Review and Response to

ICANN IDN Variant Issues Project: A Study of Issues Related to the Management of IDN Variant TLDs

John C Klensin[1], 31 January 2012

General Observations about the Report

If one were inclined to find fault with this report, it is fairly easy to attack. The sections are uneven in quality, there is considerable essentially redundant text, there are a few substantive errors, some of the assumptions are not justified by facts and others are not applied consistently. It is also long enough that I fear that few people will really read it carefully beyond the Executive Summary, an Executive Summary that is not completely consistent with the body of the report. At the same time, it attempts to be thorough, thoughtful, and is considerably better in quality, analysis and understanding of the issues, and balance than the ICANN average.

I am extremely pleased that the report tries to consider the issue from the standpoint of the user. Too many studies in the past (including, e.g., the "IDN Testbed") have concentrated primarily on the ability to place records in the DNS and retrieve them, not on whether those records and that process meet user needs.

This review attempts to avoid the problem of deterring people from reading it because of its length and in-depth analysis. For much the same reason, it avoids identifying particular sections of the report except when that is clearly necessary. There are also places where this review is in general agreement with the report (or at least with one of the ways that it can be interpreted) and I have generally not tried to identify those separately. Instead, it tries to identify the main issues with the report that require further, or different, analysis, and options the report omitted, particularly for those issues that affect next-step decisions and policy-making. Its author would be happy to engage with the ICANN community on more specific details if that were useful.

This Review and the Report

This overview is not a substitute for the report's Executive Summary, which should be considered mandatory reading, even for those not inclined to read the full report, and is not further summarized here. However, the Executive Summary does not address several of the issues with the report that discussed below. This review will be most useful to those who have actually read, or at least skimmed, the report itself.

[1] [1] Preparation of this review and report was partially supported by ICANN. The opinions expressed are those of the author. They do not represent an official ICANN position and may not reflect the opinions of ICANN Staff or Management.

The Key Issues

1. What is the Question?

If one judges from both this report and most of the team reports, the question that was asked was close to "If ICANN provides a variant mechanism that can easily provide additional names to solve whatever problems can be identified, how would you use that mechanism?" The report addresses what those variant mechanisms might be, what problems or issues exist with various scripts, and what other issues and opportunities "variants" cause (in most of this review, the term "variant" appears in quotes to emphasize the point made in Issue 3 below). After a review of the report, it seems clear that other questions, if taken seriously, would have produced different responses from the script teams and a very different final report. For example, suppose the teams had been told "If it is even possible to create a variant mechanism other than blocking, it is likely to impose significant additional restrictions and costs on the management of the domain(s) and is likely to delay evaluation of the TLD application, possibly for years rather than months. Given that constraint, what are the needs of your script and what variant mechanisms are worth considering?"

The analysis in the report provides most of the motivation for the latter question. Other than blocking and multiple delegated domains that are managed in some specific way, we do not understand the possible alternate mechanisms that could, in theory, be deployed today well enough for deployment to be wise (we do understand, and the report includes, some of the issues and risks). New mechanisms (and adequate deployment to make them usable) would be, at best, years, and probably a decade or more, off.

I do suggest that there is at least one criterion for "variant" deployment that the report does not emphasize strongly enough and that should also have been part of the question: Other than blocking (see issue 2 below), however "variants" are conceived of and implemented, they are going to be difficult to manage in a way that is entirely beneficial and problem-free. As such, it is useful to differentiate between proposed variants that are really important and variants that would be "nice to have" under one scenario or another. If a variant set that does not depend entirely on blocking is really important, then we should assume that deployment of one of the labels in the set without deployment of the others is not acceptable. In other words, the approach generally used in the ccTLD Fast Track of delegating a primary label but reserving identified variants for future consideration should not be considered acceptable. If the additional labels are really necessary to create an acceptable user experience, then installation in the root should be for all of the requested labels or none of them, with no assumptions about adding other variants later. I note that was exactly the situation with the Han-script labels delegated under the ccTLD Fast Track: the position of the applicants was that it was simply not going to be acceptable from a user experience standpoint to delegate one of the pair of labels without the other.

2. The "Blocking" Option

There is certainly one "variant" management mechanism that we know how to implement today. It is to somehow create a list of names, delegate one, and then block all of the others. The report mentions it in several places but then ignores it as an option in discussions that, while appearing to be more general, focus on different ways to make alternate labels visible to the user. It also assumes that "variant" models based on blocking require the same processes for creating variant sets, etc., that the other models do.

