再说变体字，CGP-JGP

Date：2014-09-20~21

To: ICANN/CGP members

From: Zhang Zhoucai

Subject : Again，Variants Definition & Concept

Note：我现在没有时间将此文改写为英文，同时，用英文有的地方也难以表达。抱歉暂时如此。

首先，我觉得对于这个“变体字”的基本概念、基本术语还缺少沟通、缺少共识。虽然英文都是Variant，但在我们TLD 的范畴内，概念已经变了，不再是传统说法的“异体字”了。不知哪位先生从何时起翻译为“变体字”了，Chris 建议英文Lexical Alternate，也是比较接近的。这种表达是明智的，避免了很多政治性问题，更加符合我们项目的实际了。

在讨论Mapping, Allocattable/Block之前，建议我们彻底把Chinese Variants/CJK Variants的定义概念讨论清楚。

因此，建议下次双周电话会议工作语言用中文，起码先在说中文的组员中形成共识。我写的那个CJK Variant定义稿建议可以当靶子讨论。

这不是Variant广义/侠义的问题，而是“转义”了：对中文而言，本来简体/繁体谁也不认为是Variant，大陆不以为然，港台也不以为然，现在都同意了是“变体”；有的异体字，两岸就是互为正异的，在TLD 没必要争论孰正孰异，比如决U+51B3和決U+6C7A，反正对应了，音和义都一样，互为变体字。扩大到CJK，主要是C-J，跨语境了，过去从来没人定义过跨语境的异体字，但客观上是存在着对应关系的，只不过现在靠“同音+同义”判断不行了，只能靠“同义”+“同源同用”来界定。这样的CJK变体字可能会有很多，但从TLD的角度，我们可以只选择那些常用的、高频的，还有易混淆的汉字编成CJK 变体字组。

下面是一些常用CJK Variants的实例。

Hans-----HanT-----Jpan

事U+4E8B事U+4E8B亊U+4E8A

处U+5904處U+8655処U+51E6

壤U+58E4壤U+58E4壌U+58CC

对U+5BF9對U+5C0D対U+5BFE

专U+4E13專U+5C08専U+5C02

恼U+607C惱U+60F1悩U+60A9

插U+63D2插U+63D2挿U+633F

曾U+66FE曾U+66FE曽U+66FD

樱U+6A31櫻U+6AFB桜U+685C

栈U+6808棧U+68E7桟U+685F

气U+6C14氣U+6C23気U+6C17

焰U+7130焰U+7130焔U+7114

烧U+70E7燒U+71D2焼U+713C

兽U+517D獸U+7378獣U+7363

瘦U+7626瘦U+7626痩U+75E9

发U+53D1發U+767C発U+767A

矿U+77FF礦U+7926砿U+783F

团U+56E2團U+5718団U+56E3

脑U+8111腦U+8166脳U+8133

图U+56FE圖U+5716図U+56F3

厅U+5385廳U+5EF3庁U+5E81

稳U+7A33穩U+7A69穏U+7A4F

在中文变体字组中引入日文变体字（主要是日文新字体字）有利也有弊。弊端是带来了异体字组的复杂性，但我们选择高频常用字来关联应当可以减少复杂性；有利的方面是，第一，扩大了TLD Label的国际性，第二，也可以避免有意无意或恶意的域名混淆。

以图书馆为例：

图书馆

圖書館

図書館

圗 舘

啚

其中 图有5个Variants（不算日文図是4个）,书有两个Variants,馆有3个Variants(从文字学角度远比此数字多)；

可能的Label组合不算日文有4\*2\*3=24个，算上日文有 5\*2\*3=30个。根据各国家地区的实际应用和频度，刨除“ill-formed string”,实际Allocatable Labels可能只有这三个：图书馆for hans，圖書館 for hant ，図書館for jpan。

同样的例子，

医学会for hans 醫學會for hant 医学会 for jpan

音乐艺术 for hans 音樂藝術for hant 音楽芸術for jpan

开发for hans 開發for hant 開発 for jpan

类似这样的CJK Variants 衍生的CJK Variant Labels，视为一个TLD entity好呢？还是多个entity 好？这里的利弊需要在CGP–JGP讨论,也需要ICANN的总策略来决定。

我们可以注意到，在跨语境的情形，Varian 的对称性symmetry和可传递性transivity 在绝大多数情况是成立的。

个别的非对称、不可传递的问题，多半与上下文有关，这种情况是极其个别的，如发-發-発 与髪的关系，需要人为处理，不应总为这种个例影响总体的规则的讨论。