Remote Participation

• Jabber room is open:
  – IDNQUESTIONS@jabber.icann.org
  – Frank Fowlie will manage questions posted to the room
Agenda

• IDN General Information
  – Definition
  – IDN Status Quo Overview
  – The Need for IDNs
  – Internationalization
  – Protocol and Functionality
  – Punycode, stored form vs. displayed form
  – Languages and scripts
  – Unicode and ASCII

• Confusible IDN Issues
  – Same script different language
  – Same language multiple and mixed scripts
  – Visual confusables

• IDN Program Plan

• Sao Paulo Activities

• Summary
What is an IDN?

• IDN stands for Internationalized Domain Name
  – Domain name labels containing non-host name characters.
    • Valid hostname characters are: a-z, 0-9, “-”
    • Valid hostname characters sometimes referred to as ASCII or LDH
  – Only host name strings are entered into the DNS
  – IDN in general refers to both displayed form (Unicode) and stored form (punycode) of the domain name

• Example: rødgrød.tld → xn--rdgrd-vuad.tld
  – ø is LATIN SMALL LETTER ø WITH STROKE: U+00F8
  – Used in for example Danish, Norwegian, Faroese
Domain Names in General

• Domain names are not general natural language expressions
• Domain names that are not lexically words in a language are possible and quite common
• Domain names are identifiers that help users uniquely reference information in the Internet using sequence of characters into strings
• Domain names must be unique
• Not all words in all languages will be available as domain name labels
Internationalization Overview

Domain Names Based on ASCII / LDH Rule

- IDN second level
- Internationalized top level

ASCII based browser/email clients/…

- Application upgrades to get web access in local chars + IDN enabled emails…

Content have been available in many languages for some time

- Expected to continue to expand

example.test → 실례.test and 실례.테스트

(stored form: example.test → xn--9n2bp8q.test and xn--9n2bp8q.xn--9t4b11yi5a)

Aim: An internationalized Internet
Internationalization cont.

• Internationalization of the internet means that the internet is equally accessible from all languages and scripts

• Domain names represent only a small part of internationalization of the internet

• Controversy about how important the domain names are compared to search capabilities…etc…
  – Accessibility from all languages is important which means that the way IDNs are handled is very important
  – Continuously making characters available as much as possible as these are added to Unicode
  – Disagreement about whether domain names are used by typing into browsers and usability of IDNs
    • But agreement that email addresses based on local characters are necessary for large parts of the world,
    • and URL’s listed in offline documents need to be usable by local communities
The Need for IDNs and Internationalization

• Geographic expansion of the Internet
  – IDNs match needs of increased use by linguistic groups
  – IDNs used for identification of content reflecting linguistic diversity

• Internationalization is
  – A means to localization
  – Necessary given the global nature of the Internet

• Localized system adapted to
  – Language
  – Writing system and character codes
  – Location
  – Interests

• Global Interoperability
  – Network strength is to interoperate globally
  – Security and stability is primary focus
  – Avoid fragmentation of the Internet
IDNA – Protocol Functionality

- Domain Name Resolution Process:

1. User types in 실례.test in for example browser
2. 실례.test gets converted to codepoint
3. Case-folding and normalization
4. Stringprep filter
5. Punycode conversion → xn--9n2bp8q.test
More Protocol Information

- IDNA is the acronym for the IDN protocol, developed within the IETF and published in June 2003.
- IDNA stands for Internationalized Domain Names in Application.
- Technical details are available in the IETF RFCs: RFCs 3490, 3491, and 3492.
- IDNA is currently under revision:
  - RFC4690 and associated internet drafts suggesting revisions and solutions to some problems.
  - More about this later…
Displayed Form vs. Stored Form

• Historically the domain name you register is also the domain names stored and usable in the DNS
• This is changed with introduction of IDNs
• Usually the stored form does not make any meaning
  – Example: .tld → xn--mgbtbg2evaoi.tld
• However, there are exceptions:
  – xn--gibberish - decodes into the Arabic characters ﺖ٧٩٨فﺎﻟ١
  – xn--trademark - with different versions of trademarks
  – This is coincidentally and hence not intentionally

• xn-- prefix specifically designates a system called Punycode
• xn-- prefix indicates to application software that the label needs to be decoded back into Unicode for proper display to the user
More Punycode and Some User Perspective

• Intention that Punycode (xn--...) never be exposed to users, but there are exceptions
  – situations where IDNs could not be displayed as Unicode characters
  – in such cases the utility of IDN depends on user recognition and understanding of Punycode

• Otherwise, as a user all you need is the name you want to register
  – TLD Registries will supply a list over available characters, usually in Unicode
  – Registries will handle all encodings needed during registration process

• May be useful to consider usability of the name, keyboards, business cards, and other practical limitations

• Encodings by for example:
  – http://josefsson.org/idn.php
  – Others are made available by TLD registries
Language and Script

• Languages are used by humans to interact
  – Best guesses estimate 5000-7000 languages worldwide, of which 100-200 are mainly used
  – RFC3066 discusses languages in more detail
  – Examples: Arabic, Greek, Portuguese

