Comments on "Label Generation Ruleset for the Root Zone Version 1 (LGR-1)"
John C Klensin, 2016-01-27

Summary Comment and Overall Conclusion

Acceptance of a single script as LGR-1 appears to be unwise, inconsistent with assumptions in the Procedure, and to pose an unacceptable risk to the predictability and consistency of behavior of the DNS root. That predictability issue was one of, if not the, primary motivation for the Variant Information Project that was the origin of the Procedure and this work. Using Arabic script, given its many special characteristics as compared to many other scripts, coding properties that are different from long-established published Unicode norms (ones on which IDNA2008 is dependent), and the wide number of languages written with it and resulting variations in usage as a writing system as the sole initial example appears to introduce additional risks. That is especially true if most of those writing system variations have not been studied and considered, which appears to be the case. For the generation of top-level IDNs without the special restrictions of the ccTLD Fast Track, the combination of these factors (including others discussed below) to appears to pose an unacceptable risk to Internet stability and security as the use of other scripts in the DNS evolves.

Specific Discussion and Details

Thanks for the invitation to comment on this proposal. My apologies for submission at the eleventh hour. Many other tasks intervened and, as ICANN is certainly aware, processes like this one impose a significant burden on members of the community who are only professionally interested in the work rather than supported for it.

Several of the comments below are based on the understanding that, despite a good deal of discussion during the development period, the Variant Information Program / Label Generation Rules model has never actually been tested before this set of proposals. This version should be considered as that test as well as a specific proposal. In particular, the initial set of proposals and rules have to be considered in the light of whether or not they meet the needs of the Internet, particularly for security, stability, and a minimal (or at least acceptable) level of possible name confusion or ambiguity in the root.

In a few cases, the document titled "Integration Panel: Root Zone Label Generation Rules Generation Rules — LGR-1 Overview and Summary" ("the report", below) leads me to believe that staff and the Integration Panel have interpreted some of the provisions of the Procedure [Procedure] differently than what I understood to be intended during the time that Procedure was being developed. However, I believe the important evaluation criterion at this stage is not how the Procedure is interpreted in those areas but what is right for the Internet. If the Procedure seems ambiguous, then this is the time to clarify it (as prerequisite to LGR adoption) and it should be clarified in the direction of a well-working DNS that provides the desired protections for the Internet.

The issues below are numbered for the convenience of the reader. Neither the numbers nor the order of items have any significance, especially with regard to relative importance.

(1) The user community for a string

The statement of the "Least Astonishment Principle" in the report includes "should not present recognition difficulties to the zone's intended user population". Although it is likely that the most severe possible problems would be picked up as "lend[ing] itself to malicious use", the issue remains that the "intended user population" for the DNS root zone is the entire Internet. One can make a case that labels at the second level and below have enough context that it is possible to talk about intended user populations in terms of scripts or languages. Even that position requires assuming that registries will impose and enforce restrictive naming rules and insist on their propagation down the tree, but the distributed administrative hierarchy of the DNS (as well as recent traditions under ICANN administration) argues against even that. For the root, there is no such context and the user population, intended or not, includes anyone who could encounter a label at the root level.

(2) What is being "integrated"?

As noted in Section 4.3 ("Comprehensiveness and Staging") of the current report, an ideal LGR set would be comprehensive, covering all scripts that might ever be of interest with regard to the root zone. As the report points out (and as was anticipated when the Procedure was being defined), that is not a realistic goal. Not only will some script communities make faster progress than others, but, as the Internet progresses from models that have encouraged homogenization toward Latin Script and European Languages (primarily, but not exclusively basic (undecorated) Latin characters and English) toward a force for preservation and even re-introduction of languages and scripts that are not as well known in industrialized countries, it may be necessary to consider scripts for the root zone that do not even have Unicode encodings today. While opinions probably differ as to how much responsibility ICANN should try to assume for language preservation and recovery efforts, it seems clear that ICANN and its rules should not become an obstruction to such developments.

At the same time, there are shared characters and relationships among scripts, at least some of them due to derivation from common origins. The function of the Integration Panel should be to "integrate" -- to look carefully at any cross-script issues, not simply to provide an additional check on proposals about specific scripts (or collections of scripts that are closely-enough related to justify having the same script generation panel deal with several of them). If scripts are added one at a time with a rule that no code point once allowed can later be disallowed, then there is significant risk of missing an inter-script conflict and giving the same advantages to early root-zone IDN adopters that users of other than Basic Latin labels have complained for years has been given to ASCII. In the case of the LGR process, "early adopters" would be defined, not my readiness of the scripts and interest in them but by the order in which the Integration Panel chooses to consider the scripts in question.

ICANN must strike a balance between "wait until all scripts are ready" and "do things one at a time". Tipping that balance entirely toward the latter carries its own set of risks (discussed in more detail elsewhere in these comments) but, in any event, does not represent "integration" or expose and test the assumptions that go with it (the largely untested tutorial in the present report notwithstanding).

