J3

Comments to IP's feedback on Yoneya's algorithm

(1) about the annotation on slide #3

Q *Could you elaborate more on what this annotation means? Especially, we need to know exactly how 'script' should be used in RootLGR project.*

For example, Japanese language uses Kanji, Hiragana and Katakana. In this case, is 'Japanese script' a proper usage? Or, is 'Kanji script' a proper usage?

A "Script" is used in two senses. A "pure" sense, mostly used in Unicode, where Han, Hiragana and Katakana are "scripts". But also loosely, in the sense of ISO 15924, where "Jpan" is a code for "script" that represents the mixture of these three element scripts.

When speaking about Unicode properties (and classification of characters by them) then the IP uses the term script in the former sense. When speaking about the scope of a "single-script" GP, the IP uses the term "script" in the latter sense. (The IP usually avoids speaking of a 'Japanese script', preferring expressions such as "repertoire corresponding to script code 'Jpan'" or similar wording).

(2) about WLE (to annotation on slide #4)

Q *Does this annotation intend to be based on the situation "some of the WLE rules may be for strings that contain characters in multiple scripts"? For example, it refers to WLE rules for strings containing both Kanji and Hiragana, which are very common as Japanese words ?*

The annotation says that "WLE rules are not script-specific in the procedure: there is only a single, common set.". Does this mean every LGR-2 has the same WLE? If so, development of RootLGR won't finish until the last Script LGR finished. Or does it mean every LGR-2 which shares the same script has the same WLE?
If the answer is latter, CJK coordination steps has to have WLE integration step which will require human intervention.
I would like to see the concrete image of LGR-2.

A In practice, not all WLEs apply to all labels. The various <rule> and <action> elements are often crafted so that they get triggered only for code points or variants from specific subsets of the Root

Zone repertoire.

Formally, though, "WLE rules are not script-specific". That means, when the IP integrates all LGR files from the various GPs, it will create a single <rules> element (which contains all the individual <rule> and <action> elements).

For the CJK GP's this issue has more importance, because so much of the Han script repertoire is shared. Any <rule> that is effective for strings from the Han script, must be the same for C, J and K. However, note, that any <action> that is triggered solely by variant-types can be written to be selective. If a certain variant type is only defined by the J GP, then any <action> triggered by some combination of this type is only ever triggered by a J label.

Because of this, in practice, the problem is not as difficult as it could theoretically be. It is entirely like that the IP would be able to release, for example, a first version of the Root Zone LGR covering the Arabic and Armenian scripts, without risk of introducing compatibility issues, if the second version of the Root Zone LGR were to add the C, J, and K LGRs later. It is extremely unlikely that any Arabic WLE would ever be triggered by a CJK label or vice versa.

Yes, WLE integration will use human intervention, and this is one of the areas that the IP will help. One of the things the IP will do, if necessary, is to rewrite the XML so that any apparent conflict between unrelated WLE rules can be avoided --- as long as that can be done without changing the intended result of the rules.

(3) about the term 'disposition' (to annotation on slide #6)

Q *The annotation says that "Note: in the XML schema this is done via the "type" attribute, on the <var> element and only labels have dispositions". When we say "disposition" in LGR context, does it mean label disposition? How can we say "attribute on the <var> element" shortly? Attribute? Type? Disposition type?*
It is very helpful if we (IP and GPs) use the same terminology.

A The integration panel usually speaks of "variant types". Or, if several types can lead to an allocatable label, they might be referred to as "subtypes".

(4) about (o) (disposition) type (to annotate on slide #11)

Q *The annotation says that "Should be (b)?". Someone feels it should be (o), and someone feels it should be (b). Whether it is (b) or (o) doesn't make the result of WLE different (the label is invalid). Is our understanding correct?*

A For a reflexive variant, where code point C is mapped to itself, as in C-->C, the type should always be (o), whenever C is not part of the repertoire.

In the unlikely case that a reflexive variant is defined for a code point C that is in the repertoire, but should not be part of a valid label, then "blocked" (or some other type) can be used.

There is a small difference between "blocked" and "invalid". A blocked label does exist and is tested for collisions against registered labels (from any script). An "invalid" label does not exist and for that reason would not be tested for collisions.

