**Comments on "Procedure to Develop and Maintain the Label Generation Rules for the Root Zone in Respect of IDNA Labels", Version 2012-12-07b**
John C Klensin, 2013-01-07[1]

This report is obviously carefully-researched and well-thought-out.  While one can quibble about details, it seems likely that changes would not lead to significant improvements given that one accepts the objectives and context of the report and a large number of starting assumptions.  Unfortunately, all of those are questionable and should be examined by the community.   After deliberating as to whether to include a detailed analysis in this public comment, I've concluded that I should defer to my own observation that long reports in the ICANN context are rarely read and studied carefully, leading to decisions that ignore important issues (I hope these four pages of notes are not, themselves, too long).   Consequently, the format that appears below is intended to facilitate identification of key issues and a deeper discussion of them without going into great detail.  The details have been shared, albeit often informally and incrementally, with relevant staff members and the project team.  Some have been posted or published in other forums.  I would be happy to discuss any of them further if there is interest from Board members, key stakeholders, or others who are materially involved.

*What is the report about?*

> While the report is identified as about Label Generation Rules (LGR) and is the product of a project devoted to variants and variant handling, its real topic is the determination of a character repertoire for the root. While variant-specific topics are covered, they are very nearly incidental and, in my opinion, leave many open issues and risks.

> Whether the process model of this report is a reasonable byproduct of the Variant Project or not, it is clear that some mechanism for determining an appropriate character repertoire for the root is needed (such a determination is actually an IDNA2008 requirement).   Merely using what IDNA2008 and a current version of Unicode allows would not only violate the standard but would  either create considerable security and stability risks or would require case-by-case evaluations that are unnecessary in practice.   Although some of the considerations outlined below still apply, the proposal of this "LGR" report would almost certainly be among the better models ICANN could invent with for defining that repertoire if either the "variant" components that make up a small part of it were eliminated or simply used to flag a label as needed further case-by-case evaluation (as noted below, despite the added complexity of the variant considerations, it may not accomplish much more than that anyway).

*Are "variants", especially in the root, consistent with the design of the DNS and ICANN's security and stability mission?*

> While there may be legitimate exceptions for specific script or language relationships (or not) the report (and the Variant Information Project generally) makes the *a priori* assumption that variants should be carefully considered for legitimacy and appropriateness for all scripts.  The community has never examined, via an open discussion in which all of the major issues and tradeoffs have been exposed, that question, nor has the Board ever initiated a public review or made a documented decision on that subject.   The "Variant Information Project" was constituted to gather information, which was entirely appropriate, but there has never been any public evaluation of whether that information justified moving forward with more work on, and structure around, variants.  Instead, that conclusion has been assumed and this round of work was initiated on that basis with a scope that prevented analysis of the more basic questions.  This (and other second-round reports) should be considered, along with the "Integrated Issues Report" of the first round, to determine whether the complexity, risks, and costs of a variant system are justified by the utility of variants and the number and kinds of situations they can and cannot actually accommodate and resolve.

[1] PO Box 400197, Cambridge, MA 02140, USA.  The production of this review and associated comments have been supported, in part, by ICANN but the comments are strictly those of the author and do not necessarily reflect the views of ICANN Staff or any other members or groups in the ICANN community.

I continue to believe that the default action should be a slightly updated version of the one recommended to the Board by the committees headed by Masanobu Katoh almost a decade ago: that those who desire names about which they see some relationship should apply for them separately and that those who see conflicts or risks about particular names or relationships should have access to an efficient and inexpensive objection process.    The complexities of the "variant" process implied by this and other reports appear to me to be justified only if the community, after a careful evaluation, concludes that they will be appreciably less expensive, more efficient, and/or will spare ICANN and its stakeholders and participants the risks and costs of difficult, case-by-case decisions.  Not only has that default position not been evaluated by the community (Appendix I of the current report notwithstanding), but the complexity of this proposal, and that of the Integrated Issues Report, somewhat enforces the view that such an evaluation is not only necessary but that is might well favor an alternate course of action.

It falls well outside the scope of the LGR report and this review, but, yes, there are alternatives to variants. Some of them open up fundamental questions about the role of internationalization and localization in the DNS and the Internet more generally.

