Email addresses and domain names are *non*-latin! Now what?

Jim DeLaHunt / IUC44 / 14 October 2020

Universal Acceptance

Internationalized domain names, email addresses

+1,000,000,000

The next one billion internet users use a wide variety of languages and scripts. Standards allow email addresses, and domain names, in scripts they can easily read. This is an introduction to those standards.

To: مانيش@ أشوكا. الهند

http:// 普遍接受 - 测试。世界

To: données@fußballplatz.technology

Abstract

Email addresses, and domain names, are no longer limited to ASCII Latin script. They can now be

http:// 普遍接受 - 测试。世界 or مانيش@أشوكا.الـهند or données@fußballplatz.technology .

Software, frameworks, and workflows will need to change to accommodate. What are Internationalized Domain Names (IDN) and Email Address Internationalization (EAI)? What do you need to know? What do you do next? This tutorial brings you up to speed. It explains IDN and EAI. It shows you the implications. It connects you to sources of information. It helps you understand what this will mean for you. Suitable for software developers, QA, marketers, system administrators, and management.

Agenda

Slides: http://go.jdlh.com/iuc44t3t3 (links in slides)

* Who we are: UASG, Jim DeLaHunt
* Context: the next one billion, and previous domain names, email
* What's new: so many top-level domain names, Internationalized Domain Names (IDNs), Email Address Internationalization (EAI)
* Benefits of, Issues with IDNs, EAI
* More resources
* Next steps
* Q&A

Who we are

Who we are

Universal Acceptance Steering Group (UASG)

* http://www.uasg.tech
* Community-led initiative, world-wide
* Raise awareness, identify problems, solve them
* Project of ICANN, the domain name system organisation

Jim DeLaHunt

* http://jdlh.com, ☎ +1-604-376-8953
* Vancouver, Canada
* Consultant in multilingual websites; software engineer
* UASG volunteer participant

UASG materials available

UASG operates primarily by public education. Participants write outreach materials, technical notes. They give presentations to industry meetings. They evaluate, report, and follow up on UA issue reports.

Technical Notes (selection)

* UASG004 Use Cases for UA Readiness Evaluation
* UASG010 Quick Guide to Linkification
* UASG018 Programming Languages Evaluation Criteria

Plus C-level outreach papers, magazine articles, presentations, ...

Who you are

This talk is a tutorial for those who know email addresses and Internet domain names primarily as ASCII-only. We introduce internationalised domain names (IDNs) and email addresses (EAI). Software development skills helpful for some advanced material, to which we link.

Primary audience

* Users of domain names and email addresses, technically inquisitive
* Application developers handling domain name and email addresses
* Dev, QA, marketers, system administrators, and management

Context

The next 1,000,000,000 Internet users

5th (next) billion

China, India, Third World.
Large share use non-Latin script.
Little marginal North American, European increase.
Mostly mobile and small-screen, lower share on desktop, laptop.
Extending to mid-, lower-educated, less comfortable with Latin script.

VS

First 1 billion

First world, N. America, Europe.
Large share use Latin script.
Includes large share of North American, European potential.
Mostly desktop & laptop computers, mobile only later.
Cream of highly-educated in each market, the best at Latin script

Domain names

Domain names are the primary way to locate things on the internet. Original standards limited domain names an ASCII subset, and thus to Latin script. This obstructs users of non-Latin languages. Names aren't just on-line (see: ads), or written (see: saying a domain name)

Domain name standards

* ASCII Letters, Digits, and Hyphen, max 63 (RFC1035)
* Well known Top-Level Domains: .com, .org. .net, .jp, .ru, .cn, .in, ...
* e.g. Amazon.com, XgenPlus.com,
* Appear in many areas, e.g. email addresses, URLs, billboards, speech

Domain names, extended

Recent changes permit Internationalized Domain Names for Apps (IDNA). This allows new non-Latin TLDs, and non-Latin characters in rest of name. Parallel changes permit Latin TLDs with more than three characters. Thousands have been registered.

Domain name extensions

* Internationalized Domain Names for Apps “IDNA2008” (RFC5890)
  * Replaces earlier IDNA2003
  * e.g. http:// 普遍接受 - 测试。世界
* .भारत (“bharat”, India), . 中国 (China), 「。」 as well as ‘.’
* .tech, .museum, and hundreds more

Email addresses

Still a mainstay of Internet communication. Actually a stack of related specifications, including SMTP, POP3, IMAP, *etc*. Original standards limited email addresses to an ASCII subset, and thus to Latin script. This obstructs users with names from non-Latin-script languages.

