Top Issues in Universal Acceptance
of Non-Latin Email Addresses and Domain Names

Jim DeLaHunt • IUC45 S8T2 • 15. Oct 2021
Slides, links: http://go.jdlh.com/iuc45s8t2

Abstract

The next one billion internet users use a wide variety of languages and scripts. They will demand email addresses, and domain names, in scripts they can easily read. This challenges apps and systems to provide Universal Acceptance (UA) — of all domain names and email addresses, from http:// 普遍适用测试 . 我爱你 to تجرية-بريد-الكتروني@تجرية-القبول-الشاملموريتانيا to सार्वभौमिक-स्वीकृति-परीक्षण.संगठन . We explain the most troubling obstacles and the most inspiring successes in Universal Acceptance encountered by the Universal Acceptance Steering Group. From major email platforms launching support of internationalized addresses to improving support by programming languages and libraries, it has been an exciting year.

Agenda

- Introduction
- What is Universal Acceptance?
- State of UA, 2021 News
- Top issues, strategic situation
- Q&A

Introduction

- Jim DeLaHunt
  - http://jdlh.com, ☎ +1-604-376-8953, he/him, जिम@डाटामेल.भारत

  - Squamish, Musqueam, and Tsleil-Waututh unceded territories: Vancouver, Canada
  - Consultant in multilingual websites; software engr
- Slides, links: http://go.jdlh.com/iuc45s8t2

Universal Acceptance

Universal Acceptance

"Universal Acceptance (UA)… [aims] to ensure that all domain names and email addresses are supported equally by all Internet-enabled applications on all devices and systems."

- Universal Acceptance Readiness Report FY21, UASG034, p.4.

What is Universal Acceptance?

- Internationalised Domain Names (IDNs)
  - New short top-level domain names: example.sky
  - New long top-level domain names: example.berlin
  - Internationalized Domain Names: παράδειγμα.ευ

What is Universal Acceptance?

- Email Address Internationalisation (EAI)
  - ASCII@ASCII; new short or long TLD: ekrem@misal.istanbul
  - ASCII@IDN: john@société.org
  - Unicode@ASCII: 测试 @example.com
  - Unicode@IDN: ईमेल@उदाहरण.भारत

  - Unicode@IDN, right-to-left scripts: ايميل@مثال-الموقع

UA Steering Group (UASG)

- Community-based group
  - with support of ICANN – https://uasg.tech/
- Dedicated to promoting UA
- Analysis of tech, services, applications, …
- Raises awareness, trains, writes papers …
- Remediation of apps and services …

UA Steering Group (UASG)

- Working Groups
  - Measurement WG   – Technology WG
  - Email Address Internationalization (EAI) WG
  - Communications WG
- Outreach groups
  - Local Initiatives WG – UA Ambassadors WG
- (We would love you to join us…)

State of UA, 2021 News

Fundamental specifications

- Specifications for IDNs, EAI are published
- >150 IDNs are in operation
  - .شــبكة (.network), 。我爱你 (.iloveyou), …
  - >1400 top-level domains (including Latin-script)
- Over 36,000 mail server IPs support EAI

But, fundamentals not perfect

- HTML `<input type="email">`

- Spec defines address as ASCII@ASCII !
  html.spec.whatwg.org/#valid-e-mail-address

- Discussion, proposed change
  github.com/whatwg/html/issues/4562 , thank you Addison Phillips

- Stagnant since Aug 2020
  - "needs implementer interest"

Fundamental libraries

- Many languages have adequate libraries for EAI, IDNs
- You can write apps with Universal Acceptance if you really want to
  - *i.e.* "it is possible"
  - not, "it is easy" or "every option leads to UA"

UASG gap analysis of fundamentals

- UA Compliance of… Libs and Frameworks
  - UASG018A uasg.tech/…/UASG018A-en-digital.pdf

- EAI Software Test Results
  - UASG030A uasg.tech/…/EAI-Software-Test Results-UASG030A.pdf

- UA of (CMS) Phase 1 – WordPress
  - UASG032 uasg.tech/…/UASG032-en-digital.pdf

- others.…

Customer adoption of UA

- Karnataka state government officials
  - @ಕರ್ನಾಟಕ.ಭಾರತ in Kannada language (2020?)

- Earlier:
  - Rajmail: @राजस्थान.भारत in Hindi, Rajasthan state, India https://uasg.tech/case-studies/ (2017)
  - etc….