*Are the "variants" addressed by the proposed rules the most important and relevant ones given user and community expectations?*

The system proposed in the document is based entirely on relationships among characters, or short sequences of characters.  The variant relationships with which we have the most experience are actually based on semantic and orthographic relationships even though, as a result of the properties of "ideographic" scripts, they have been characterized in the form of character relationships[2].    The report has treated any sort of semantic, and most orthographic, relationships as out of scope unless they can be characterized as strictly character relationships.

*Will the mechanism proposed meet community expectations?*

It seems to me that the community's assumption has been that having a good set of Label Generation Rules would largely eliminate the need for subjective, case-by-case, evaluations of whether a label, or set of labels, are appropriate.  In large part because it does not address relationships among complete labels, this draft doesn't accomplish that goal.  Indeed, the rule framework it proposes would be insufficient to deal with several cases ICANN has handled already in the context of the ccTLD Fast Track.

*Is the process outlined consistent with ICANN community norms and responsibilities?*

The mechanism the report describes is very technocratic and extremely dependent on a small group of experts ("the integration panel") who are appointed by staff (not a public process) and from whose decisions there is no appeal.  The individuals from that panel can be subject to public abuse (through the public

---

[2] Because many of the discussions about "variants" have started from extrapolations of the JET specification (RFC 3743) and there has been widespread misunderstanding within ICANN on this point, it is perhaps worth a brief explanation.  In an alphabetic script (ignoring distinctions within that broad category), the character-symbols usually represent phonemes (not necessarily the same phonemes in different languages) and are associated with glyphs (that may be more or less different or similar depending on type styles) but, regardless of their origins, they do not convey character-specific meanings.  To the extent to which a script is "ideographic", the characters actually have meanings (sometimes more than one).  The decisions that make the relationship between Simplified Chinese (SC) and Traditional Chinese (TC) characters other than one-to-one were associated with different meanings, making some choices valid in contact and others not.  While the SC↔TC relationships are expressed in the CDNC tables (see RFC 4713) in character form, the closed analogy in, e.g., Latin script would be cases like the difference between British and American spelling of the same word, not, e.g., evolution of the forms of characters.
That understanding is different from the one I had when I wrote RFC 4290 (cited in the report).  To the extent to which it makes a difference, I believe that 4290 was a serious mistake.

comment process) and, under conditions the report doesn't seem to specify, fired and replaced by staff, but neither of those provisions is a safeguard against a small set of bad decisions. Having people fired by staff — especially if they are paid consultants— obviously creates its own set of risks, especially if staff has some investment in outcomes, such as via a link between ICANN funding and the number of applications considered or accepted. It is also not clear that a fully-appropriate panel can be chosen (as the report notes, some of us have observed that the requirements include omniscience) although it seems certain that, if staff is told "appoint a panel", they will do so with no plausible possibility of their returning and saying "no appropriate panel can be constituted". Such a technocratic process has considerable appeal to those who, at least on some days, believe that expertise is more important than public or stakeholder opinion, but it is not how ICANN has made (or claimed to make) key decisions that affect multiple user and stakeholder groups in the past.

In particular, there should be no delusions that the proposed LGR model is somehow going to avoid setting policy for allowable labels and applications without the use of any of the usual policy development processes.

### Can a label generation rule process define the "disposition" of characters and labels?

One of the key issues with "variants" is what should happen to a set of labels that should be treated as a group. All but one of them can be blocked (not permitted to be allocated under any circumstances in the same zone) or some or all of them can be candidates for allocation and, possibly, delegation. The report presumes that the decision between blocking and possible allocation can be made on a per-character basis (more specifically on the basis of the LGR-assigned properties of the characters in the proposed label). Experience in the ccTLD Fast Track indicates that there are labels that should be allowed or not allowed at all based on the whole string, not just the properties of individual characters, and that there is every reason to believe that issue would apply to variant handling as well. Of course, as some members of the team that have produced this report have indicated, a careful and responsible integration panel might conclude that all variant characters (except possibly Chinese ones) should be blocked. If that were the case, then ICANN might have to deal with exception requests through some subject process (see above) unless it were politically feasible to reject all such requests.

