关于 CJK 变体字 Variant 的定义问题

DATE：2014-08-18~08-20

CGP 张轴材

最近 ICANN 顶级域名生成规则(TLD LGR)的讨论，促使我对两个问题进行了反思。一个是中文变体字的定义，另一个是与此相关的变体字的 Block 问题。这两个问题涉及比较根本的规则。本文先就 Variant 定义做一初步修改建议，期望引起大家注意，更深入地讨论。

摘自三年前中文变体字研究报告中的定义：

Selected from Report on Chinese Variants in Internationalized Top-Level Domains (2011). Chinese (character) variants are:

> "characters with different visual forms but with the same pronunciations and with the same meanings as the corresponding official forms in the given language contexts.[1]"

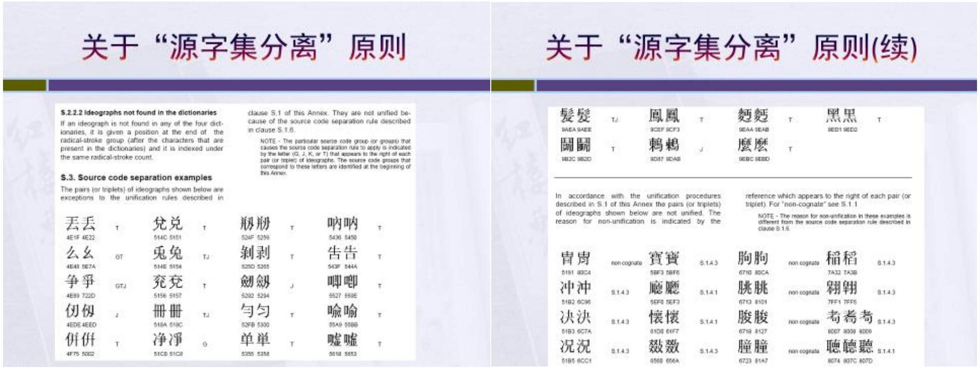
[参考中译]与给定语境下的规范字/正字音义相同而字形有异的汉字。

在我们 TLD 的特定环境下，目前看来，这个定义有一些缺陷：

1 这个定义中没有指明我们是在"异码"汉字的范围讨论问题；因为许多"可见的字形差异"已经在 Unicode 中被认同了（比如下图）。

它们不再是我们讨论的异体字或变体字。

而少数字形很相近，但由于"source code separation"rule 而分别编码了。而它们很可能是互为变体的。（参见下图）



为了更严谨，在 Unicode 的前提下，Variant 的定义建议修改为：

> “characters with (different visual forms *hence) with different codes* but with the same pronunciations and with the same meanings as the corresponding official forms in the given language contexts.[2]”

2 每个汉字都可能有不止一个读音，不止一个字义。注意定义中的 pronunciations 和 meanings 都是复数。实际情况是，两个变体字之间的两组音义可能全等，也可能部分相等；部分相等的情况，也可能是主要音义相同，也可能是个别音义相同。这些情况应当如何在变体字的定义中界定呢？如果不加限定，Variant cluster 可能会很庞大，比如：钟锺鍾鐘鈡；參葠蔘叄叅蓡薓叁。我主张进一步修改 the same pronunciations and with the same meanings 为 the same *major* pronunciations and with the same *principal* meanings，这样可以大大减少这种 big variant cluster,简化变体字的处理。

3 该定义指出了“给定语境”的条件，但是还没有涉及“跨语境”（across language context）的问题。在跨语境的情况，由于文化差异和政策差异，没有必要强调孰正孰异。所谓规范字/正字(official)和“异体字”(variant)都是相对的。因此，在跨语境的条件下，as the corresponding official forms in the given language contexts 这段文字似可删去。另外，由于语境的不同，字音差异也很大，特别是中日汉字之间。因此，在我们的 TLD 项目中，变体字的定义可以不考虑“字音”(pronunciations)的因素。

于是，变体字定义似可进一步修改为：**基本字义相同的 CJK 异码汉字**。

> *CJK Ideographic Variants are those separately encoded Chinese Hanzi, Korean Hanja or Japanese Kanji with the same basic meanings.*

4 CJK 变体字的关系具有三个重要属性：

4.1 耦合强度 The coupling strength amongst variants 变体字之间的关系是有强弱之分的。比如，在一组 Variants 中，岛-島-嶋-嶌-隝，前三者可能是耦合较强的，而后面两个就是耦合较弱的；在另一组，并-並-幷-竝-傡-併，前三个耦合可能比较紧，后面四个则比较松弛。这从个国家地区的字频统计数据，或者字符集的编号、级别都可大致看

出。

4.2 语境相关 Dependency on language context

对一组变体字中，如：气-氣-気-炁，对 hans,hant,主要变体是气-氣，对 jpan 则是：气-氣-気。気在 hans,hant 语境中不出现，炁在现代语境中完全不出现。

4.3 基本对称少数不对称 Basically Symmetric with minor Asymmetric

从一个语境向另一个语境映射时，大多数 Variants 之间的关系是 1:1 的对称关系；少数 Variants 之间是非 1:1 的（1 对多，1 代多，或者多对多 M:N）。值得注意的是，这些非 1:1 的 Variants 往往是用频度较高。（这通过压缩 MSR 和 Block 若干 Variant Mapping，可以大大减少和限制 Asymmetry ）