Email standards

* Subset of ASCII, typically letters, digits, punctuation (RFC2822)
* *mailbox* @ *domain.name*, e.g. info@unicode.org
* *mailbox* preferably similar to user's own name in own script
* Many implementations, some deviating from standards

Email addresses, extended

Domain name extensions brings change to the domain.name part of email addresses. Extensions to email address syntax permit almost any Unicode character in mailbox. Consequences ripple through SMTP, MIME, IMAP, POP3, and more.

Email Address Internationalization (EAI) standards

* EAI Overview and Framework (RFC6530) + 6 more RFCs
* EAI requires changes to several protocols and components
* Change takes time, so EAI must interoperate with legacy email

What’s new: So many top-level domain names!

The older, simpler top-level domain names

The top-level domain name is the part after the final '.' Until 2001, there used to be a small set of 3-letter generic top-level domains, plus 2-letter country code top-level domains. They all consisted of latin letters.

Top-level domains, up to 2001

* generic: com, edu, gov, mil, org
* country-code, 2-letter: e.g. .ca, .uk, .eu
  * Based on ISO 3166-1 standard, with supplements
* Latin script, letters only

Top-level domains today

Resource

* http://data.iana.org/TLD/tlds-alpha-by-domain.txt
* Consider analysing with spreadsheet or Python code.

Top-level domains today

Resource

* http://data.iana.org/TLD/tlds-alpha-by-domain.txt (excerpt below)
* Consider analysing with spreadsheet or Python code.

Version 2020101200…
AAA
AARP …
BZH
CA
CAB …
COM …
NORTHWESTERNMUTUAL …
XN--CLCHC0EA0B2G2A9GCD …
XN--VERMGENSBERATER-CTB
XN--VERMGENSBERATUNG-PWB …
ZUERICH
ZW

Top-level domains today

Resource

* http://data.iana.org/TLD/tlds-alpha-by-domain.txt
* Consider analysing with spreadsheet or Perl/Python code.

Questions:

* How many top-level domain names (TLDs) now?
* How many begin with “XN--” prefix? How many don’t?
* What is the longest TLD not having “XN--” prefix?
* How many 3-character TLDs are there now?
* How many TLDs not having “XN--” prefix include digits or ‘-’?

Top-level domains today

Resource

* http://data.iana.org/TLD/tlds-alpha-by-domain.txt
* Consider analysing with spreadsheet or Perl/Python code.

Answers:

* How many top-level domain names (TLDs) now? A: **1507**
* # Begin with “XN--” prefix? How many don’t? A: **153** (11%), **1354**
* Longest TLD not having “XN--” prefix? A: **NORTHWESTERNMUTUAL**
* How many 3-character TLDs are there now? A: **223**
* How many TLDs not having “XN--” prefix include digits or ‘-’? A: **0**

Internationalized Domain Names (IDNs)

IDN: Unicode names, LDH infrastructure

The Domain Name System was designed to permit only Letters, Digits, and Hyphens (LDH). It was reliable, but so important, change was cautious. When internationalising, rather than add more characters to the DNS, they…

http:// 普遍接受测试。世界

IDN: Unicode names, LDH infrastructure

...mapped Unicode characters to LDH.

- Internationalized Domain Name for Applications (IDNA)
- 普遍接受 : U-Label
- NamePrep to normalise
- Punycode to separate non-ASCII and map to LDH

- Prefix 'XN--'
- xn--uorr18ad6bbt1e: A-Label
- e.g.
  http:// 普遍接受测试。世界
  http://xn--uorr18ad6bbt1e.xn--rhqv96g

IDNA U-Labels, A-Labels, NR-LDH labels

Domain names are separated by period '.' into labels. A label using anything outside Letters, Digits, and Hyphen (LDH) is a U-Label. The IDNA algorithm converts to a corresponding A-Label made of LDH. The familiar LDH labels are "NR-LDH".

