“*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 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

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
  - risk factors
- 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-IDN 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
  - Every Culture/Country/ Company/ Person makes their own decisions independently and does things their 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

IDNs May Expand Old Risks

- Many characters can be confused with others
  - Problem even exists in ASCII
    - Digit “1” and lower-case “l”
    - Digit “0” and upper-case “O”
  - IDNs increasing the character collection
    - From 64 in ASCII (LDH)
    - To tens of thousands in Unicode (nameprep)
- “Confusables”
  - Create opportunities for user mistakes
  - and fraud

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
  - Artistic fonts are ignored

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).
  - Precision has been very low but improving with recent versions

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

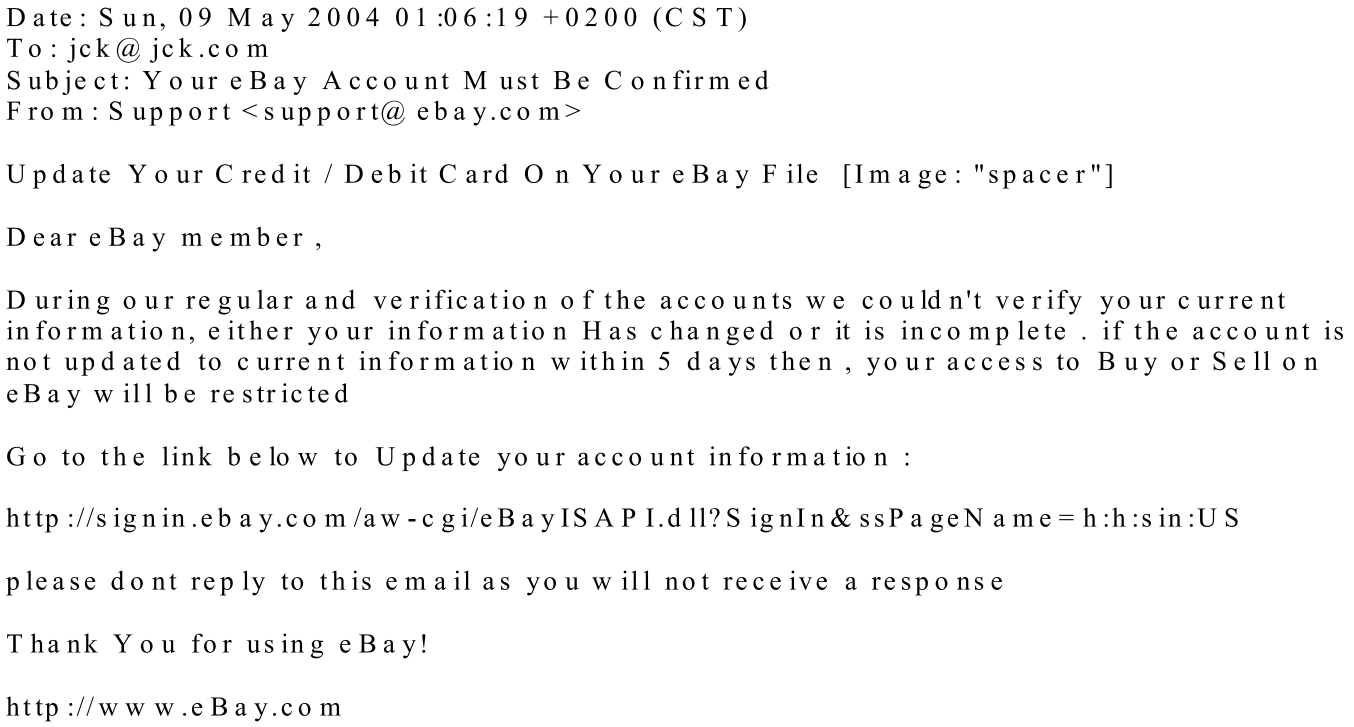
The Problem: Deployment

- 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
- Generally, 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.

