.3 按文字的种类而不是按国家／民族分类编码

这一趋势清晰地体现在 ISO 8859 之中。

在 8859－1 中，SC2 从 14 个语言中抽取了共同的拉丁字符加以归并，形成了适用于44 个国家的字符集；在 8859－2 中，则是从10个语言中抽取了共同的字符加以归并，形成了适用于 9 个国家的字符集。

在这里，即令是法语、德语、西班牙语这样一些著名的语言，也没有分别独立编码。阿拉伯文、俄文亦然。

这对于各国信息的沟通、文化的交流，无疑是很大的推动。显然，CJK的归并不应与 CLG 的归并相提并论，相反的，它符合英、法、德等文字按文字种类编码的国际潮流。

4.4 表意文字编码规则尚待研究

如何处理传承字(通用字)与繁体字／异体字之间的关系，如何对多音字进行编码，这些问题在各国家标准中已有不同方式的处理。但是在国际标准中，如何统一地掌握这些尺度，还是一个待商榷的问题。

4.5 严格控制冗余编码

4.1，4.2，4.3 均体现了这一点。在宝贵的基本多文种平面，应当更好地贯彻这一原则。为了实现与使用的方便，建议将这一原则修改为"有控制的冗余"，但这种冗余仍不应发生在基本多文种平面内。

5. 兼容性问题

5.1 字符集的兼容性

字符集的兼容性可分为以下四个层次：

- 字量
- 字形
- 字序（字集中各字之间的相对关系）
- 字位（字集中各字相对于代码坐标的关系）

尽管 HCC 会对兼容性带来损失，但在以下几点，仍然是有益的或无损的：

- 字量，不但不减，反而会大幅度增加。
- 字形，求同存异，从各国角度看，都是不变的。
- 字位，没有违背 C0，C1 限定的格局。
- 对日文假名，朝鲜谚文，完全保留了从字量、字形、字序直至字位的兼容性。

当然，最重大的损失是汉字字序的变化。

由于各国家标准和 HCC 之间，不存在简单的转换关系，所以大对照表必需使用，转换的效率会为此而降低。

WG2 的任务便是，如何在这种字序的变化和字量的增加之间作出权衡，斟酌得失。

5.2 兼容性的损失是发展的代价

上述兼容性的损失可换取的利益主要是：

- 字量的巨额增加，可高达 10,000 之多，在这种情况下，中文繁体字，日文 JIS Ⅲ 以及 GB 3 的汉字都有可能全部收入 B.M.P。
- 避免了3 Octet 或不等长码的处理，保持了现存亚洲文字双字节的最基本模式，大大地简化了软件的实现。
- 简化了字库 (FONT) 的设计与实现。

6. 结束语－中国的建议

ISO 10646 中表意文字的编码是一个极其重要、极其复杂的问题。这个问题处理得当，可能会对信息技术的发展产生深远的影响。在可预见的将来，我们已很难再找到制订国际字符集标准的机会。所以，我们应极其慎重地，从长远的需求出发，捕捉住这"最后的"机会，实现真正国际化的表意文字编码。

我们一向认为，汉字不仅是亚洲人民的宝贵文化遗产，也是世界文明的共同财富。中国标准化组织和信息产业，愿意和各国专家/组织一道，共同作出贡献，努力推动 ISO 10646 的实现。

由于 ISO 10646 是一个具有战略性的标准，涉及一系列重要的原则，我们建议，SC2/WG2 的专家认真地研究我们提交的文件和资料，并将有关问题提交到 SC2，SC18 以及 TC46/SC4 共同研究讨论。