Technical Issues with IDNs

John C Klensin, Ph.D.

ICANN Marrakech 2006.06.27

Issue Identification

- IAB Report …and this presentation…
  - Identify issues
  - Sometimes identify possibilities
  - Try to identify who should look at them
  - Do not propose solutions
- Some issues..
  - Do not have solutions other than education and awareness or
  - Getting the problem out of the DNS

Quick Terminology Review

- IDN: Internationalized Domain Name (“label”)
- Domain names consist of labels
- ISO 646
  - ASCII, ITU T.50 (IA5)
  - Upper and lower case undecorated Roman-derived alphabetics, digits, some specials
- URL: Uniform Resource Locator
  - The much more general and internationalized “IRI” form still contains ASCII syntax
- Examples mostly Roman-based for convenience

Technical and Other Issues

- IDNs have become a mix of
  - Technical issues in implementation
  - User interface and Internet Navigation
  - Cultural issues in evolving to a multilingual Internet
  - Issues in competition and profitability
  - Social, national, and political symbols
- This talk addresses only the first three, focusing on the first one.
- This is *not* a technical presentation

Presentation Drawn From…

- IAB report
  - "Review and Recommendations for Internationalized Domain Names (IDN)"
  - Approved for publication
  - http://www.ietf.org/internet-drafts/draft-iab-idn-nextstep
- Related internationalization and Internet Navigation work
- Personal impressions

IDNs are the Solution to a Problem

- What is the problem?
  - Better mnemonic value for names in non-Latin (undecorated Roman-based) scripts
- National pride and recognition???.

- Some other problems with no solution in IDNs
  - Content availability
  - Connectivity and Access
  - User-friendly URLs
  - Understanding each other's languages

The DNS Constraints (part 1)

- Exact match
  - No “close enough” or “do you mean” option
- Traditional Upper and Lower Case
  - Simple definition for ISO646BV: Exact correspondence and reversible
  - Case-sensitive storage and replies, case insensitive queries
- Characters, not
  - “names”, languages, or even scripts

The DNS Constraints (part 2)

- Harder to understand, but no less important
- More subtle issues
  - Strict administrative hierarchy
  - Inflexible aliasing (no “see also”)
- Technically complex and subtle, but important, e.g.,
  - “RR set consistency”

Names In the Real World

- Languages, dialects, and scripts are a complicated business
  - Relationships can be debated… passionately
  - Often no clear answers

Name and Character Matching

- Subjective Decisions
  - People are better at them than computers
  - Contemporary, rule-based, computer systems are better at them than the DNS.
  - DNS doesn't have enough information to even try most approaches.
- If "linguistic correctness" is the question…
  - IDNs are not the answer

IDNA

- Standard for encoding IDNs into DNS
  - Unicode mapped ("nameprep") Unicode
  - Nameprepped Unicode -> "Punycode" ACE
  - Punycode -> Nameprepped Unicode
- Somewhat over three years old
- Turned to be a little naïve in several ways
- ICANN policy statements and plans a little more so

Problems with Implementing IDNA

- …
- … *(none here)*
- …

- Has proven easy to implement and deploy … if policy issues are ignored
- Big problem is getting around to it
- Supported in production versions of all major Internet web browsers except one
- Little support in other applications so far – other work comes first.

Problems with *Using* IDNA and IDNs

- Character spoofing and similarities
  - Can't be "fixed" technically
  - Hard to design policies that help for many cases
  - Impossible to prevent all cases
- Transcription from written form
- Human expectations and DNS ones
  - Different
  - DNS much less flexible

Education about the Possible

- Do ø and ö match? Maß and Mass? oe and œ?
  - Users may think so… or not
  - Depends on language context and perhaps more
  - Not possible to get this right in DNS or coding
- Easier to match “color” and “colour”
  - DNS cannot do this either…
  - But there is a business opportunity

As Soon As Characters Get More Complicated than ISO 646IRV

- Case-matching becomes imprecise and requires tables
- Character list inevitably expands over time.
- Matching new and old characters, and new and old tables, is going to be version-sensitive.
- Some matching is in the eye of the beholder

Transcribing a URL

- In what domain does one look for
  - http://www.example.py/
  - Cyrillic names in Paraguay?
  - Note that “one script per label” does not fix this
- Does the following violate any important policy?
  - www. раурал.сом
  - Is that a large enough hint for the SLD? The TLD?