• Script is a set of graphic characters used for the written form of one or more languages (ISO10646 definition)
  – Examples: Arabic, Cyrillic, Greek, Han

• Computers don’t understand languages instead any characters will have an associated code-point
Unicode and ASCII

• Unicode is one of many character encoding systems in use.
  – Encoding systems are lists that assign a unique number to each character in the list
• Unicode accommodate a Universal Character Set and contains different ways for representing characters
  – Not all is adequate for handling IDNs partly due to variations in language and user perceptions
  – http://www.unicode.org, technical reports UTR36 and UTR39, and more details in RFC4690
• The DNS uses a different encoding system, ACE is an ASCII Compatible Encoding
  – American Standard Code for Information Interchange
  – Punycode (the xn- - form) is the ACE used for IDNs
• This is what we saw before with the displayed form in Unicode and the stored form in Punycode (ASCII)
How far did we make it....

• IDN General Information
  – Definition
  – IDN Status Quo Overview
  – The Need for IDNs
  – Internationalization
  – Protocol and Functionality
  – Punycode, stored form vs. displayed form
  – Languages and scripts
  – Unicode and ASCII

• Confusable IDN Issues
  – Same script different language
  – Same language multiple and mixed scripts
  – Visual confusables

• IDN Program Plan
• Sao Paulo Activities
• Summary
Same Script Different Language Issue

• Language specific character issues
  – Jorgen = Jørgen = Jörgen in Danish, Swedish, Norwegian
  – But users don’t always think that o equal ø and ö
    – ø is LATIN SMALL LETTER o WITH STROKE (U+00F8)
    – ö is 'LATIN SMALL LETTER o WITH DIAERESIS' (U+00D6)

• Not possible to make generic rule at the protocol level
• Need for specific rules at TLD registry level

• Some registries have submitted character tables to the IANA repository to show variants
  – Example: the .se table displays that:
    • The letter Ü is referred to in Swedish as a # "German Y" and is # considered to be a variant of the letter Y.
    • The letter Å is not considered to be a variant of the letter A...Earlier practice substituted AA, which is no longer recommended but will still be encountered

• http://www.iana.org
  – (link to IANA Repository at bottom left of main page)
Same Language Multiple Scripts Issues

- Some languages can be expressed by multiple scripts
  - Eastern European and Central Asian languages can be expressed in Cyrillic or Latin characters
  - African and Southeast Asian languages can be expressed in Arabic or Latin characters
  - Other languages are written in a combination of scripts- Kanji, Kana, Romanji for Japanese & Hangul and Hanji for Korean

- Hence, same word, same language can be expressed in different ways
  - Some words can only be expressed use a single script
  - Some words are expressed by mixing of scripts

- Result is that script definition is very important and sensitive in terms of IDNs
Visual Confusion Issues

• Well-known example: paypal.com
  – Second character is U+0430, Cyrillic small a
  – Looks like Roman/ASCII “a”
  – This is now prevented by “one label, one script” rule per the IDN Guidelines with exceptions for mixed script languages

• Other example:
  – Russian ccTLD is .ru
    • Cyrillic “r” and “u” is: p and y
    • Which looks like p y (in latin) is ccTLD for Paraguay
    • **Note: Russia did not ask for .py, this is just an example**
  – Process needed to determine labels matching ccTLDs
General Overview of User Confusion Issues

- IDNs Expanding Risk of Known Problems
- Many characters can be confused with others
  - Problem exists in ASCII as well
    - 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
- This kind of confusion
  - create opportunities for user mistakes
  - and fraud
Mid-way Summary

We have looked at some of the main issues related to IDNS – what about solutions…

Some user confusion is being solved by
- protocol adjustments
- IDN guidelines revisions
- implementation of adequate registry policies

Remaining user confusion need to be solved by
- education of community
IDN Program Plan

• A new program within ICANN
  – IDN Program recently established within ICANN to achieve the possibility to insert internationalized top level labels in the root zone.

• IDN dedicated staff
  – Existing Technical, Policy, IANA Staff
  – New positions of CTO, Writer, Project Coordinator, etc

• Goals with program includes
  – Enable introduction of internationalized top level labels
  – Response to increased geographic use of the internet
  – Global interoperability and keeping the internet secure and stable
Towards Introduction of Internationalized TLDs

• The Program Plan is comprised of several Projects that may be planned and managed separately but have independencies.

