L2/17-289

Recent developments in IDNs

ICANN *8/3/17 – Asmus Freytag*

Root Zone Label Generation Rules

There is an ongoing project at ICANN to define Label Generation Rules (LGRs) for the Root Zone. Label Generation Rules define what labels are valid and are available for delegation in a given zone. These rules contain an enumeration of repertoire elements (both code points and sequences) together with context and whole label rules that may restrict some code points from occurring in particular combinations. Up to here, this is broadly similar, in principle, to the way the underlying protocol, IDNA 2008 defines valid code points and their contexts. However, Label Generation Rules also allow the definition of variant code points and the assignment of a disposition to the resulting variant labels.

There are two types of variant labels in the root: “blocked”, which are not available for delegation to anyone and “allocatable”, which may be delegated, but only to the same entity to which the original label was delegated of which they are variants.

Blocked variants are a useful tool to compute certain kinds of strongly confusable labels and to prevent their registration. For the root zone, these include strict homoglyphs (identical appearance), semantically equivalent characters (e.g. some ideographs) and certain homophones (where they are used indiscriminately and interchangeably as in Ethiopic).

Allocatable variants are needed where different parts of the community use different code points to designate what is intended to be the same label. The typical examples for this are Arabic and Chinese. In these cases, someone registering a domain name will almost always want to have it available in more than one form (such as traditional and simplified versions of the same name, or using Arabic-language

vs. Urdu choices for digits or some characters). As this undercuts the exact match nature of the DNS to an extent, strict limits on the number of allocatable labels will be placed. No such limits are placed on blocked variants.

Variants may be of code points or code point sequences. All variants will be defined such that they are symmetric and transitive, which rules out the use of blocked variants to cover similar code points where the similarity is some threshold of full interchangeability.

The Root Zone will support only modern scripts and only characters in widespread modern use, as defined by the Maximal Starting Subset: https://www.icann.org/en/system/files/files/msr-2-overview-14apr15-en.pdf (and files linked within).

Root Zone labels will be limited to be in a single script, except that for Japanese and Korean, the definition of "script" follows that of ISO 15924 rather than Unicode, so the CJK scripts exceptionally make shared use of the Han ideographs. For each script, there is a local panel developing the actual subset and identifying variants, contexts and whole label rules. The first half dozen of these have completed their work and the results are available as: https://www.icann.org/sites/default/files/lgr/lgr-2-overview-01jun17-en.pdf (and files linked within).

Note that this is a draft for which the comment period has just closed. However, the final version, which is expected soon, will not substantially deviate. Work is underway for the CJK scripts, the scripts of India and the main European scripts.

Reference Label Generation Rules

ICANN is engaged in a separate project to formulate Reference Label Generation Rules for the Second Level. These can be seen as "templates" for actual LGRs used by second level registries. They are

intended to promote a more consistent user experience. As some second level registries operate in the context of a single language, the Reference LGRs will exists both as per-script and per-language LGRs

See https://www.icann.org/resources/pages/second-level-lgr-2015-06-21-en (for language-based reference LGRs)

IETF

RFC 7940 – Representing Label Generation Rules in XML

This RFC defines and XML format for LGRs. Up to now, label generation rules were described in various ways, from full-text descriptions in registry policies to a combination of plain text tables and comments, some of them adhering to one of several RFCs, some of them more free-form or archived only as PDFs of formatted versions of such tables. The new format supports all known label generation rules and allows them to be expressed in a machine readable format that can be used directly as input to processors implementing that part of a registry's policies. It also allows common tools to edit or to create more human-readable presentations of the data. (See the Root Zone LGR for examples of both).

https://tools.ietf.org/html/rfc7940

Draft-freytag-lager-variant-rules

This is an internet draft in the final stages before being published as an RFC, titled "Guidance on Designing Label Generation Rulesets (LGR) Supporting Variant Labels". It discusses in some detail the considerations that need to be taken into account when designing well-behaved LGRs supporting variant labels.

https://www.ietf.org/id/draft-freytag-lager-variant-rules-06.txt

Draft-klensin-idna-rfc5891bis

Following the IAB statement that noticed the issue with Arabic character U+08A1 in Unicode 7.0 (IAB Statement on Identifiers and Unicode 7.0.0, document L2/15-026), the IDNA 2008 process of updating to new Unicode versions has come to a halt. There are now several RFCs being drafted in an effort to come to an understanding that allows the normal IDNA 2008 update process to resume. The first of these is a draft titled "Internationalized Domain Names in Applications (IDNA): Registry Restrictions and Recommendations" that spells out that IDNA 2008 all along contained the requirement for registries to apply their own vetting of code points on top of IDNA 2008. Therefore, it is up to registries to safely handle PVALID code points that may be rendered identically to other PVALID code points or sequences, whether by restricting the repertoire, or by making some code points or sequences blocked variants of each other.

While IDNA 2008 contains this requirement today, it is spread out across several RFCs. If this approach is found acceptable, it would avoid the need to redraft the heart of IDNA2008 (RFC 5892) that spells out the rules for basic assignment of PVALID code points.

Draft-freytag-troublesome-characters

As part of stressing the requirement that registries take positive steps of dealing with problematic code points that are nevertheless PVALID, a separate draft titled "Those Troublesome Characters: A Registry of Unicode Code Points Needing Special Consideration When Used in Network Identifiers" attempts to give guidance to registries by collecting known information about code points that must be positively addressed in registry policies in order to avoid issues.

https://www.ietf.org/id/draft-freytag-troublesome-characters-01.txt

This draft contains a tentative list of code points and suggests the creation of a registry that can be updated as additional information emerges.