DNS and IDNA "labels"

* e.g. www.uasg.tech has three labels: "www", "uasg", and "tech"
* NR-LDH labels: must not start or end with "-", LDH only, max length 63
* A-Labels: LDH labels, start with "xn--", valid Punycode output
* U-Labels: Unicode string from reversing Punycode on A-Label
* A-Label ← IDNA (Nameprep, Punycode) algorithm → U-Label

Example U-Labels, A-Labels, NR-LDH labels

U-Label, A-Label pairs

* 中国 ⇔ xn--fiqs8s
* भारत ⇔ xn--h2brj9c
* résumé ⇔ xn--rsum-bpad
* après-ski ⇔ xn--aprs-ski-30a

NR-LDH labels

* com, gov, ca
* unicodeconference, iuc44
* apres-ski

What are these?

* munchen, münchen
* museum
* xn-trik-bpad, xn--trik-bpad

Try it!

* https://eai.xgenplus.com/Multilanguage-To-Punycode-Convertor.jsp

IDN uptake

* 8.3m IDNs registered (January 2020)
  * 2.3% of approx 366m total domain names registered
  * 60% IDNs under Latin-script TLDs e.g. .com, .xyz
  * 40% under IDN top-level domain names
* Examples
  * .中国 .中國 .中信 .网址 (Chinese), .рус, .рф (Russian),
  * Sources: DomainTools, IDNworldReport.eu, ntldstats.com.

Learn the IDNA reference knowledge

Learn about Internationalized Domain Names for Applications, understand Nameprep and Punycode, know when to use U-Labels and A-Labels. Look for libraries to take over some of this for you.

IDNA references

* RFC5890 IDNA: Definitions and Document Framework
* RFC5891 IDNA: Protocol
* RFC5892 The Unicode Code Points and IDNA
* RFC3492 Punycode: A Bootstring encoding of Unicode for IDNA
    * etc....

Email Address Internationalization (EAI)

Email Addresses, Internationalised

* Structure: **อีเมลทดสอบ@ยูเอทดสอบ.ไทย**
  → Mailbox ‘@’ →Domain Name
* Right-to-left: تجربة-بريد-الكتروني@تجربة-القبول-الشامل.موريتاني
  Domain Name ← ‘@’ Mailbox ←
* A-labels: **อีเมลทดสอบ@**xn--l3cfk3a5bpd5gxc.xn--o3cw4h

Email Addresses, Internationalised

* Domain name part (after ‘@’)
  * May be IDN, A-Labels, NR-LDH, any combination
* Mailbox part (before ‘@’)
  * Interpretation is determined by email server
  * There are issues, and best practices

Learn the EAI reference knowledge

Learn about Email Address Internationalization. There are many formal standards covering email messages, SMTP, IMAP, etc. RFC6530 is the root of this tree. UASG012 is a good, technical starting point.

EAI references

* UASG006 Standards, RFCs Related to Universal Acceptance
* UASG012 EAI: A Technical Overview
* RFC6530 Overview and Framework for Internationalized Email
* etc....

Benefits of IDNs and EAI

Universal Acceptance (UA)

* “the state where all valid domain names and email addresses are accepted, validated, stored, processed and displayed correctly and consistently by all Internet- enabled applications, devices and systems.”
* i.e., IDNs, EAI just work

Benefits to internet user for IDN, EAI

* Easier to access websites, people
  * Familiar language and spelling of addresses
  * Less confusion from transcription to Latin script
* Positive impression from familiarity
* Next 1 billion users different from 1st billion
  * Less comfort with foreign languages (e.g. English)

Benefits to website/organisation for IDN, EAI

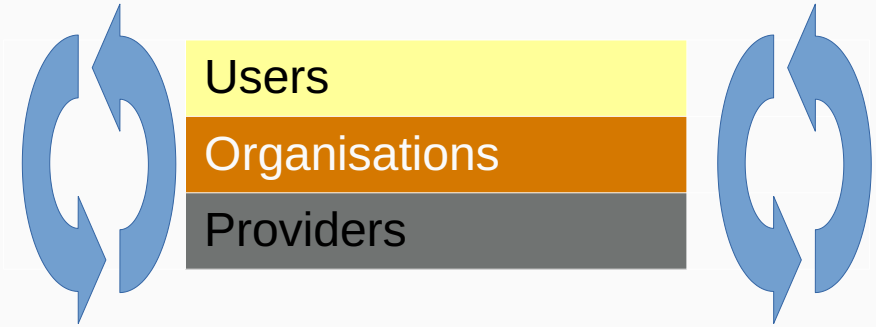
* Easier for their customers/users
* Competitive advantage (better service)
* Opens new markets
  * IDN, EAI as part of push into new markets
  * New IDN top-level domains are new market space