(5) about minimization of allocatable variant labels (to annotate on slide #12 & #13 & #16)

Q *The annotation says that "Overall, the goal of the IP is to help the CJK-GPs to arrive at coordinated LGRs, such that the number of allocatable variant labels is minimized.". My understanding is that selection of delegated label(s) from "allocatable label(s)" is by decision of applicant. And it is out of scope of RootLGR.*

A As you point out, there are two sets of labels "allocatable" and "delegated". The IP understands the required minimization to minimize the first set, to make it as small as is required. The applicant may then decide to select out of that small set.

For some scripts, initial drafts of the LGR would allow hundreds of allocatable labels. The IP does not look favorably on any LGR that has the potential to generate a large number of allocatable variants; it would make it completely unpredictable which variant or variants might be chosen by the applicant. In contrast, if a well-chosen set of variant types can be used to limit allocatable C-LGR labels to "all traditional", "all simplified" and "original label", the applicant has effectively three items to choose from (a bit more, because the applicant is free to pick a particular "original" label).

With such a design, users would know to attempt to reproduce a label in at most three forms (as seen, as written in traditional and as written in simplified form). One, or perhaps all, of these can be expected to be successful.

Random mixtures of variants, on the other hand, would seem to rarely lead to an improved user experience. That was behind the procedure's imperative to minimize the number of allocatable labels.

Q *For Japanese language, it is impossible to predict which character is allocatable and which is blocked individually because all characters are allocatable independently. It depends on the applicant's thought. JGP doesn't say all the variant labels should be delegated. The number of delegated labels will be at most 2 or 3. But, for example, which labels are appropriate among the 16 variant labels to be allocated cannot be predefined in the RootLGR level. Such limitation should be done in application proposal and its evaluation because all the variant labels are valid in Japanese case.*

A The question would be, whether, in this situation, there really is a overriding need for the same Japanese applicant to apply for multiple variants of the same label. And whether such situation is really common.

Without interaction of the C-GP's LGR, and its definition of variant, we understand that Japanese usually view these characters as unrelated, and that in the second level, two labels might independently be registered in a Japanese domain that would be variants in a Chinese domain.

If these labels are really considered independent, wouldn't it be the case that they commonly are not registered by the same applicant, but by different applicants, just like the domains .color and .colour in English can be registered fully independently?

If that is indeed the common case, then it would seem that making all of the variants introduced by the C-LGR "allocatable" in the J-LGR doesn't really address the need. Because they would then be restricted to the _same_ applicant.

The IP would want to have evidence that such a scenario is not only common, but essential.

Finally, if the applicant ends up picking 2 or 3 variant labels out of a set of 16, for example, how will the users be able to guess which ones to try?

Allocatable variants create definite costs in the DNS. This cost should be offset by a clear benefit to the users, by meeting an essential need for both applicants and users.

Q *As a summary, we should not try to solve all the issuees by RootLGR,we should think about appropriate mixture of RootLGR and application evaluation rules/guidelines.*

A For the Root Zone, it would be prudent for the LGR to be a bit more restrictive, which would also make it more predictable. This is carefully addressed in the procedure, for example in the Conservatism Principle.

(6) about Han and Kana mixed labels (to annotate on slide #16)

Q *The annotation says that "What about the case where a label is applied for that combines Han and Kana characters in Japanese. Which cases do we expect that the Han-variant can be substituted for the applied for Han code point in such a mixed label? -- This issue is specific to JGP, but it is affected by the design of the Han variants -- unless JGP has a WLE rule to "block" variants for mixed labels altogether.". For Japanese language, Han and Kana can use with any combination. As mentioned above, it is impossible to predict which character is allocatable and which is blocked individually. Mixed use of Han and Kana does not affect it.*

A It would be instructive to know of examples of actual registrations from the second level where the same applicant had registered two labels differing on only the Han component -- and where that applicant is treating them as variants (leading to the same content).

It would also be instructive to know how common this phenomenon is, and whether it is "defensive" in nature. (Similar to registering a "typo" for an English label, so as to avoid it being captured by an unrelated party). Such "defensive" use would be unnecessary if the variant is "blocked".