

关于 CGP 的目前工作

2014-08-15~2014-08-20

张轴材 CGP

各位：

1. 我个人认为目前阶段我们应集中力量尽快解决 Generation Panel 的字汇 (Repertoire) 问题。

Variant Mapping 可以考虑，但目前还不是重点。

当前的重点应是清理削减现有的字汇，严格控制它的 Size。大家都可理解，我们面对的是一个现代的、通用的、全球的、狭义的文字应用领域-顶级域名用字。注册者、用户追求的是域名的特异性（当然还有意义、易懂、易记、响亮等等需求），而不应是单个汉字的特异性，不应追求生僻字、罕用字。Internet 用户不希望遇到认读、输入、显印的困扰。压缩 TLD 的字汇，使之更加通用和实用，这是简化异体字处理的关键。

那些不成字的笔画、包括吏读在内的古字不应当留在 GP Repertoire 里。前几次会议已经指出过。但这只是少量的问题。

更大量的，主要的是非现代的、非通用的罕用字，而且他们往往是某种规范字/正字的异体字。应当根据各国家/地区的现代通用字频统计数据作为依据，尽量滤除。我测算这类字在 MSR 中占 3000-6000 字。目前要下决心解决，否则是个大包袱。特别是那些变体过多的汉字，比如：(岩) 巖巖巖巖巖巖，(國) 國國

囯囯囯囯，(回) 迴迴回回迴，这些变体字对于 TLD 几乎没有实际的使用价值，徒然增加变体字处理的复杂性。

2. 关于日文用字。实际上很多是中文古代用汉字，只有少量的是日本国字、略字。原来 CDNC Table 中已经吸收了大量的日文 unique 用字，大约 300 字（看附件 JP1.XSLX）。现在日本又提出了近 400 字（看附件 JP2.xlsx）。CGP Repertoire 处于有增无减的态势，我们必须认真研究。

有一些日文汉字或者与之对应的中文汉字，确实是比较常用的高频字。比如 勞

单 団 囯 実 広 弾 応 栄 黒 (in JP1.xlsx) 和 処 対 桜 気 発 砒 脳 积 駅 (in JP2.xlsx)，收入这些字应无大的异议；但是大多数 JP1 和 JP2 中的日文汉字看来都不是现代常用字，尽管它们都处于基本多文种平面。

我建议：

第一步，如果日本同意 CGP, KGP, JGP 共用一个统一的 Repertoire，我们当然不反对把 JP1 和 JP2 都纳入 CDNC 同一个字表，形成一个并集 Union。

第二步，我们共同用各家的现代字频统计数据对这个 Union 进行筛选，滤除那些过于古老、频度过低的汉字，无论是来自何方。只要没有任何一方支持它的实用性，就删除之。

第三步，在各语言环境分头 Block。对 Variant Mapping 的 Block，有两个问题需要讨论：

1. Block a mapping to a specific script only or to ALL CJK script ?
If Japan blocks 发-発, does China is unable to map 发 to 発 ,
or China maybe still able to link 发 and 発 ?
2. Is register able to have Local or International preference ?
From the Government' s position, the TLD label be in the
normalized form, say 检察厅 (in hans) or 檢察廳(in hant), not
檢察庁(in Jpan). On the other hand, a non-governmental register
would like to have its website more internationalized. e.g.
the mappings amongst 厅-廳-庁 are all permitted, hence the
labels 音乐厅, 音樂廳 and 音樂庁 are all allocatable.

More examples: under hans-hant-jpan

图书馆-圖書館-図書館

医学会-醫學會-医学会 [0]

心脑血管-心腦血管-心脑血管

围棋对决-圍棋對決-囲碁対決[1]

期刊杂志-期刊雜誌-期刊雜誌[2]

快乐驿站-快樂驛站-快樂駅[3]

Note [0] More combinations would be ill-formed ,hence may be
limited/blocked by others rules.

Note [1] 棋-碁 is mapped in the combination 囲碁 under jpan.

Note [2] 刊-棐 is blocked because of the low usage frequency
of 棐.

Note [3] This is not translation from Chinese to Japanese, but
only Hanzi to Kanji mapping. The better expression would be 楽
しい宿場?

补充：

听说有日本同事认为国/國，發/発之间在日文语境中不是 Variant 的关系。这使我反思到 ICANN Chinese VIP Report 对于 Variant 的定义甚至译法都需要重新斟酌、细化、修改。请参见我另外一篇关于 Variant 定义的文件。

我为此对着两对汉字在日文 JIS 中的资料。显然他们的“同义”是无疑的，只是用法、用处不同。这不影响他们的异体关系。

日本工業規格（案）
 JIS X0208-1996?
 附属書 1・2
 公開レビュー資料

1995.12

INSTAC

40-15 767A	発 482F C8AF	105(ア)-4	S5193	M22662'	[常] ハツ, ホツ	→66-04(發)
---------------	-------------------	----------	-------	---------	------------	-----------

66-04 767C	發 6224 E2A4	105(ア)-7	S5194	M22669	あはく, おこる, つかわす, ハツ, はなつ, ホツ	→40-15(發)
---------------	-------------------	----------	-------	--------	-----------------------------	-----------

66 X 0208-1996?	25 区 81 点 ~					
25-81 56FD	國 3971 B9F1	31(口)-5	S1318	M4752	[常] コク, くに	→51-91(圖), 52-02(國)

52-02 570B	國 5422 D4A2	31(口)-8	S1319	M4798	くに, コク	→25-81(國), 51-91(圖) 注 1
---------------	-------------------	---------	-------	-------	--------	-------------------------