Benefits to software providers for IDN/EAI features

* Larger addressable market
* Universal acceptance has “bug fix” cost (relative to other features)
* Competitive advantage vs non-UA providers

Total benefit US $9.8 billon

* Virtuous cycle
  * Providers support IDN/EAI
  * Orgs use IDN/EAI
  * Users benefit from IDN/EAI



* Total benefit: US $9.8 billon/year
  * From new and existing users
  * Source: *Unleashing the Power of All Domains*, white paper by Analysys Mason for UASG, 2016

Example: “Kai sells chicken eggs” in Thai

* UASG case study
  * Video “**KaiKaiKaiKai Dot Thai – ใครขายไข่ไก่.ไทย**”
  * By Thai NIC, operators of .**ไทย** top level IDN
* Funny transliteration of **ใคร** (*Khır*, a name), **ขาย** (*khāy*, sells) **ไข่ไก่** (*khỳ kị̀*, chicken eggs).
  * **ใครขายไข่ไก่.ไทย** vs **kai-kai-kai-kai.com**

IDN and EAI as localisation tool

* Go to new market in other country, already
  * Localise product, its text, numbers, calendars, etc.
  * Set up in-country sales, support
* Consider also using IDN, EAI for better in-country presence

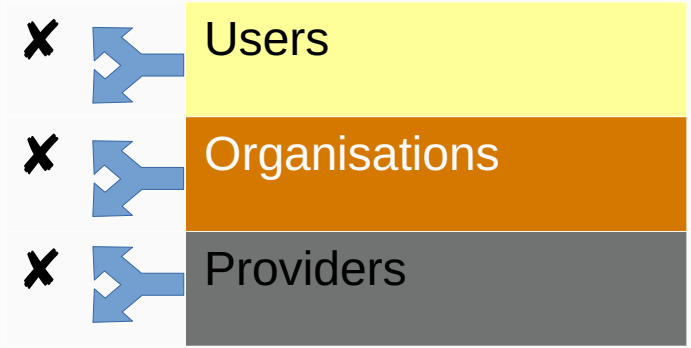
Services supporting EAI

* Hosting EAI mailboxes
  * XgenPlus (Datamail), Coremail
* Reading EAI mailboxes hosted elsewhere
  * Microsoft, Apple iOS 14, GMail.com

Issues with IDNs and EAI

Issue with adopting IDNs, EAI: supply-demand paradox

* Vicious cycle
  * Orgs see no demand from users
  * Providers see no demand from orgs
  * Users see no IDN/EAIs to use



* “The Chicken and Egg problem”

Issues with adopting IDNs

* Training customers, staff about IDN
* Compatibility problems in your products
  * Do they universally accept IDNs, EAI?
* Compatibility problems at your customers
  * Do their tools universally accept your IDNs, EAI?

Typical compatibility problems with IDNs

* Rejecting URLs based on limited syntax
  * Any regular expression testing URL is probably wrong
* Use good libraries for domain names, URLs
  * Normalization, U-Label ⇔ A-Label, live DNS lookup
  * UASG.tech has technical advice for developers
* Linkification (auto-convert URL-like string)

Issue with IDNs: fear of phishing

* Confusable characters: e.g. paypal.com
  * (U+0430 Cyrillic Small Letter A, not U+0061 Latin Small Letter A)
* This concern is overblown
  * Most phishing uses other attacks, e.g. paypal-com.bad.xyz)
* Many resources, e.g. Unicode Security Issues FAQ

EAI challenges: email as identifier

* Where email address is used as identifier, are EAI addresses accepted?
  * Any regular expression validating email address is probably wrong
* Test login process, user profile fields
* One person may have multiple addresses