I suggest the community consider whether handling blocking that way just creates more mechanism, structure, complexity, and the associated costs without any real advantages. The Applicant Guidebook already includes procedures by which people can object to proposed names for any of a wide range of reasons (some claim effectively "any reason at all"). Rather than putting applicants, ICANN, and others to the trouble to establish and utilize databases of variant sets, might it not be more sensible to create a "variant" objection category (possibly with a different fee structure) and let TLD applicants and those to whom TLD labels have been delegated monitor their own labels for conflicts or find entrepreneurs to do it for them? Nothing would prevent such a strategy from incorporating announcements from TLD applicants or operators of lists of names to which they would object should someone ask for them. And an objection-based strategy (whether proactive in the form of announcements or reactive in the form of objections after applications were announced) would have the further advantage that ICANN would not need to figure out what a "variant" is or how to identify one (see Issue 3 below) or how to agree upon and establish "variant label generation rules" (see Section 4 of the report).

This approach to blocking is consistent with, and perhaps even recommended by, the report, but the recommendation is somewhat buried, and very easy to miss in the much more lengthy discussions of how to do more difficult or impossible things. Worse, things that can be accomplished in the general case only by subjective evaluation or objection processes are described in ways that some may read as indicating they are straightforward (for example, see "It is expected that there will be mechanisms to detect these,…" in Section 3.2, Class 2(b)).

3. What is a "Variant" Anyway?

The various documents that preceded the report, and parts of the report itself, make it clear that the term "variant" is used in different communities for a number of different concepts. One of the initial goals for the report was to establish a definition for that term that would be usable across the community. For a variety of reasons, including disagreements among the various script teams, the effort was not successful in that goal: we have a better list of the different things different communities mean by "variant", but no clear definition or even a closed set of categories. ICANN should consider that a cautionary message vis-à-vis giving "variants" special treatment.

While the report explores these issues at some length, many elements of the discussion are essentially hand waving that hides important cases. As a superficial example, the pseudo-definition involving "conflated with each other" in Section 3.1 would allow both words (sic, see

Issue 4) with national or dialect-based orthographic differences ("colour" and "color") and translations from one language to another to be treated as variants (that issue is discussed in Section 3.5, but is not mentioned in the Executive Summary and may be easy to miss). It is worth noting in this context that the implicit assumption that the presumed need for variants is limited to IDNs may not be correct: while the DNS has apparently worked well for the last quarter-century without special mechanisms to treat British and American spellings of the same conceptual word as equivalent, if national or dialect-based orthographic variations start being given special treatment in other scripts, there is every reason to believe that similar demands will arise within the communities who have been comfortable using only Basic Latin characters.

4. Mnemonics and Words

There continues to be a huge amount of confusion in the ICANN community about what a DNS label (at any level of the tree) represents. The traditional DNS view is that such labels are simply mnemonics. If that is the case, then there is no requirement that it be possible to express all of the words of any given language, or even any of the words of some particular language, as labels. Seen that way, the "internationalization" requirement is only that it be possible to use characters drawn from locally-recognized scripts to construct such mnemonics; fine points of orthography or typographic rendering are simply irrelevant. At the other extreme, there is an assumption that no system is adequate unless the vast majority of words (or even phrases) in each relevant language can be written in the DNS and understood in the ways in which people can understand them. The difficulties with that assumption are that the DNS as we know it simply cannot accommodate it and that "variants" won't help. For example, while many search engines today can successfully look up words and phrases despite spelling errors, the rigid, known item search, mechanism of the DNS could not do so unless all possible spelling variations and errors were somehow identified, listed, and treated as equivalent (whatever that actually means).

Perhaps worse, there is considerable desire to do some things differently within a given script based on the language in use, but, not only does the DNS have no mechanism for identifying the language associated with a particular label, but there is no feasible mechanism that would not cause problems on lookup if casual users misidentified the relevant language for a string they were trying to copy. Again, the report recognizes this problem. But then several sections of it proceed on the basis that the problem does not exist and that it is reasonable to talk about language distinctions.