### How does this relate to "visual confusion"?

At least since the "Paypal demonstration", the ICANN community has been painfully aware that, when users are expecting to see one script, carefully-chosen characters from another one can be used to aid in name-spoofing and associated attacks. The first line of defense against that problem is to prohibit labels containing characters from more than one script or to allow them only in well-understood and carefully-specified circumstances. The LGR report contains suggested mechanisms that should get that right. However, in the years since the community started talking about the paypal exploit, most of the people in the world who are interested in such attacks have figured out that it is not difficult to select labels, all of the characters in which are members of the same script, that will provide equal or better opportunities for confusing the unwary (several examples of these have been given at talks at ICANN meetings). The proposed LGR process doesn't address labels that contain nothing but characters permitted with a consistent "tag". To the extent to which ICANN still feels responsibility for those confusing full-label cases and the attack (phishing and otherwise) opportunities they present, it needs to understand that the proposed LGR system won't help and other mechanisms will be necessary. To some extent, this is a just a special case of the more general question above about community expectations.

### What happens with endangered languages and language communities who are not yet active in the Internet?

The procedural model specified in the report carries with it significant risk of having decisions made that would make it far more difficult to construct sensible mnemonics using languages and writing systems that are not active enough in today's Internet (or ICANN's community) to get early requests for, and participation in,

generation panels. As the report suggests, there may be no practical way to avoid that problem (I agree with the report that waiting until there is full knowledge of every script, writing system, and language is not practical). On the other hand, the only protection against such problems is the general knowledge available in the integration panel (this situation is one of those that has produced comments about a requirement for omniscience). To the extent to which the ICANN community believes and is on record as supporting the goal of allowing every language community, especially those who are now at a disadvantage, to get on the Internet and to create DNS mnemonics that make sense for their languages and writing systems and for a "level playing field" for DNS TLD applicants in the future, it needs to be generally understood that a "charge ahead with the languages and scripts that now have an constituency within ICANN" approach is inconsistent with those goals.

*How does the proposed LGR process interact with IDN applications that are now in the system?*

There seems to be a general understanding among the participants in drafting this report and in those involved in variant policy development more generally that there is no practical way to apply these policies to TLD IDNs that have already been applied for, at least without retroactively stopping processing or all such applications (Appendix A should perhaps note this, but it is clear from other parts of the report and independent discussions). ICANN's history has been one of carefully adopting narrow and restrictive rules and then relaxing them as that proves safe. Going ahead and approving labels under current procedures that might then be prohibited by the results of the LGR (but, as the report suggests, retained in the root) is a potential narrowing of a too-liberal policy. Moving in that direction creates a very high likelihood of complaints about early-applicant advantages and unfair treatment of those who apply later. I see no alternative other than stopping processing of existing gTLD IDN applications until the label generation rules (or even the non-variant character repertoire it contemplates) are complete for the relevant scripts, but that risk should be understood and explicitly accepted if this report is to be adopted.

*Risks and risk assessment*

The comments above, and several sections of the report, identify cases where the process outlined could fail, either to meet expectations or more seriously. From the standpoint of the report, the most serious risk is that of the integration panel discovering that it has to change a then-existing rule. From ICANN's, the possibility of a serious security or stability problem than can be directly attributed to negligence in the design or development might be even worse. At least the first set of risks is quite real as scripts that support multiple writing systems are considered and pragmatic decisions are made (see, for example, the comments in Appendix A about Arabic and note that there has been considerable demand for IDNs in variations on that script). That particular situation, in which the integration panel could be faced with a choice between making a high-risk decision that is politically necessary and making important constituencies very unhappy (especially with some strings in that script already in the root) is a particularly good example in the present community but by no means the only area of risk.

I can only hope that the ICANN decision-making process can take these risks into account and, if appropriate, accepts them explicitly. That explicit understanding should include the knowledge that some of the risks of going ahead with a general "variant" policy on the basis proposed could easily fall, not into the category of problems that could easily be corrected down the road, but into the "you bet the organization" one.