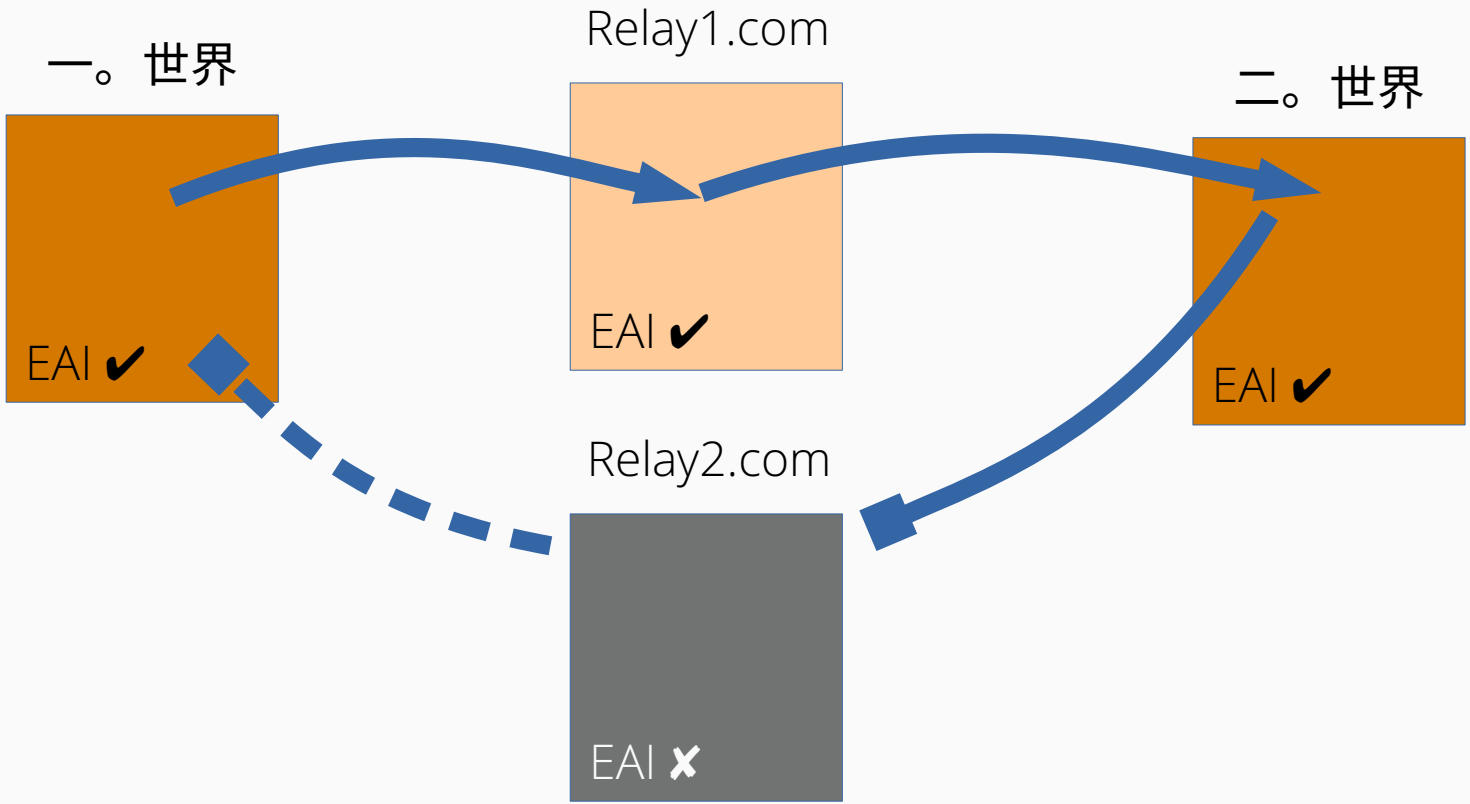
EAI challenges: send/receive to EAI address

* Can your email systems receive email from an EAI address?
* Can it send back to that EAI address?
* EAI support in email fields of customer records

EAI challenges: host EAI addresses

* Can your email systems host mailboxes with EAI addresses?
* Many related systems need to support EAI
  * Spam filters, Address books, Calendars, To-do lists
  * Mailing list servers, email marketing services

EAI challenges: delivery path obstacles



EAI challenges: delivery path obstacles

* Will relays deliver your EAI email to, receive from correspondents?
* Intermediate email relays may reject EAI
  * Outgoing & incoming email may take different routes, and so encounter different obstacles
  * Bounce messages might be discarded by intermediates

EAI challenges: EAI mailbox name considerations

* Which mailbox names do you allow?
  * Which scripts to allow? Script mixing?
  * Invalid or harmful names to prevent? Etc.
* Alternate addresses for global legibility?
  * e.g. info@jdlh.com as alternate for जिम@डाटामेल.भारत
* Paper on Best Practices for Mailbox Names, from UASG soon

