"Everyone has the right...
to seek, receive and impart information and ideas through any media regardless of frontiers"

-- Universal Declaration of Human Rights

There is no universal "right to language". But there are human rights with an implicit linguistic content that multilingual states must acknowledge in order to comply with their international obligations under such instruments as the International Covenant on Civil and Political Rights.

– UNDP, *Human Development Report 2004: Cultural liberty in today's diverse world*, New York: United Nations Development Programme, 2004, page 60.

كأره انترنت انتارابعسا

(عالميه‌الغة الانترنت)

John C Klensin

டான் டஇன் வீ

庄振宏

С големия принос на

Patrik Fältström

黃勝雄

Internationalization and the Internet

John C Klensin

Tan Tin Wee

James Seng

with major contributions from

Patrik Faltstrom

Kenny Huang

كأره انترنت انتارابعسا
(عالميه‌الغة الانترنت)

John C Klensin

டான் டீன் வீ

庄振宏

С големия принос на

Patrik Fältström

黃勝雄

Internationalization and the Internet

John C Klensin

Tan Tin Wee

James Seng

with major contributions from

Patrik Faltstrom

Kenny Huang

Internationalization and the Internet

An Internet Society Tutorial

Preface – Design of the ARPANET and Internet

Two Communities

- Technical challenges and sharing of computer resources
- Facilities for expanding human communications and knowledge

Global Accessibility and Global Interoperability

- Tutorial is about questions and decisions, not answers
- Many easy answers for internationalization for a isolated, homogeneous population – but all of them (so far) tend to fragment the net and impede global communication
- The global solutions all involve policy tradeoffs with no clear “correct” answers

Goals for the Tutorial

- Examine IDNs in the general contexts of
  - internationalization and localization
  - navigation on the Internet
- Describe the "physics" of the environment: properties of the DNS or the Internet generally that constrain solutions
- Identify some key policy issues and the associated tradeoffs: the afternoon session may *begin* the process of developing comprehensive policy.

Internationalization and the Internet

- Outline
  - Internationalization and other general issues
  - History of IDN efforts and context
  - The IDNA standards and their implications
  - Policy issues, tradeoffs, and unsolved problems
- Order of material
  - Everything is connected to everything else, so we will, of necessity, talk generally about some topics and then come back and define them properly.

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" or even "we have a solution for our language, why should anyone else care?"
- 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
  - Each Culture/Country/ Company/ Person makes its own decisions independently and does things its way.
  - 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

But accomplishing both…

Requires that we work together in good faith and with due respect for each other and for the many linguistic and cultural differences that these problems involve.

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
- Even where it is possibly optimal, it may be
  - Inconsistent with familiar coding methods
  - Inconsistent internally
- But… all of the alternatives are worse.

Characters and Character Sets

- In a “character set” coded for information processing use, fairly abstract characters are assigned “code points”
  - Essentially, characters are grouped, ordered, and then numbered
  - “Glyphs” – the form of the characters – are rarely standardized

Scripts and Languages

- A “script” is an (often poorly-defined) collection of related characters
  - It is common for several languages to share most, but not all, characters from a given script
  - Scripts are often given the same name as one of the languages that uses them, creating much confusion.
    - Cyrillic script, but Russian, Ukrainian, … languages
    - Arabic script, but Arabic, Farsi, Urdu,… languages
- Unicode consortium gives script names and language bindings (UTR 24), but precision is very low

Languages and Countries

- People migrate and take languages with them
- Most languages are used in many countries, not just those where they are dominant or "official"
- Over enough time, most languages evolve differently in different locations

A Content Problem

- Even when we can use tagging and the rules are well-specified there can be unexpected large difficulties
- “Please type out your name in Chinese characters and send it to me” is not a simple request with a simple response today.

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 for global DNS use
  - 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.Б.ג.ئ ?
- 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

The Problem: Deployment

- The Internet is not just the world wide web and its “http” and “https” protocols
- For content, the web and email share descriptive structure – cannot change one without affecting the other
- For any development that requires changing something that already exists, it takes a long time to deploy new, fully-compatible application software.

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://ABH.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://ABH.PL/

  in lower case, it would be http://αβη.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

   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

The DNS and “languages”

- DNS labels are
  - traditionally just arbitrary strings of permitted characters
  - not “words” or language elements except by convention
- IDNA simply expands the range of permitted characters
- Requirement for non-ASCII strings 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

DNS Constraints

- Name lookup would be more workable with "yes/no/nearly/maybe"
  - But the DNS is only "yes" or "no" – no hints
- Localized systems tend to fragment network
- Character translation and transliteration are important sometimes (or not)
  - Simplified and Traditional Chinese
  - Kanji and Kana – Vowels or not
  - British and American – Typographic conventions

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

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

Ancient Network History

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

MIME is a system for structuring and identifying content other than simple ASCII text – multimedia, national character sets, applications structures.

Internationalization and the Internet

- Consideration given to “international characters” in the 1970s
  - Character set standards weren’t ready other than
  - “National use” positions in what became ISO 646
- Project that led to MIME
  - “multimedia email” capability
  - Initiated largely to standardize and permit non-ASCII characters

Internationalization Developments

- 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

Alternatives to Upgrading Applications

- Plug-ins and patches do not yield a consistent user experience
  - One user to the next
  - One application to the nextb
- Looking at “punycode”:

Changing

  - from a miserable, but memorable, transliteration to a
  - incomprehensible and ugly code

is not an improvement

History of IDN Efforts

Description of IDNA