IDNA Protocol Status Review

John C Klensin
6 December 2006
About This….

• Not a tutorial, but…
  – Review of developments about the IDNA protocol
  – Guesses about the future

• Some terminology assumed from prior tutorials, including Sunday
Standard Developed Under Stress

• Design model
  – No changes to DNS itself
  – Processing of Unicode to DNS-compatible form
  – Normalization to deal with different Unicode structural forms
  – Actual DNS entry is an LDH-compatible ACE “punycode”
  – Strong ties to Unicode 3.2

• Assumptions
  – Universal implementation
  – Users would never see punycode
  – All Unicode characters unless a reason ("exclusion")
  – Extensive Unicode mappings
  – No major “confusion” problems
  – Could specify a version of Unicode

• All Wrong
Where are We Today?

• Experience of last few years
  – Confusing character pairs
  – User and registrant confusion about mappings
• Growing sense of risk when IDNs are used outside of “native” context
• Implementers making their own rules to protect users
  – Punycode if bad combinations (Mozilla, Opera, Safari,…)
  – Punycode if not user-configured script (IE7)
  – Net: Lots of punycode
• Users don’t see it as an improvement.
Issues Identification: The IAB “nextsteps” Report

• RFC4690 –Highlights
• Many User Expectations Not Realistic
  – Language-dependent matching
  – Preventing mixed language writing systems in a label
  – Complete cure for confusable characters
  – Fully-comfortable and culturally appropriate solution for mixed R-to-L and L-to-R strings
  – Ability to use Internet without any Roman-based characters and without significant “presentation” work.
The Context of RFC 4690

• A few obvious recommendations
  – Unicode version-agile
    • Can’t stay stuck at 3.2
    • Can’t move to 5.0 and get stuck there
  – Inclusion list
  – Review and update protocol and tables
• Content is Issues, not proposals
• Not all problems identified can be solved
• IDNs won’t make the Internet multilingual, but they may be an important piece of a larger picture
Fairy Tales and IDNs

- Can’t use the Internet except in English
- If only we had top-level IDNs, my country would be well-connected
- Users use IDNs
  - URIs need protocol identifiers, only one can be default
  - http:// (or https: or ftp: or mailto:) and tail syntax are not going away.
- People can transcribe arbitrary characters from printed form.
More Unreasonable Expectations

- Any valid word in any language…
- Writing sentences (or novels)…
- Ability to mix scripts, especially related ones, without causing confusion and risk.
And Some Unrealistic DNS Ideas

• Two trees with translations or transliterations all the way down
• Homogeneity between labels of a fully-qualified domain name (FQDN)
• Different resolution methods based on TLD
Recommendations for Solutions

• Three components, all important
  – Protocol adjustments
  – Registration models and restrictions
  – Presentation

• There will always be risks (there are with LDH too, but many more characters)
First Component: Protocol

• Protocol
  – Applies to all levels of DNS and all domains
  – Otherwise, lose interoperability and global references

• Reformulate IDNA
  – Few substantive changes
  – Model more easy to understand and better tied to concepts
  – Separate sequences of steps for registration and lookup
  – Unlock new Unicode versions
Other Protocol-Related Changes

• Character list changes
  – Inclusion list with ability to add incrementally
  – Prohibit non-language characters
  – Remove mappings from protocol
    • If a character is mapped out under IDNA2003, prohibit
    • Some “prohibited” characters will become matters for local user interface mapping
    • Some necessary exceptions
  – Allow some things that IDNA2003 prohibits to permit a wider range of characters and scripts
    • Bidirectional improvements
    • Zero-width breaking and non-breaking spaces
• Not clear how to do some of this yet.
Efforts in Progress or Coming Soon

• IDNA Reformulation
  – Existing procedure conforms to new definition

• New tables
  – Characters permitted, prohibited, and pending
  – Joint work between IETF and Unicode Tech Committee groups

• Fix the bidirectional rules
  – Allow a larger range of languages
  – Clarify edge cases
Protocol Change Impact

• No fundamental change to algorithm
• No change to prefix
• Little effect on existing non-test registrations that conform to existing guidelines
• Some strings now prohibited by guidelines will be prohibited by protocol
• Ability to register more languages and more practical strings and names
Second Component: Registration

• Requirement: Anything requiring context
  – Language rules
  – Cultural limitations
  … Must be handled as restrictions on what can be registered
• Specifically…
  – Elimination of language and context-based confusion
  – “Variant” linked-registration rules
  – “Not found” must be ok… and preserved
• *But probably not effective below second level*
Third Component: Presentation and Input

• **What the user sees ... and types or utters**
• Issues include…
  – How much can we localize consistent with a global network?
  – When is it unsafe to display native characters?
  – What to do when characters cannot be displayed?
  – Input of user-unrecognized characters
  – Localized users who travel
  – Mixing Right to Left and Left to Right strings
  – How to see an invisible character

• **Ultimately not an IETF or ICANN decision**
Other Issues

• Balance among
  – Usability
  – Maximum localization
  – Requirements of a global Internet
  – DNS stability and referential integrity

• Some risks will remain

• Must be realistic about the problems IDNs can solve

• Leave the door open for DNS support of other navigation techniques
Finding Solutions

• This is a problem with
  – Many constraints, notably preserving DNS
    • Stable operation
    • Referential integrity
  – Many desires
    • Linguistically and orthographically correct representation of words in any language
    • Universal comprehensibility of all labels
    • Users never see punycode
• These goals *cannot* be completely realized and are not independent
• Tradeoffs must be balanced instead.
Personal Editorial

• IDNs are very important
  – For some limited, but critical, purposes
  – They will not, alone, make the Internet multilingual

• Community could kill IDNs by accident
  – Overreactions to risks
  – Too much punycode in front of users
  – Non-interoperable “better solutions”
  – Overwhelming serious work and design with
    • Agendas that use IDNs as a platform, not a DNS-related goal.
    • Discussions and Decisions based on passion mixed with extremes of ignorance

• We need to focus on those risks and avoid them.
Summary

• IDNs provide an opportunity
  – To make the Internet more accessible to many communities
  – To help with the important task of preserving cultures and languages

• But they also pose risks including
  – The risk of violating DNS constraints and ending with something that does not work
  – The risk of incompatible implementations that would cause names to mean different things in different places: especially difficult if the DNS is infrastructure for other navigational techniques
  – The risk of impeding innovation in other techniques
What Next?

Optimistic View

• IETF moves swiftly ahead on protocol adjustments
• ICANN takes risks seriously and
  – Invests in understanding
  – Makes decisions based on maximum IDN capability consistent with a DNS that is fully-functional for users.
  – Avoids decisions based on trying to satisfy those with unrealistic demands or simply an ability to register and retrieve names

Or We Risk

• Next-generation navigation that
  – Doesn’t work or
  – Doesn’t rely on DNS
• Having to discard the present DNS tree and start over.