EAI challenges: bi-di email addresses

* IDNs with right-to-left content can get displayed in confusing order
  * Strong RTL labels + weakly directional ‘@’, ‘.’
  * Surrounding digits might get sucked in, e.g. ‘78 #’
* e.g. تـــجربة@تـــجربة.موريتاني and تـــجربة@تـــجربة.mr
* A Best Practices guide would be helpful

EAI challenges: globally legible aliases

* How to communicate with a world which can't read your script?
  * Alternate address in another script?
  * If message delivered, how to read the contents? Translation?
* Precedent: postal mail international convention to always accept address in Latin script

Benefits of, and Issues with, IDNs and EAI

* Great benefits await ($9.8bn!)
* Issues still exist
  * More business and administrative than technical
  * Supply-demand paradox: which comes first
* IDNs, EAI, UA are advancing

More developer resources

Tools & Resources for Developers

Authoritative Tables:

* http://www.internic.net/domain/root.zone
* http://www.dns.icann.org/services/authoritative-dns/index.html
* http://data.iana.org/TLD/tlds-alpha-by-domain.txt
* See SAC070 on static TLD / suffix lists: https://tinyurl.com/sac070

Internationalized Domain Names for Applications:

* Tables: https://tools.ietf.org/html/rfc5892
* Rationale: https://tools.ietf.org/html/rfc5894
* Protocol: https://tools.ietf.org/html/rfc5891

Unicode:

* Security Considerations: http://unicode.org/reports/tr36/
* IDNA Compatibility Processing: http://unicode.org/reports/tr46/

Universal Acceptance Steering Group info & recent developments: www.uasg.tech

Five Key Tasks of Universal Acceptance

Accept. Validate. Store. Process. Display. For all domain names.
Make wise end-to-end decisions about using A-Labels, U-Labels.

UASG guides

* UASG006 Universal Acceptance Quick Guide

http:// 普遍接受 - 测试。世界

EAI case studies

We have case studies of organizations which have already supported EAI. Their experience helps you know what to expect. They may have tools you can use to help test your EAI.

UASG guides (partial)

* UASG013D Case Study: Data Xgen Technologies Pvt Ltd
  * “updating… at least 12 major elements… webmail, IMAP, POP, SMTP, contacts, calendar, antispam, search, logger and rules.”
* UASG013C Case Study: ICANN
  * Phased approach, 87 components = 46 in-house + 41 from vendors

UA use cases

IDNA Pattern	Example
ascii.long	ua-test.technology
idn.idn	普遍接受 - 测试 . 世界
idn-rtl.idn-rtl	اختبار -القبولالعالمي .شبكة
idn.ascii/unicode	普遍接受 - 测试 .top/ 我的页面
EAI Pattern	Example
unicode@idn.idn	युएअसजी@डाटामेल.भारत
ascii@ascii.idn	info4@ua-test 。世界
unicode@rtl.rtl	دون@رسيل.السعودية

UASG004 Use Cases
These domains are registered, ready to use in test suites.
Total 45 cases.

XgenPlus tools

Software developer XgenPlus has made a number of EAI- and IDNA-related tools available to developers free of charge. Here are some links to start exploring.

XgenPlus tools (partial)

* https://eai.xgenplus.com/
    * Puny Code Converter, Mix Script test, Mail Delivery Test
* Datamail multilingual email service https://www.datamail.in/
    * Email addresses in 12 scripts for iOS, Android, and web.

Next steps

Learning more about IDNs and EAI

The UASG stands ready to help support your use of IDNs and EAI, and to help you support others. There is a general email list, and working groups. Join us!

Suggested next steps for you

* Follow the UASG at https://uasg.tech/ .
* Subscribe to the *ua-discuss@uasg.tech* email list.
  * https://mm.icann.org/mailman/listinfo/ua-discuss

Play with IDNs and EAI

* Email Jim DeLaHunt <जिम@डाटामेल.भारत>
* Get your own non-Latin email address
  * e.g. https://www.datamail.in/ .
* Experiment with U-Label ⇔ A-Label
  * e.g. https://eai.xgenplus.com/Multilanguage-To-Punycode-Convertor.jsp

Q&A

Thank you!

Q&A

Slides: http://go.jdlh.com/iuc44t3t3

Evaluation: http://unicodeconference.org/eval-sp/

Email addresses and domain names are *non*-latin! Now what?

Jim DeLaHunt / IUC44 / 14 October 2020