• Projects focuses on following objectives:
  – Security and Stability of the DNS
  – Results and recommendations from the IETF’s Review of IDNA
  – Promoting consumer choice and avoiding user confusion
  – Developing consensus policy to guide implementation
  – Increasing Outreach and communication plans
IDN Laboratory Testing Goals

- Demonstrate that the insertion of IDN strings into the root has no appreciable negative impact on existing resolutions

- Obtain agreement of US DoC that internationalized top level labels can be inserted (potentially initially for test purposes) live in the DNS

- Reach consensus opinion with RSSAC and the root-ops that internationalized top level labels can be inserted (potentially initially for test purposes) in the DNS
IDN Laboratory Testing: Project Milestones

- **July 2006:**
  - Meeting with IDN-PAC and root-server operators during Marrakesh and Montreal meetings
  - Plan NS and DNAME testing as two parallel running tracks

- **September 2006:**
  - ICANN retained Autonomica to perform laboratory test
    - Highly DNS experienced staff
    - Test plans will be made publicly available for replication opportunities

- **October 2006**
  - IDN-PAC agrees on method to select the strings for the laboratory test
  - Set of strings are provided Autonomica and initial testing are commenced
    - Preliminary tests already performed and while successful, demonstrated that some applications have not implemented IDNA in accordance with the existing protocol standard

- **December 2006**
  - More test details expected to be provided
IDN Laboratory Testing Details

• Autonomica will develop and ICANN will publish the test procedure
  – plan detail will be sufficient so that others may replicate the test
  – ICANN will publish the results received of any other test performed in accordance with the publish test plan

• The laboratory test plans includes the following:
  – insertion of NS records into a copy of the root zone
  – tests performed in closed laboratory environment with a series of systems implemented to replicate as closely as possible the server software of the various root servers. This includes:
    • versions of BIND server software, and
    • use of the most popular DNS resolver software packages

• No further end-user or application testing is included as the laboratory environment is closed and not accessible from outside
Development of Laboratory test strings

• Test strings was delivered by ICANN as coordinated through the IDN-PAC

• Normal Unicode-Punycode conversion
  – flod18häst → .xn--flod18hst-12a

• Performance with a 63-character long TLD string
  – .hippo18potamusshipippo18potamusshipippo18potamusshipippo18po

• Right to left, embedded characters with opposing directional properties

• Left to right script with sophisticated shaping properties

• Non-alphabetic script
First IDN Test Complete

- First IDN Test Run successfully completed in October 2006

- 63 letter Top Level Domain
- Conducted at the .museum IDN lab, in association with Autonomica

- Preliminary Results
  - Resolver software in test environment worked without problems
  - End-user software showed difference that was not related to implementation of the IDNA protocol, and is currently being corrected
Application Software Testing

• A positive result from the laboratory tests will allow move to a “Live” IDN TLD test

• These additional tests are intended to ensure that application software will work with internationalized domain names
  - Introduce <.test> in various scripts to ensure participant understanding that this is for testing only
  - Test scripts are intended to be determined after consultation with Internet community
  - Plans will be main topic for IDN-PAC meeting in Sao Paulo
  - Plans will need further discussion with technical community
IDNA Protocol Revision,

By IETF
Proposed Revisions to IDNA Protocol

- Revising the IDNA protocol will build an “inclusion” based model for determining what scripts may be used for IDNs and potentially increase the number of scripts available for IDN deployment.

- The revision will base the protocol on Unicode 5.0 (containing 64 scripts), the existing protocol is based on Unicode 3.2 (containing 45 scripts).

- The revision to the protocol will:
  - Potentially increase available blocks of characters
  - Include revision process to include additional scripts in the future
  - Include technical review of protocol functionality

- The revision effort is being managed through the IAB/IETF

- The Basic Framework was published Sept-06
  - RFC4690
Revisions suggestions of IDNA Protocol

• Three internet-drafts were published providing suggestions for solutions to the issues raised in RFC4690:
  
  • An overview with proposed issues and changes for IDNA
  
  • A suggestion for solving an IDNA problem in right-to-left scripts by revising the stringprep profile
  
  • An overview of suggested inclusion based IDNA Unicode Codepoints based on Unicode 5.0
  
• A status report will be provided in the IDN workshop
  – Wednesday, 6 December 2006, 17.30-19.30
Communication and Outreach
IDN Outreach and Communication Focus

• ICANN regional road-trip in Middle East, October 2006
  – Arabic script vs. language issues
• Internet Days Forum
  – Stockholm 24-26 October 2006
• Internet Governance Forum, IDN workshop
  – Athens 31 October 2006
• APTLD meeting with IDN focus
  – 14 November 2006
• ccTLD meeting for Middle East
  – Dubai 20 November 2006
• Sao Paolo ICANN Meeting with IDN sessions
  – 2-8 December 2006

• RSS feed available for IDN Communications
• Online Calendar for IDN events available
Sao Paulo Activities

- GNSO IDN working group
- IDN Tutorial
- IDN workshop
  - Statuses on various projects
- GNSO, ccNSO joint work
- GAC IDN working group
- IDN Presidents Advisory Committee
- Additional other meetings…
- Resources:
  - http://www.icann.org/topics/idn
Summary of IDN Principles

• Global uniqueness and interoperability of the DNS
  – unique and unambiguous domain names
  – Same functionality regardless of geographic placement of access
  – URLs and emails connect as expected regardless of geographic placement of access

• Promote “Future-Proof” solutions
  – Define Unicode characters to be allowed
  – Provides ability for adding new languages, new characters far in the future

• Avoid or diminish as much as possible user confusion
  – Technical limitations
  – Implementation requirements
  – Registry restricted list and policies
  – User education

• Promote multi-stakeholder involvement