P1 — Label Generation Ruleset Tool
(a.k.a. IDN Table Format)

Kim Davies

30 August 2012

Internet Corporation for
Assigned Names & Numbers

What we have today

- Many TLD registries use some form of “table” to define allowable code points for registration
  - Simple codepoint eligibility
  - Calculating variant sets
- ICANN (on the IANA website) publishes a “repository” of those tables that are public

Problem

- No consistent format
- To analyse and implement tables for the purposes of LGR work, want to be able to implement tools to read and parse the tables

Goal

- MUST be in a format that can be implemented in a reasonably straightforward manner in software;
- The format SHOULD be able to be checked for formatting errors, such that common mistakes can be caught;
- An IDN Table MUST be able to express the set of valid code points that are allowed for registration under a specific zone administrator's policies;
- MUST be able to express computed alternatives to a given domain name based on a one-to-one, or one-to-many relationship. These computed alternatives are commonly known as "IDN variants";
- IDN Variants SHOULD be able to be tagged with specific categories, such that the categories can be used to support registry policy (such as whether to list the computed variant in the zone, or to merely block it from registration);
- IDN Variants MUST be able to stipulated based on contextual information. For example, specific variants may only be applicable when they follow another specific code point, or when the code point is displayed in a specific presentation form;
- The data contained within the table MUST be unambiguous, such that independent implementations that utilise the contents will arrive at the same results;
- IDN Tables SHOULD be suitable for comparison and re-use, such that one could easily compare the contents of two or more to see the differences, to merge them, and so on.
- As many existing IDN Tables are practicable SHOULD be able to be migrated to the new format with all applicable logic retained.

Non-goals

- Stipulate what code points should be listed in an IDN Table by a zone administrator. What registration policies are used for a particular zone is outside the scope of this memo.
- Stipulate what a consumer of an IDN Table must do when they determine a particular domain is valid or invalid; or arrive at a set of computed IDN variants. IDN Tables are only used to describe rules for computing code points, but does not prescribe how registries and other parties utilise them.

Current status

- Internet Draft first published earlier this year
  - http://tools.ietf.org/html/draft-davies-idntables
  - Good feedback, mostly format nits, some Arabic feedback.
- Partially complete implementation created
  - github.com/kjd/idntables
- Seeking more feedback

idntables / samples

name	age	message	history
..			
README	7 months ago	Initial commit [kjd]	
asia_ja_1.1.xml	6 months ago	Update IDN Table Samples [kjd]	
asia_ko_1.1.xml	7 months ago	Initial commit [kjd]	
asia_zh_1.1.xml	6 months ago	Update IDN Table Samples [kjd]	
bg_bg-BG_1.0.xml	7 months ago	Initial commit [kjd]	
bg_ru-BG_1.0.xml	7 months ago	Initial commit [kjd]	
biz_da_1.0.xml	7 months ago	Initial commit [kjd]	
biz_fi_1.0.xml	7 months ago	Initial commit [kjd]	
biz_zh_1.0.xml	6 months ago	Update IDN Table Samples [kjd]	
br_pt-BR_1.0.xml	7 months ago	Initial commit [kjd]	
cl_Latn_1.0.xml	6 months ago	Update IDN Table Samples [kjd]	
cn_zh-CN_4.0.xml	6 months ago	Update IDN Table Samples [kjd]	
com_zh_1.0.xml	7 months ago	Initial commit [kjd]	
eu_Cyrl_1.0.xml	6 months ago	Update IDN Table Samples [kjd]	
eu_Grek_1.0.xml	6 months ago	Update IDN Table Samples [kjd]	

name_ar_1.0.xml

```
<?xml version="1.0"?>
<idntable xmlns="http://www.iana.org/idn-tables/0.1">
  <meta>
    <language>ar</language>
    <domain>name</domain>
    <date>2011-07-01</date>
    <version>1.0</version>
    <description type="text/plain"># TLD:            NAME
# Script:         Arabic
# Version Number: 1.0
# Effective Date: July 1st, 2011
# Registry:       Verisign, Inc.
# Address:        12061 Bluemont Way, Reston VA 20190, USA
# Telephone:      +1 (703) 925-6999
# Email:          info@verisign-grs.com
# URL:            http://www.verisigninc.com</description>
  </meta>
  <data>
    <range first-cp="0030" last-cp="0039" />
    <char cp="00B7" />
    <range first-cp="02B9" last-cp="02C1" />
    <range first-cp="02C6" last-cp="02D1" />
    <char cp="02EC" />
    <char cp="02EE" />
    <range first-cp="0300" last-cp="033F" />
    <char cp="0342" />
    <range first-cp="0346" last-cp="034E" />
    <range first-cp="0350" last-cp="036F" />
    <range first-cp="0485" last-cp="0486" />
    <range first-cp="0610" last-cp="061A" />
    <range first-cp="0621" last-cp="063F" />
    <range first-cp="0641" last-cp="065E" />
    <range first-cp="0660" last-cp="0669" />
    <range first-cp="066E" last-cp="0674" />
    <range first-cp="0679" last-cp="06D3" />
    <range first-cp="06D5" last-cp="06DC" />
    <range first-cp="06DF" last-cp="06E8" />
    <range first-cp="06EA" last-cp="06FF" />
    <range first-cp="0750" last-cp="077F" />
    <range first-cp="0951" last-cp="0952" />
    <range first-cp="1CD0" last-cp="1CD2" />
    <range first-cp="1CD4" last-cp="1CF2" />
```

Line: 1 Column: 1 XML Tab Size: 4

```
$ python
Python 2.7.2 (default, Jun 20 2012, 16:23:33)
[GCC 4.2.1 Compatible Apple Clang 4.0 (tags/Apple/clang-418.0.60)] on darw
Type "help", "copyright", "credits" or "license" for more information.
>>> from idntables import *
>>> greek = IDNTable("eu_Grek_1.0.xml")
>>> len(greek)
35
>>> chinese = IDNTable("cn_zh-CN_4.0.xml")
>>> len(chinese)
19556
>>> u"测试" in chinese
True
>>> u"測试" in greek
False
>>> u"δοκιμή" in greek
True
>>> for variant in chinese.variants(u"测试"):
...     print variant, variant.ulabel
...
xn--0zwm56d 测试
xn--0zww43d 測試
xn--g6ws64d 测试
xn--g6w251d 測試
```

Current known issues/feature requests

- http://github.com/kjd/idntables/issues

Future

- Wrap this up this year, unless significant new requirements are identified
- In tandem, release proof-of-concept code that works with it
- Longer term: convert existing IANA registry to use format, automated posting with validation?
- Would this be used as the basis for identifying variants in client/server software?