As an illustration, parts of the choice of the subset of ASCII that makes up what we now refer to as the "LDH rule" were fairly arbitrary. That subset could have been chosen to prohibit the use of any vowel, making writing of any of the words of English (and most or all other languages that use Latin script) impossible. It would, however, have had no significant negative effect on the utility of the DNS for mnemonics. Interestingly, it would have made the issues and choices facing us today much more clear.

The report recognizes the "words" problem, mentions it, and then, in several sections, lapses into talking about "words" and issues that are relevant only if the goal is orthographically and typographically-correct presentation of those words.

5. Localization and "Let the DNS Do It"

For many of the situations for which "variants" have been seen as a solution in the ICANN community, there are alternatives that are completely different from trying to put multiple labels into the DNS and establish some sort of equivalence relationship. One of the best examples is mentioned in the report: for domain name labels that contain digits, one strategy is to put only a single digit character form into the DNS but then treat input and display of those digits in the local script as a local presentation matter. Digits are not now permitted in TLD labels, as the report notes, but the approach could, in principle, be generalized to permit a certain amount of canonicalization of strings to match local requirements to what is stored in the DNS – processing that would go beyond what IDNA already requires. Taking this type of approach would require that we revise somewhat what we mean by a "consistent user experience", but the user experience is already not completely consistent when two different people expect to be able to type in two different strings and have them treated as more or less the same name.

This comment is not a criticism of the report in any way. Instead it is an attempt to ask whether we have slipped into the territory of "variants and the DNS are the answer, now what is the question?", rather than examining the actual requirements and then the full range of options that might satisfy them.

6. The "Script Team" Approach

This effort was based on the assumption that careful examination of six scripts by script-specific teams with staff and external expert support would shed considerable light on the "variant" issues. That approach was more successful than some had predicted, but its inherent limitations, the limitations created by the composition of the teams and how they functioned, and the risks implied by those limitations should be understood by anyone trying to understand the wisdom of making policy decisions on the basis of the report. In particular, the following issues arise as a consequence of those limited patterns of expertise (most of these are identified in the report, but worth reviewing):

(i) Most of the teams were unable to capture and reflect experience with all of the languages that use the designated script, so we just don't know whether their reports actually cover the script or only some particular cluster of languages that use it. If rules are made for particular scripts based on the composition of the teams (and hence on the groups participating actively in ICANN today), there is a significant risk of those rules working to the disadvantage of communities who might come along later and be dependent on the same script.

(ii) To the extent to which one is concerned about "variants" as a mechanism for dealing with confusability, some scripts must be considered in conjunction with closely-related ones as well as vis-à-vis internal conflicts. The Devanāgarī team recognized this and considered closely-related Brāhmī-derived scripts. At least one joint meeting was held to consider issues that might affect the Greek-Latin-Cyrillic combination.

But similar efforts did not occur for other scripts, even when they might have been helpful, and, while the report discusses cross-script "variants", it essentially acknowledges that no general solutions are likely to be feasible and, for one case, recommends further study.

(iii) For the purpose of this study and others, the boundary of what is included in a script and hence where characters are considered as belonging to a different script is often at least controversial and may be arbitrary. If one needs to pick a classification system with global scope, the ISO-Unicode one is probably as good as any, but any study of this type should recognize that application of the rules and assumptions (other than the largely-accidental history of national coding standards) that caused Latin, Greek, and Cyrillic to be considered as three separate scripts, might, if applied to Arabic or Han scripts, have produced separate script categories for Arabic and Perso-Arabic and perhaps even for Japanese and Chinese (and maybe Han script Korean). Conversely, if the same "unification" norms that were applied to Han script were applied to Western European ones, Latin, Greek, and Cyrillic would almost certainly be one script with different typographic and rendering conventions. The consequences of this issue are much broader than this particular report and affect any ICANN requirement or guideline that treats strings or label applications differently depending on the script to which they belong.

As the report notes, the issue of the limitations of the teams that contributed to this report calls the feasibility of the different models for creating label generation rules (Sections 4.1 and 4.2) into question unless ICANN has a mechanism for establishing teams and expert groups with significantly broader expertise than the teams who contributed to this report. While the report suggests that the problems could be minimized by concentrating on the characters, scripts, and languages most likely to be encountered in TLD applications, the consequences of thereby excluding, or setting up barriers to applications involving developing countries, minority populations, or mnemonics based on endangered language, could be profound.

