

Internationalized Domain Names Technology & Status Review

John C Klensin
APTLD Manila
23 February 2009

Goals

- ◆ Provide mechanism for non-ASCII domain name labels
- ◆ Without disrupting
 - ◆ DNS operations
 - ◆ Existing applications
- ◆ Still obey other DNS constraints
 - ◆ Length limits on labels and FQDNs
 - ◆ Severe limits on punctuation and special characters
 - ◆ Language-based mnemonics, not “words”

Getting Here

- ◆ Original standards published in 2003
- ◆ Fairly widely implemented and deployed in many zones (registries)
- ◆ Clear by early 2005 that there were some issues, both conceptual and protocol
 - ◆ Upgrade path for new versions of Unicode and new scripts
 - ◆ Confusing terminology
 - ◆ Opportunities for mischief
 - ◆ Misunderstandings of DNS

Current Status

- ◆ Revision effort underway
 - ◆ Mostly tuning
 - ◆ Largely invisible to careful existing applications
 - ◆ Some changes still being debated
 - ◆ Especially IDNA2003 compatibility for strings in files
 - ◆ Nature of conversion/ migration advice
 - ◆ Lookup restrictions (unassigned code points)
- ◆ This presentation talks about 2003 standards using 2008/9 terminology

The Mechanism

- ◆ Introduce reserved form of labels
 - ◆ “xn--” prefix
 - ◆ All labels with “--” in third and fourth positions reserved in IDNA-aware applications and zones
 - ◆ Unicode characters encoded with special algorithm (“Punycode”)
- ◆ Every label containing non-ASCII characters has two forms
 - ◆ U-label: Native characters
 - ◆ A-label: Encoded form with prefix
 - ◆ May also have mapped equivalent forms (IDNA2003)

The Mechanism

- ◆ Only A-labels go into DNS zones
- ◆ Interpretation, coding, and decoding are application matters, not DNS ones
- ◆ Non-IDNA-aware applications or implementations
 - ◆ Just see the A-labels
 - ◆ Understand them as ordinary hostname-format (“LDH”) labels

Why a Revision

- A number of issues with IDNA2003
- Some very important to specific communities even though less so to others
- Others have general impact
- Goal is to keep and improve balance between
 - Mnemonics for as many languages/ scripts as possible
 - Identifiers that are
 - Safe and reliable
 - Easy to use
 - Well-behaved and more predictable

Some Key Issues I

- Some important characters and scripts excluded from IDNA2003
 - To a user of a script, it is important
 - To a user of a language, the entire orthography is important
 - Even though the DNS still cannot always represent words, much less sentences
 - Need to balance different uses and global accessibility
 - Globally-dangerous characters
- Some characters *not* excluded – back to LDH-extended!

Some Key Issues II

- Some scripts perceived as badly handled for mnemonic purposes
- Confusing terminology has made problems worse.

Some Key Goals

- Get a standard that is
 - Unicode version agnostic
 - Easier to understand
 - More predictable with regard to what happens
 - More adaptable to local conditions – realistic interoperability, not just theory
 - Deals with issues discussed earlier to extent possible

DNS Internationalization

- Not just IDNA
- Many areas of responsibility to make things work well
 - Standard Protocol
 - Registries and Registry Restrictions
 - Registrants – Increasing or decreasing confusion
 - Lookup implementors and Applications
 - Common sense by users and others
- Still mnemonics, not literature
- Can't solve confusion, but can provide better tools

New Terminology

- Much confusion in IDNA2003
- Standard is about labels, not FQDNs...
 - Still some Right-to-left (bidi) questions
 - Hard issues with label separators
- Native Character Label Form
 - U-label (if valid)
 - Status issue (outside IDNA Standard)
 - Status *in* URIs
 - Status *of* IRIs
- For protocol, “registry” = “zone”

Reclassification of ASCII Labels

- LDH label – Traditional definition, ASCII
 - IDN Reserved LDH Label (Hyphens in 3rd and 4th positions)
 - XN label form (“xn--”)
 - A-labels (must be valid)
 - Fake A-labels
 - (invalid)
 - Non-reserved LDH labels (everything else)
- Non-reserved LDH-labels
 - Different from Non-LDH-labels

Warning: this is still brand-new and may change again

Summary of Rationale

- Trying to move ahead to make IDNs far more useful and predictable as mnemonics
- Improve understanding, usability, marketability
- No big-jump revisions in future
- Still can't write literature

Registration Under IDNA2008

- Labels registered are final strings
 - No more mapping in protocol
 - Registration must be based on A-labels and validate them
- Explicit requirement to develop policies to keep identifiers safe and reliable
 - May want to think more about character exclusions and variants
 - Labels that violate standards will probably not be looked up by applications
- Registries and registrars share responsibility for keeping DNS useful
 - In ASCII DNS, protocol didn't prevent registering unusable strings

Some Scripts Used for Many Languages

- Uses are often different
 - Conventions
 - Rendering
 - Specific characters
- May require tradeoffs among languages using script
 - Need for cooperation within script community
 - JET work for CJK – introduced “variants”
 - Current Arabic Script IDN effort

Summary of Proposed Changes

- ◆ The differences
 - ◆ Clarifying terminology
 - ◆ Unicode version independence – new scripts
 - ◆ Return to letter-digit-hyphen concept (character inclusion)
 - ◆ Elimination of all mapping from protocol
 - ◆ Implications still *very* controversial
 - ◆ Correction of 'bidi' bug
 - ◆ Contextually-valid characters
 - ◆ Most invalid labels will not be looked up

Implications for Registries

- ◆ More explicit responsibility to validate
 - ◆ No more “guess what I mean”
 - ◆ Many doing that already
- ◆ New characters
 - ◆ May require special consideration relative to existing labels and/or confusing label pairs
 - ◆ Potential for much better handling of a few scripts, including some uses of Arabic and most Indic scripts (but contextual rules)
- ◆ Inclusion list
 - ◆ No more punctuation, symbols, compatibility characters

More Reading

- ◆ New Definitions
 - ◆ <https://datatracker.ietf.org/drafts/draft-ietf-idnabis-defs/>
- ◆ Rationale, context, registry advice
 - ◆ <https://datatracker.ietf.org/drafts/draft-ietf-idnabis-rationale/>
- ◆ Actual protocol definition
 - ◆ <https://datatracker.ietf.org/drafts/draft-ietf-idnabis-protocol/>
- ◆ Handling of right-to-left characters
 - ◆ <https://datatracker.ietf.org/drafts/draft-ietf-idnabis-bidi/>
- ◆ Permitted character definitions and tables
 - ◆ <https://datatracker.ietf.org/drafts/draft-ietf-idnabis-tables/>