

# Internationalization and the Internet

John Klensin

Patrik Fältström (Faltstrom?)

??? ??? (Kenny Huang)

# The Problem and the Topic

- Internationalized Domain Names
  - are not the problem
  - might be part of the solution
- The problem is how to make the Internet fully international, with as little “English bias” as possible
- We will return frequently to this distinction

# Internationalized Domain Names (IDN)

- Term used in many ways
  - Strictly, domain name labels that represent names containing non-”host name” characters.
  - Only “host name” (or “LDH”) strings are actually entered into the DNS.
  - Sometimes, “IDN” is used to refer to a fully-qualified domain name that contains at least one non-LDN label/
- Sometimes used to refer to other ways of internationalization or localization
  - “Keywords”,
  - Special searching or directory mechanisms, etc.

# Internationalization and Users

- Users typically do not want internationalization (or “multilingual” capability) but...
- Systems that are “localized”: adapted to their particular
  - Language
  - Writing system and character codes
  - Location
  - Interests
- Internationalization is
  - A means to localization
  - Necessary given the global nature of the Internet

# Why is there a problem?

- Many have suggested “just put non-ASCII names in”
- Three big issues
  - Local solutions and global interoperability
  - Flexibility and safety
  - Unicode issues and alternatives
- First two impact almost every major Internet policy decision.

# Local solutions and global interoperability

- Tension between
  - Every Country/ Company/ Person makes own decisions independently and does its own thing.
  - A major strength of the network is the ability to smoothly interoperate globally, permitting the next generation of innovations.
- Both together are often possible
  - end to end principle permits more independent decision-making than previous network technologies
  - Still often a tricky and complex balance to accomplish this.
  - Simple and obvious solutions can be a global disaster

# Flexibility and safety

- Often another tradeoff between
  - Maximum freedom to interpret protocols in different ways and
  - Stability and/or security of the network

# Unicode issues and alternatives

- Several decisions made in designing Unicode make it non-optimal for DNS (and Resource Identifier) applications
- All of the alternatives are worse.

# Representing Unicode/ ISO10646

- No tagging equals no national character sets
  - Unlike applications (such as the web), no room in DNS for character set tagging, so a comprehensive, “universal” character set –UCS – is a requirement
  - Poor experience with stateful switching of character codings
- More characters, mixing scripts
  - Many opportunities for problems from look-alikes that were not present in ASCII alone

# Internationalization, IDNs, and the Problem Being Solved

- Letting people access information and the Internet in natural languages and scripts
  - The problem?
  - Yes, unless, maybe, one is a greedy participant in the “domain names market”... maximization of confusion and FUD.
- What is broken and needs fixing?

# The Problem: What is not working adequately?

- Individual domain name labels?
- The periods / full stops in x.? . .? ?
- Protocol-name strings such as “http” or “mailto”
- Special characters in, e.g., URIs ?  
/ : ? = # ...
- Or email  
@ % ! ...
- Left-to-right elements are natural in some cultures, right-to-left in others

# Confusion and Fraud

- Most of the problems are with us already with ASCII, weak software, and bad habits
- “Do no harm” may be another important principle: supplying guns and bullets to criminals is rarely a good idea.

# The eBay/ Credit Card Scam

Date: Sun, 09 May 2004 01:06:19 +0200 (CST)  
To: jck@jck.com  
Subject: Your eBay Account Must Be Confirmed  
From: Support <support@ebay.com>

Update Your Credit / Debit Card On Your eBay File [Image: "spacer"]

Dear eBay member ,

During our regular and verification of the accounts we couldn't verify your current information, either your information Has changed or it is incomplete . if the account is not updated to current information within 5 days then , your access to Buy or Sell on eBay will be restricted

Go to the link below to Update your account information :

<http://signin.ebay.com/aw-cgi/eBayISAPI.dll?SignIn&ssPageName=h:h:sin:US>

please dont reply to this email as you will not receive a response

Thank You for using eBay!

<http://www.eBay.com>

- Link appears to be <http://signin.ebay.com/aw-cgi/eBayISAPI.dll?SignIn&ssPageName=h:h:sin:US>
- But it is really <http://61.100.12.150/verify/index.php>

# What does that have to do with IDNs?

- That one is very easy to detect (by careful people or software)
- But consider the potential for  
`http://? ? ? .COM/`
- Are you sure you know what that is?

