ICANN IDN TLD Variant Issues Project

Presentation to the Unicode Technical Committee

Andrew Sullivan (consultant)

ajs@anvilwalrusden.com
I’m a consultant

*Blame me for mistakes here, not staff or ICANN*
Background

- DNS labels were always in (a subset of) ASCII
- Lots of people don’t normally use ASCII
- Internationalized Domains Names for Applications (IDNA) invented to help
Reminder: two flavours

IDNA2003
IDNA2008
Basic problem

- IDNA (2003 & 2008) expands DNS label repertoire
- The LDH pattern does not fit perfectly in other languages, scripts, or both
- People want DNS labels to work like parts of natural language
What makes a DNS label?

- DNS labels are octets
- Preferred syntax (RFC 1035) is Letters, Digits, and Hyphen (“LDH”)
- Special DNS rule for ASCII
  - Case insensitive but case-preserving
IDNA

• Permit non-LDH characters in label
• Be as compatible as practical with deployed software
• *No* changes to deployed DNS software or protocol
• Provide a list of code points that are allowed
• Map cases that are troublesome (e.g. ZWNJ, upper-to-lowercase) using Nameprep
• To the extent there’s an installed base, this is it
IDNA2008

• Attempt to address some perceived limitations of IDNA2003

• Permits or disallows code points based on code point properties

• Certain incompatibilities with IDNA2003
What’s a variant?

Exactly
Origins of variants

- Starts because of Simplified Chinese/Traditional Chinese issue
  - JET Guidelines (RFC 3743)
- Became model for other issues, not always related
Things people have claimed

- Characters that are substitutable
- “Same words” or “same meaning”
- Sometimes a constraint on child names, sometimes not
Why now?

- ccTLD IDN “Fast Track” process delegated some
  - Not uncontroversial
- New gTLDs under development
- If we’re going to create “variants”, we should be able to say what they are.
IDN Variant Issues Project

IDN Variant Issues Project Phases

- Arabic Case Study Report
- Chinese Case Study Report
- Cyrillic Case Study Report
- Devanagari Case Study Report
- Greek Case Study Report
- Latin Case Study Report

Integrated Issues Report

Development of Solutions
IDN Variant Issues Project

We are here
Comment period to 14 Nov

http://www.icann.org/en/announcements/announcement-4-03oct11-en.htm

and

http://www.icann.org/en/public-comment/
Reports are only about the root

While some of the conclusions may apply to other types of zones, the reports discuss variants for TLDs only.
A planned constraint for TLDs

Current rule is “only letters” (strictly, General Category {Ll, Lo, Lm, Mn})

• No numerals
• No HYPHEN-MINUS
• No ZWNJ/ZWJ

From the guidebook
Restrictions suggested in report

- No combining marks
- No digits
- No archaic
- No Quranic marks

Arabic team
Abdul Salam  

- Arguments for and against
- Refinement of IDNA2008 context rule
  - Issue is lack of shape change
- Questions about resulting variants
Groups of characters

Arabic team

• Identical shape at some position (e.g. YEH)
• Similar shape at some position (e.g. ALEF w/ HAMZA ABOVE)
• Interchangeable use (e.g. KAF vs SWASH KAF)
“NFC” issues

- Not exactly issue with NFC
- Example: U+06C7 vs. U+0648, U+064F
- Perhaps could be caught by “confusables” algorithms?

Arabic team
Recommendations

• Whenever there is a variant, all resulting labels are available to the applicant
• It is up to the applicant which ones to activate

Arabic team
Focus on Chinese Language

- Reports in principle about “script”, but report primarily about Chinese
- Some consideration of effects on Japanese and Korean
RFC 3743, experience

- Experience at other levels of DNS
- RFC 3743 a good fit for CJK use
Two fundamental cases

- Traditional vs Simplified
- Variation due to Source Separation Rule (e.g. U+6237 versus U+6236)
Focus on reducing confusion

• Mainly interested in confusion of strings between languages
• Unlike Chinese and Arabic, no strong recommendation that “everything works”
Different from other cases

- Many more languages than some other scripts
- Extremely fraught political environment:
  - Cyrillic vs. Latin
  - Cyrillic vs. Arabic
  - Many spelling & character reforms
One language can cause issues

- Substitutions in one language obliterate differences in others
- E.g. U+0435 vs U+0451, U+0433 vs U+0491
- Some characters not on keyboards
Interaction with other scripts

Cyrillic team

- Issue of relation to Greek and Latin raised
- Declared out of scope, but problematic
Very different issues

- Confusing similarity a high priority issue
- Especially worried about URL bar display
- Concern about ill-formed akshars
Environment issues

- Display of Devanagari script can be problematic
  - Rendering engines
  - Fonts
ZWJ and ZWNJ

- Some Devanagari-using languages rely on ZWJ
  - Even if there is a precomposed version that will do
- ZWNJ needed for noun paradigms
  - Use in TLDs not clear
Inter-script issues

Devanagari team

• Relationship between Devanagari and other Bramhi-derived scripts?
• Ruled out of scope, but may be important
Unusual case

Greek team

• Greek alone in studied scripts in being used for only one language
Additional restrictions

- Team recommends excluding ancient characters
- Team recommends sticking to Monotonic characters

Greek team
Sigma and Tonos

• IDNA2003 maps upper case to lower case: Tonos can be lost
• IDNA2003 maps away final form sigma
• Transformations in applications in IDNA2008

Greek team
Final sigma

Greek team

- Recommend registering final form sigmas wherever requested
- Also register without the final sigma (i.e. with small sigma in place of final sigma)
Greek team

- Recommend registering with Tonos where requested
- Also register with Tonos stripped
Dimotiki and Katharevousa

- Recommendation that, if Katharevousa string is requested, the “same” Dimotiki “word” is blocked
- Only report that requests variant behaviour because of whole-string meaning
The impossible dream

- There are too many relationships among characters in Latin-using languages
- There’s no way to decide
- Therefore, no variants

Latin team
Remember, please comment

Open until 14 November

http://www.icann.org/en/public-comment/