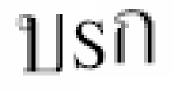
Variations for most scripts

- 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

  
  ปรก 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
  - British and American
  - Vowels or not
  - 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

Recent Experience and the "Phishing" Problem

- Article warning about risk of "look alike" characters in IDNs
  - Both web navigation and
  - SSL/TLS certificates
- The attack is old news
- But browser vendors are reacting by
  - Disabling IDNs unless known to be safe
  - Making safety decisions on a TLD basis

IDNA and How It Works

- Current Internet Standard for IDNs
  - IDNA (RFC 3940)
  - Internationalized Domain Names for Applications
- Plus tables to define and map characters
  - Nameprep (RFC 3941) … a profile of
  - Stringprep (RFC 3454)
- Details and tables being reviewed after phishing discussion – might be revised somewhat.
- IDNA and Nameprep
  - Involve *no* change to DNS itself
  - DNS still operates on ASCII LDH, names

IDNA Registration Process

- Applicant supplies label to be registered
  - Punycode or “native script” form, or both
- Optionally checked for
  - Locally-prohibited characters or words
- Processed through IDNA
- Optionally checked for “variants” and variant labels generated.
- Checked for conflicts with existing names
- Only internal (“punycode”) form entered in DNS.

IDNA Lookup Process

- Application applies IDNA’s “ToASCII”operation
  - Checking
  - Processing through nameprep (identical to registration)
  - Creation of punycode
- Punycode is then looked up in the DNS as if it were a conventional DNS name

Application Receiving an IDN

- If in punycode
  - Convert to “native script” Unicode with IDNA
  - Display if possible and safe
- If in “native script”
  - May need to convert to punycode and back to validate
  - Conversion may not produce the same string
  - Mappings imply that
    - ToUnicode(ToASCII(string)) may not equal the string
    - Examples
      - Toys-Я-Us (if permitted by the registry) would become toys-я-us
      - “Schloß” would become “schloss”

IDNA: Things to Note

- Standard encoding and mapping tables worldwide
  - Without this, the DNS will not work consistently
- All restrictions on registration of names, whether imposed by
  - Culture (no offensive terms)
  - Restrictions on mixed scripts or languages
  - Restrictions on registration of look-alike or related characters (at all or by different parties)

  just prevent names from being registered
- On lookup, there is no difference between
  - No entry due to no one having registered a particular name
  - No entry due to prohibitions on registering some names

  so different restrictions in different registries is a safe technique

Reprise: Early technical IDN approaches

- **“Just use” UTF-8 or 8859-N or GB2312, or Big5, or KOI-8, or…**
- **Tagging problem w/ DNS**
- **The IDNA Approach**
  - **Name format no one uses.**
  - **Efficient for script-homogeneous strings (UTF-7 and UTF-8 are not, especially for East Asian characters)**

Some DNS Physics

- DNS performance depends critically on caching “near” the site of the query
- Consistent and predictable DNS operations depends on caching only complete RR sets
- All known-possible methods for guaranteeing integrity of DNS data, including DNSSec, are quite sensitive to non-conforming handling of queries and responses.
- “Trick servers” are, to at least some extent, a problem for each of these.

Problems Internal to IDNA and Issues It Does Not Address

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 or languages but used differently
- Unavoidably does one-way mappings badly (e.g., a German IDN may be registered with ä, ö, or ü, but not ß)
- Important to understand that these properties are the result of tradeoffs – the alternatives are worse.
- Review underway now as to whether too many mappings have been permitted and whether non-language characters should be entirely excluded.

Applications Issues

- Email addresses
  - Local-parts more important than domain-part?
  - DNS advantage with LDH
  - Unrestricted local-parts, so ACE-like encoding cannot be completely safe
  - Envelope – header (transport) issues
- URL definition
  - Strict ASCII
  - IRI proposal and http://…
  - Status of IRIs

Traditional DNS: What Goes In, Comes Out

- Case-insensitive mapping
  - If “A” is registered, a query for “a” matches, but returns “A”.
- With IDNA,
  - “Ü” can be looked up, but not registered
  - If “ü” is registered, but the query is for “Ü”, the query will match, but “ü” will be returned.

depending on the application, this difference may result in some user astonishment.