# What does that have to do with IDNs?

- That one is very easy to detect (by careful people or software)
- But consider the potential for  
`http://? ? ? .PL/`  
in lower case, it would be  
`http://aß?.pl/`  
that obviously is not `http://abh.pl/`, but the link will be consistent with the display.

# Variations for most scripts

- Internally
  - 1 l (1 L) / 0 O (zero)
- Between related scripts
  - All, or almost all, contemporary alphabetic scripts have a common origin; character similarities are inevitable
  - “USA”, “pectopan”
- The Chinese Problem(s)

# What is the DNS to be used for?

- Tension between
  - Network-facing identifier
  - User-facing “name” (of a company, product, organization,...)
- Constraints on solutions
  - Short label strings – no reasonable way to tag
  - Uniqueness of names
  - Potential for confusion or fraud
- Requirement for non-ASCII names is clear but
  - Caution is in order – many possible traps and risks
  - Hard to go back if too permissive

# Reminder about where the DNS cannot help

- Internationalization is really a “multilingual” problem, not just “multiscript”
- Local matching rules needed
- Searching capabilities –not just exact match lookups – needed
- Attribute structures – language, location, entry or business type – needed

# And the DNS's Constraints

- Nameprep would be more workable with “yes/no/maybe”
  - But the DNS is only “yes” or “no” – no hints
- Localized systems tend to fragment network
- Translation and transliteration are important sometimes
  - Simplified and Traditional Chinese
  - Kanji and Kana
  - Vowels or not
  - British and American

By now,  
You should be at least a little bit  
frightened

So, how did we get here and what  
do we do?

# History

- **Hostnames and ISO 646 Basic Version**
- **Content internationalization - web & mime**

# Internationalization and the Internet

- Consideration given to “international characters” in the 1970s
  - Character set standards weren’t ready
- Project that led to MIME
  - “multimedia email” capability
  - Initiated largely to standardize and permit non-ASCII characters
- Web
  - Recognized requirement early
  - Details only for Western European languages until mid-90s
- All were done by “tagging”
  - Tagging is consistent with localization approaches

# Applications & International Characters

- Most Internet application protocols defined for ASCII, or at least seven-bit characters
  - Often not an accident or ignorance – consider use of IA4 and IA5 in many ITU Recommendations
- Waiting for applications to be upgraded could
  - Be a long wait
  - Involve some unpredictability with sender not knowing receiver capabilities
- Plug-ins and patches do not yield a consistent user experience

# **IDN efforts**

- **Just use UTF8 or 8859-N or GB??? or...**
- **Tagging problem w/ DNS**
- **The IDNA Approach**
  - **Name format no one uses.**
  - **Efficient for script-homogeneous strings**

# **Description of IDNA**

# Problems Internal to IDNA

# Nameprep Issues

- **Eliminates/normalizes some lookalikes & font forms**
- **Try to preserve case-mapping rule**
- **Cannot be completely successful partially due to characters shared among scripts**

# Unicode Complications

- **Unified CJK**
- **Separate European**
- **Font-specific chars**
  
- **IDNA helps with some of this, but not much**

# Traditional and Simplified Chinese

- **Characters with semantics**
- **Relationship to case mapping**
- **Cannot process Kanji and get SC**

# The Character Variant Model

- **JET: registry restrictions, variants, and reserved strings**
  - Adoption in CJK ccTLDs
    - No actual variants, yet, in two of them.
  - Analogies to alphabetic languages
- **The ICANN Guideline**
  - Language base
  - Registration of tables
- **Implementations and issues**

# Variant Roman Character Example

- Suppose we have two people with surnames  
Müller and Quinoñes
- And they have historically registered the obvious  
ASCII domain labels  
Mueller and Quinones
- Now, when IDN registrations are permitted,  
should others be permitted to register the IDNs  
with the correct spellings, or should those names  
be reserved? If not, how is the restriction  
managed?

# The Meaning of “Language”

- JET, ICANN, etc., use the term “language” to describe tables and rules.
- Not the normal usage
- Really Zone-Language-Script
  - No one really knows what the limits of a “language” are, although governments can make decisions within their territories.
  - “Scripts” actually overlap in strange ways. Neither Unicode Consortium nor ISO have been able to rigorously define scripts associated with particular languages
  - E.g., for some zones in Western Europe the appropriate language-script has been “generic European”, i.e., “Latin-1”. For others, more specific lists of characters may be needed.