7. Rules and Serial Conflicts

The report discusses several options for label generation rules and some other rules and conventions. Some of the possible processes that are discussed carry with them a risk that is not fully explored in the report: the possibility that the practical application of the rule-defining process will result in a *de facto* first come, first served policy in which communities with a significant presence in ICANN today –and, in particular, in the constituencies who are most likely to provide the resources and expert personnel to participate in various teams, panels, and expert groups -- create barriers to other script groups. The problems discussed under Issue 6 above are only part of the problem but may be predictive. Will decisions made on the basis of these teams and these six scripts be sufficient and appropriate for other languages using those scripts? For other scripts? Will establishing a set of rules and tables that are appropriate to what the present ICANN participation profile knows, understands, and can advocate for, exclude the next generation of participants because the required rules conflict and the conflicts have to be resolved in keeping established rules stable? Will an applicant with a new script be put at a disadvantage simply because rules have already been established for a similar-appearing script?

It seems to this reviewer that the report minimizes both the problems and the risks in that area, even while it notes that establishing a single and comprehensive set of rules could be horrendously difficult.

A Summary, the Katoh Reports, and a Strawman

One of the high-level inferences that can be drawn from this hundred page long report and the complexity of its considerations, categories, and options is that a comprehensive "variant" strategy is just too much for ICANN (and perhaps the DNS) to handle. Even if the necessary expertise could be gathered to produce all of the relevant rules and tables, some decisions would inevitably be seen as favoring one script or language over another. Only a few people would actually understand all of the details, and they would probably be accused (by groups who did not understand or who perceived themselves as being disadvantaged) of being arbitrary and/or discriminatory. Depending on where they came from, such accusations could easily lead to litigation, unpleasantness in the political arena, and other responses that could cause the entire new TLD program (or at least its IDN components) to bog down.

If two or more labels were delegated on condition that the registry operators preserve some sort of "mirror" relationship, ICANN would also find itself with the problem of trying to define the boundary conditions for such a relationship. As the report strongly implies, that is a difficult task at best. Trying to enforce such rules could lead down the same unpleasant paths as perceptions about script or language favoritism.

If we had a "variant" solution (or even a small cluster of them) that would actually solve the perceived user (and "word" and language) problems, those risks and costs might be worth it. But, as the report points out, we do not have such a solution; we only have partial approximations.

ICANN's very first organized studies of the issues associated with IDNs in the root were carried out by a pair of committees under the leadership of former Board Member Masanobu Katoh. The key recommendations of those reports, somewhat updated for today's terminology, can be summarized as

> Do not attempt to make special application or allocation policies for names (labels) that are somehow seen as linked together. Let groups apply for related names if they wish, but let the applications be considered on their merits, not on the basis of some attempt at making rules for any or all of the cases. Also be sure that there are effective and easy-to-use procedures in place for blocking names that pose problematic conflicts with existing or applied-for names.

At least as a strawman against which this report, proposed future work that extends from it, and possible policies can be measured, I suggest that ICANN consider the option of

- Dropping all notions of special treatment of "variants" from the vocabulary,

- Letting applicants apply for multiple names that they perceive of as related if they wish, and
- Adjusting the existing objection procedures as necessary, doing so in line with the analysis above, possibly including the possibility of proactive objections.

Whether or not it would be desirable to adjust the application model or fee structure to give some special treatment to labels that the applicants believe are linked, is a separate issue, but the answer does not affect this general suggestion.

Some sets of labels allocated this way would "work" along the various dimensions outlined in the report better than others, depending largely on the behavior of the registry operator and associated policies. But that would not be ICANN's problem. More important, it seems entirely consistent with the new TLD model: different applicants ought to be able to experiment with different strategies, with the marketplace rewarding those who get it right and punishing those who do not.

This strawman suggestion would not have a significant effect on the registrant database ("whois") or rights protection sections of the report, but would significantly reduce the requirements for IANA record-keeping and management and for ICANN evaluation of proposals.

While some of us might have guessed, it took this effort and the reports it produced to actually establish the complexity of the problems and the importance of considering that strawman as an alternative. So the effort, at least in this reviewer's opinion, was worthwhile. But it should represent a result from which we can learn, not a path down which we are obligated to go.

Simplicity is an advantage for both ICANN and the Internet. Complexity helps no one ICANN should want to help and can bias the policy process by limiting effective participation primarily to those who can invest the increased resources it requires.