Customer adoption of UA

- UASG crawl of SMTP servers
  - 208,511,439 second-level domains in 1,177 TLDs
  - 34,554,390 unique mail servers, 2,550,184 IPs
- 6.3% of IPs accept Unicode@ASCII addrs
- 1.45% accept Unicode@Unicode addrs
- Low support, lots of work to do

Provider and OS adoption of UA

- Apple iOS 14 can correspond with EAI addresses (Sept 2020)
    - "You can send and receive email using addresses in non-Latin languages — including Chinese, Japanese, Korean, Russian, Thai, and Hindi."
      https://www.apple.com/ios/ios-14/features/ via Wayback machine
    - No claim to support EAI mailboxes, but it works in some of our tests

Provider and OS adoption of UA

- ## Salesforce Classic can correspond with EAI email addresses

  - "Connect with users across the globe by using non-Latin-based characters in email addresses when sending emails from Salesforce."

    https://help.salesforce.com/s/articleView?id=sf.emailadmin_eai.htm&type=5

Provider and OS, existing UA

- Datamail https://www.datamail.in/ hosts webmail in 17 languages (India, 2016)
- Coremail http://www.coremail.cn/ email server hosting EAI addresses (HK, 2017)
- Gmail https://mail.google.com/ can correspond with EAI addresses (US, 2014)

Provider and OS, existing UA

- Chinese Domain Name registrar service
  - http://互联网.中国, Coremail, China (2018)
- XgenPlus enterprise email server (2011?)
  https://www.xgenplus.com/
  - And SpamJadoo spam filter https://www.spamjadoo.com/
  - by Data Xgen Technologies, India

UASG guidance for UA adopters

- *Considerations for Naming Internationalized Email Mailboxes*
  - UASG028 uasg.tech/…/UASG028-en-digital.pdf

- *Test Domain Names and Email Addresses*
  - UASG004 uasg.tech/…/UASG004-en-digital.pdf
  - Structured text: UASG004A uasg.tech/…/UASG004-en-digital.txt

- etc.…

Domain name promotions

- ThaiNIC: free .ไทย for .th domain name

- Godil.in: register names in .भारत and .in domains, 1 year free
  - India Day special, 2021-01-26

Label Generation Rules (LGRs)

- Domain name policy on what characters are permitted in parts (labels) of names
  https://www.icann.org/resources/pages/root-zone-lgr-2015-06-21-en

  - *e.g.* 台湾卤面， 檯灣鹵麵， 臺灣滷麵 group
    and 台檯簺臺颱， 湾灣， 卤滷鹵， 面麫麵

  - *e.g.* forbids script mixing: जिम 体 . 中国

- Makes IDNs more secure, more useful

Label Generation Rules (LGRs)

- Different rules for each script, language
  - 18 scripts covered (2020-11), *e.g.* Arabic, …, Thai
  - Japanese, Latin proposed (2021-09)
- Adopting these rules are separate issues
  - In Root zone, in 2nd level domains
  - Many different registry operators and regulations

Top issues, strategic situation

Supply-demand paradox

Demanders

Await UA support

Await UA demand

Suppliers

- Demanders = users
- Suppliers = app vendors
- ("catch-22", to N.A. Anglophones)
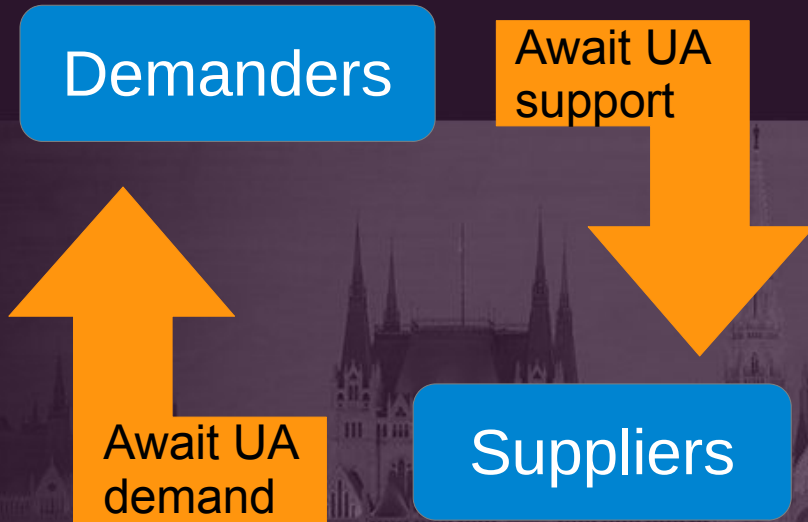
Supply-demand paradox

- Network effects are still small
- Users limited by
  - Vendor's apps, browser, email client
  - Their own readiness to settle for less
- Vendors limited by
  - App platform, libraries, email host
  - Their determination and budget

Asymmetrical costs and benefits

- Beneficiaries of UA are diffuse
  - And in part, not yet present on internet
- Costs of UA are concentrated, mostly with vendors and suppliers
- Who has an incentive to invest in UA?
  - ccTLD operators, but not a big incentive