Conclusion: This report should not be adopted or supposedly-integrated LGR tables and rules published until the Integration Panel is ready to consider and adopt several script reports, preferably including at least two of the scripts identified in Section 4.2 as supporting a wide variety of languages (Arabic, Cyrillic, Devanagari and Latin) and at least two scripts from a cluster of scripts that are known to be very closely related with a large number of similar characters and conventions (e.g., Cyrillic and Greek and maybe Latin or Indic Brahmi-derived scripts). Even if, after due (and open and transparent) consideration, ICANN and the Integration Panel judge that to be too strong a requirement, any notion of "integration" and an LGR rule set is implausible without at least two scripts being represented.

***(3) Language integration and Language-dependent Characters Within** a Script.*

At the time IDNA2008 was designed, a strong assumption was made that, if there were two or more possible ways within the same script to construct a glyph (or set of visually indistinguishable forms), normalization, notably NFC, when applied to the relevant code point sequences, would result in the same canonical form for all of them (at least within the ranges of letters, digits, and the few special characters considered Protocol-Valid in IDNA). That assumption turned out to be incorrect as discussed in the IAB Statement dated 2015-02-11 and cited as [IAB] in the report, in the preparatory notes for the LUCID BOF [LUCID-Notes], and described in more detail in an evolving Internet Draft [K-F-Unicode-7]. The decision to prohibit any combining characters (noted in Section 3.2.1 of the report) would appear to eliminate the greatest risks associated with that issue (now often knoen as "the decomposition problem"), but it is difficult to predict what might happen in the future without a more complete analysis of the issue or at least the ability to compare its handling across multiple scripts (see (2) above), the ability to predict future Unicode code point assignments for the Arabic script, and/or assurances that the Arabic Generation panel and/or the Integration Panel had a complete understanding of the various languages that use the Arabic script and how they use them (see (4) below).

(4) Languages using the Arabic Script

As noted in the report and elsewhere, Arabic script is used to write a large number of languages, some from very different language groups with regard to phonemes and how they are used. An additional large number were written in Arabic once and were largely switched to other scripts within the last few centuries. Some of those are now being switched back. For some others, Arabic script has continued to be used as a minority or special-use writing system. That implies a relatively complex environment. The notion in the Procedure, at least as I understood it, was that Generation Panels for particular scripts were expected to reflect a broad range of expertise across, all, or at least a significant and representative sample, of the languages that used that

script. One of the responsibilities of the Integration Panel was to be sure that requirement was met.

However, the composition of the Arabic Generation Panel that produced the proposal on which the report is based is another concern. I applaud the goals and list of languages covered by the Task Force [TF-IDN] and the MESWG Membership [MESWG]. At the same time, there appears to be no representation of languages in Europe (e.g., Bosnian, Crimean Tatar), Central Africa (e.g., Hausa Ajami), Sub-Mediterranean Africa (e.g., Fula), or Central Asia (e.g., Kazakh, Uzbek) all of which are sometimes written in Arabic (see [Omni-Arabic. It is unknown, at least to this author, whether the use of Arabic script in any of those language contexts raises issues not addressed by the Generation Panel. However, the observation that the addition to Unicode of a character specific to one of those languages, Fula, because the starting point for the discovery of the so-called non-decomposable character problem, a problem that challenges a fundamental IDNA2008 assumption (see (3) above), it is at least a reasonable hypothesis that there are issues with the use of Arabic in some of those languages that would challenge assumptions made by the Generation and Integration Panels. There is little or no evidence in the report that those issues, or even the possibility of them, were carefully considered. That hypothesis is particularly important given that similar-looking characters and variants are being considered and, indeed, are the core of the LGR (VIP) work. I do not know whether this would be less of a concern if we had more experience with IDNs in scripts that serve a large number of languages and are used differently in some clusters of them than others, but, with Arabic as the first and only case under consideration for a global label generation rule set, it would seem prudent for the security and stability of the Internet to be extra-careful that the full range of language groups and geography be studied and covered.

References

All citations above that are not accompanied by references below are citations to the items of the same name in the Report. To save time, links and abbreviated citations of well-known works are given in several places rather than complete references in traditional form.

[Daniels-Sect62] P.T. Daniels and W.Bright, _The World's Writing Systems_, 1996, Section 62.

[K-F-Unicode-7] https://datatracker.ietf.org/doc/draft-klensin-idna-5892upd-unicode70/

[LUCID-Notes] https://datatracker.ietf.org/doc/draft-sullivan-lucid-prob-stmt/

[MESWG] https://community.icann.org/download/attachments/40935327/MESWG-members%5B1%5D.pdf?version=1&modificationDate=1382456369000&api=v2

[Omni-Arabic] http://www.omniglot.com/writing/arabic.htm

[TF-IDN] https://community.icann.org/display/MES/Task+Force+on+Arabic+Script+IDNs