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 Simplified Chinese**

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

Dispute Resolution or Conflict Prevention

- Key principles
- Character variants and other evolving systems: prevention of conflicting/ confusing registrations
- Dispute resolution policies and mechanisms: “register first, then straighten it out”

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, IETF, ICANN, etc., use the term “language” to describe tables and rules.

this is *not* the normal usage

The Meaning of “Language”

- Really Zone-Language-Script
  - No one really knows what the limits of a “language” are, although governments can make decisions within their territories.
  - “Scripts” overlap in strange ways. Neither Unicode Consortium nor ISO have been able to rigorously define scripts associated with particular languages (there are some broad, descriptive, definitions)
  - For example., 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.

Authoritative Policies about Scripts

International Bodies: Consensus about Language

- Authority
    - National sovereignty issue for ccTLDs
    - Rules generally cannot be enforced below level two or three (similar to trademarks)
    - International issue for gTLDs
- Scripts and Languages
    - If one script is used by several languages, language authority is not sufficient

Authoritative Policies about Languages

- If a good-quality recommendation is available, will registries use it?
  - Foolish not to: saves a lot of work, trouble, and looking silly
  - Compulsion is another matter
- Multiple-language scripts can be a major gTLD challenge
- So can multiple-script languages

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. That makes it 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-like 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 all confusion, just some of it
- 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
- Windows may get fixed, but not this year

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 host, using a globally-available name

  would 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”
will rapidly give way to
“being depressed”

The Cure for that Depression

Working cooperatively with each other to both

- internationalize and
- preserve global interoperability

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?

(More) Questions for Thought

- 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?
- 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?

(Still more) Questions for Thought

- Should we be giving serious consideration to inter-language 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?

Major Issues We Have Barely Touched

- Email addresses
- Names and domains in digital certificates
- A fully internationalized alternative to the URL or URI
- Special problems with “multilingual” TLD names
- Hundreds or thousands of other protocols and how to internationalize applications that use them
- Finding and navigating to resources with non-ASCII names
- User interface issues

Summary – The Protocol Foundation

- From a technical/ protocol standpoint, IDNA is ready to deploy today and being deployed.
- IDNA is ultimately rooted in Unicode, which can represent, in some plausible way, almost every character in contemporary use for writing a language in today's world.
- Where Unicode does not properly represent a the characters needed for a language, efforts should be initiated to remedy that problem.
- IDNA is essentially a coding standard, not a "solution".

Summary – The Policy Challenge

- 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.
- Avoiding or dealing with confusion and name conflicts will require a good deal of thought.
- Whatever is done, must be done with great sensitivity to cultures and traditions
- It may be time to think about "non-DNS" or "above-DNS" approaches that really do solve the problems.

Internationalization of the Internet

A Great Opportunity

A Great Risk of Fragmentation

and a Great Challenge for all of us.

Internationalization of the Internet

A Great Opportunity

A Great Risk of Fragmentation

and a Great Challenge for all of us.

Copy of these slides

- PowerPoint versions:
  - http://ws.edu.isoc.org/workshops/2004/ICANN-KL/
    - ICANN-ISOC-KL-IDN.ppt
    - ICANN-ISOC-KL-IDN-part2.ppt
- PDF will be in the same directory in a few days, as will the other sets of slides

References and Additional Reading

- The IDNA Standard: Faltstrom, P., Hoffman, P. and A. Costello, "Internationalizing Domain Names in Applications (IDNA)", RFC 3490, March 2003; Hoffman, P. and M. Blanchet, "Nameprep: A Stringprep Profile for Internationalized Domain Names (IDN)", RFC 3491, March 2003; Costello, A., "Punycode: A Bootstring encoding of Unicode for Internationalized Domain Names in Applications (IDNA)", RFC 3492, March 2003.
- Variant names and registry restrictions: Konishi, K., Huang, K., Qian, H. and Y. Ko, "Joint Engineering Team (JET) Guidelines for Internationalized Domain Names (IDN) Registration and Administration for Chinese, Japanese, and Korean", RFC 3743, April 2004; Klensin, J., “Registration of Internationalized Domain Names: Overview and Method”, work in progress, July 2004 version is draft-klensin-reg-guidelines-04.txt.
- Uses and abuses of the DNS: Klensin, J., "Role of the Domain Name System (DNS), RFC 3467, February 2003.
- A different view of the non-ASCII TLD issue: Klensin, J., “National and Local Characters in DNS TLD Names”, work in progress, Niv 2004 version is draft-klensin-idn-tld-04.txt. Also ISOC Member Briefing