  See *White Paper on The Role of ccTLDs in Achieving Universal Acceptance*

Email addresses as identifiers

- Even non-email apps can be a UA blocker!
  - Special case of supply-demand paradox
- User often has one email for everything
- Some apps use email address as user ID
  - *e.g.* Facebook
- If a major app rejects EAI address as ID, user will not use that address for anything

Escaping supply-demand paradox

- Commercial gains, new markets (vendors)
  - Profit: reduced costs; more value → higher price
- Much greater value, ease of use (users)
- Persuading, facilitiating (*e.g.* UASG)
- Government as requirements imposer
- Cultural pride of own language, script

History: Unicode vs code pages

- 1990's. OS's and apps use different code pages for different locales, markets
  - CP1252, ISO8859-1, Shift-JIS, …
- Separate code bases (esp 1byte vs 2byte)
- Unicode: one text format, for all locales
  - Allows multilingual documents
  - But complex, libraries not yet adequate, …

History: Unicode vs code pages

- Paradox: Unicode did not offer much value to single-language user
- Escape: OS (and app) developers
  - Entered more and more locales/markets
  - Cost savings from single global code base offset Unicode adoption cost
  - World wide web exposed users to other lang.s

History: Unicode beyond BMP

- Early 2000's. App developers use Unicode
  - Actually UCS2 (*i.e.* UTF-16 ≤ 0xFFFF, Base Multilingual Plane "BMP" only)
  - Avoiding surrogate pair handling, font switching
- Support "Unicode beyond the BMP"
  - Opens gate to GB18030:2000 and China market
  - The "right way" to support Unicode

History: Unicode beyond BMP

- Paradox: existing users satisfied with BMP characters
  - little market benefit, lots of dev cost
- Escape: emoji 🎎 👯, encoded past BMP!
  - New value for existing users
  - Justified cost of supporting rest of Unicode planes

Applying history lessons

- What UA usage could create value for users like emoji did?

- What UA usage could reduce vendor cost or increase addressable market like Unicode and the "single code base" did?

Threat: irrelevance

- UA assumes that humans care about reading domain names, URLs, email addrs
- Mobiles bring chat apps, walled gardens
- Chat apps: email addr → phone number
- Walled gardens: URL → in-garden links
- Search engines: domain name → search

Weakness: promoting UA support

- Datamail.in shows EAI support clearly
- "World's first IDN compliant free email service… in their native language with their preferred language email address. The vernacular email service is available in 16 languages/22 Scripts."

Weakness: promoting UA support

- Apple iOS: apple.com/ios/ios-15/ ???
  - support.apple.com/en-us/HT211808 7 words
- Microsoft: outlook.live.com/owa/ ???
  - news.microsoft.com/…indian-language… 2018
- Gmail, Salesforce, Spamjadoo, etc…
- Components like Postfix, Courier, Halon…

Weakness: promoting UA support

- No widely understood UA support labels
  - "Accepts IDN URLs", "hosts EAI mailboxes", etc.
- Labels help vendors promote UA, EAI support, and let customers ask for it
- UASG EAI working group: preparing EAI support self-certification guide

Weakness: UA, EAI terminology

- Our terminology is jargon, unclear to those outside our bubble
  - "UA"? "EAI"? "IDNs"?
- Need clear terms which we can promote
  - *e.g.* "can correspond with but not host EAI addrs"
- UASG (esp EAI WG) grappling with this

Weakness: cross-script email usage

- User A sends email, URL in language A to user B, who doesn't read language A
- What should happen?
  - Alias or fall-back email address?
  - Users won't try this? EAI is within region only?
- Analogies: how to use emoji. Int'l post.

Weaknesses: briefly

- Risk of homograph attacks via IDNs is exaggerated
  - Other risks are much larger
  - Label Generation Rules will reduce the low risk
- Software engineering education is ASCII-retentive

Opportunity: consumer campaign

- Where to look for UA value?
- Idea: a consumer campaign
- To a user base preferring non-Latin script
- Off-internet ads (radio, TV, signs) driving traffic to web pages, email addresses
- Preferred language for URLs, addresses

Conclusion

Agenda

- Introduction
- What is Universal Acceptance?
- State of UA, 2021 News
- Top issues, strategic situation
- Q&A

Conclusion

- Universal Acceptance is important
- EAI addresses and IDNs are real
- Adoption is difficult, but rising
- What is your UA opportunity?
- Participate in the UASG! https://uasg.tech/

Thank you! Q&A

Thank you!

- Q&A
- Slides: http://go.jdlh.com/iuc45s8t2

Top Issues in Universal Acceptance
of Non-Latin Email Addresses and Domain Names

Jim DeLaHunt • IUC45 S8T2 • 15. Oct 2021
Slides, links: http://go.jdlh.com/iuc45s8t2