The Variant Model

- Within a given domain…
  - Collect labels that contain similar characters
  - Register one, block others or
  - All must belong to the same registrant
- “Similar” is registry-defined… might be
  - Appearance
  - Meaning
  - Sound
  - Etc.

Variant System Status

- Strongly developed for CJK
  - Obvious applications for decorated Roman-based characters
  - Other applications across scripts
- No impact on queries

Perception that Policies are not Protective Enough

- Leads to reactions from software writers
- Those reactions will
  - Attach warnings to names perceived of as risky --or--
  - Render risky names in “punycode” form, defeating the value of IDNs --or--
  - Do other creative things
- Definition of “risky” will differ by vendor

Separate Matching Trees Do Not

- Genetic Variation
- Populating one tree with translations of another
  - Might almost work
  - “Almost” == “Unpredictable”
- But mean separate zone files at 3rd level
  - Very difficult to keep synchronized…
  - Especially with different labels

Consistency and Astonishment

- Different implementation choices about what to support
  - Leads to different behavior as seen by user.
  - If some behavior is inconsistent, registrant and user will be unable to predict
  - They won't be happy
- Violations of the Law of Least Astonishment

Unicode Normalization and IDNs

- Main protection against problems with different ways to code characters
  - “Normalize” to a single form
  - Normalization rules are designed for stability.
- IDNs have other issues
  - IDNA/Nameprep are a superset of the normalization used
  - Unnormalized strings are permitted and persist so some normalization-stability rules do not apply

Nameprep Stability Across Unicode Versions

- If Nameprep is not stable – strictly upward-compatible
  - Migrating from one version of Unicode to another is hard
  - Some methods require versioning in the DNS
  - New prefix??
- If cannot migrate
  - No recently-coded scripts as IDNs

What Next – IETF Issues

- IDNA review
  - More restrictive Nameprep … less mapping?
  - Codepoint review… fewer charaters accepted?
  - Upgrading to match versions of Unicode?
  - DNS-based IDNA versioning or script labeling?
- Recognize DNS Limits during these tasks

DNS Limitations

- IDNs will not solve URL problems
  - Structure ASCII keywords
  - Long and complex tail syntax
- IDNs do not address “near match”
  - “Near match” may be the only real solution to similar characters
- Rigid administrative hierarchy
  - Limits “similar tree” ideas
- Solutions lie “above DNS”, not in it

Characters and Security

- Any issue with confusable characters or surprising matching…
  - Probably has greater impact on security and certificates than on the DNS
  - When DNS names are used to establish identifier locales, any problems multiply

The Costs of Change

- Making changes has consequences
  - May invalidate now-valid names
  - Any prefix change would require software changes and careful study
- When is the price too high?
  - If the price is not trivial, may require broad community consultation
  - Users who are hurt by *not* making the change must be considered

Some Key ICANN Issues

- New kinds of disputes and dispute resolution issues
- Decisions by registries imply registry responsibility
  - Technically, each registry can have different policies about permitted names (within IDNA scope)
  - Some restrictions might make things easier for everyone

The IDN TLD Issue

- Naming and Delegating Decisions
  - Not as easy as seem to be believed
  - If some decisions are made, others may be impossible in practice
- Multiple Labels for “the same” TLD
  - Real aliases and their implications
  - Attempts at replicating or translating trees
- Coding and Presentation Questions
  - May not exist… or require IETF interaction

Next Steps

- Reduction of permitted character list - consider
  - Remove non-language characters
  - Remove word separators
- Update to Unicode 5.0
- Reexamine non-DNS and above-DNS approaches
- Examine “whois” again

Summary

- This isn't easy
- We got it a little bit wrong the first time
- We need to get it fixed *before* deployment is broader
- "We" will require IETF and ICANN
  - To work together
  - Not just toss demands based on assumptions about how thing work over the wall