# Major Issues with variant models

- “Multilingual” strings
- Labels and “names”
- Variant charging in JET-like models
  - Cost of a reserved label
  - Cost of activation given that the label has no value to anyone else
- DNS as an administrative hierarchy
- New types of conflict/ dispute problems

# Technical Interoperability

- IDNA is entirely a client algorithm and procedure, hence depends on correct client implementations and is hard to verify.
- JET Guidelines and similar approaches are registry-dependent
  - They do not raise interoperability issues.
  - May raise user experience ones

# Administrative Hierarchy Issues

- Policy and trust relationships
- No cross-tree cross-references to branches of hierarchy
- Maintaining parallel trees
  - Workable if really identical and have a single coordinating database.
- Organizational branding
  - <http://www.product.tld/> or
  - <http://www.organization.tld/product>

# New Dispute and Resolution Issues

- ICANN-WIPO UDRP assumes
  - Homogeneous scripts and language characters
  - Conflicts about rights to identical names
- but not...
  - Labels constructed from line or box-drawing characters
  - Look-alike characters and strings from different scripts unless they meet trademark criteria for “confusingly similar”
  - Translations, transcriptions, transcodings
- Is the relevant “name” the IDNA encoding or its display/presentation form?

# Problems IDNs Don't Solve

- Registration policy issues
  - “This language is more important”
  - The gTLD problem
- Applications and local character sets
- Even JET Guidelines won't eliminate confusion
- DNS is a poor “search” mechanism... and getting worse.

# The Whois Policy Issues

- Registration in non-ASCII and data in ???
- Searching of a multilingual/ multiscript database
- Reading the records
- Information about variants and IDN  
Package contents

# Competition and Policy

- Policy tradeoff between
  - More flexibility of registrations
  - Less risk of conflicts, deception, or fraud
- Each domain or zone will need to develop its own policy, and there will probably be wide variations.
- Implications of a country deciding to go its own way with, e.g., local character codings.
- User-exposed punycode between people using very different scripts is probably forever.

# What was that Problem Again?

- **Domain-name guessing is becoming less useful**
  - Effectiveness reduced with more names
  - Effectiveness reduced with more possibly-relevant TLDs
- **Guessing in a multiple script (“multilingual”) environment will be *much* harder.**

# The Application Interface Problem and Unicode

- Windows, Internet Explorer, Outlook, and...
  - Winsock and UTF-8 conversion of UTF-8
  - Localized versions with local character codings and different behavior
- Better if you have a Mac
- Maybe better if you have a Unix or Linux system

# Global Interoperability Again

- Giving up the ideas of
  - Any two Internet users being able to communicate, regardless of language
  - Any Internet user being able to access any public hostwould make many of these problems much easier, but...
- It would be a high price to pay.

For some of us...

This is where  
“being frightened”  
gives way to  
“being depressed”

# And we still have not solved the problem

- If IDNs are this hard  
and do not solve the problem  
– and slogans do not solve it either
- Maybe it is time to go back to the problem  
and do some serious thinking about models  
and approaches.

# Questions for Thought

- Several studies indicate that search engine use is rising rapidly and even replacing name-guessing in some areas. Does that suggest opportunities?
- Can we get past the marketing hype, scaling problems, and need for a name-conflict “judge” and take another look at alternate naming systems with fewer constraints about characters and cross-references than the DNS?
- Is it time to look again at “yellow pages”-like systems, perhaps with the multihierarchical structure of contemporary classification systems, as an alternative to both the DNS and search engines for some purposes?

# (More) Questions for Thought

- Are IDNs of primary importance for communication within a country or language rather than between them? Can we accept the use of Roman-based characters – or even ASCII or IA4 – between language groups?
- Should we be giving serious consideration to translation of DNS names in applications in addition to IDNA mapping to and from DNS names in those applications?
- If IDNA had been designed with knowledge of the registry restriction and variant models, would its mappings and restrictions be the same? If not, is it too late to fix?

# Summary

- From a technical/ protocol standpoint, IDNA is ready to deploy today and being deployed.
- But it is essentially a coding standard, not a “solution”.
- The interesting issues and opportunities are best found by examining the user experience at the application interface: putting names in the DNS and getting them out is easy and always has been.
- It may be time to think about “non-DNS” or “above-DNS” approaches that really do solve the problems.