Email Address Internationalization

CSA 2019 | Cologne | April 2019

John Levine, john.levine@standcore.com
My new e-mail address

yés@nø.sp.am
A very short history of e-mail

In three acts
From: Boris <boris@example.com>
To: Ines <ines@example.org>
Subject: Lunch cooperation

How about 1 PM at the cafe?

All text is ASCII
From: Борис <boris@example.com>
To: Inés <ines@example.org>
Subject: Когда будет ланч?

How about 1 PM at the café?

Non-ASCII in most headers
Non-ASCII bodies
From: Борис <Борис@пример.com>
To: Inés <inés@example.org>
Subject: Когда будет ланч?

How about 1 PM at the café?

• UTF-8 everywhere
• In all visible headers and bodies
Make mail work for global users

The Internet’s users are increasingly diverse

* Many languages and scripts, including Arabic, Chinese and over 100 others.
* Users want domain names and email addresses in their own scripts.
  * weiß.koeln, landstraße.wien, ...
  * 動物を殺さなくても僕らは履きやすい靴をつくれる.tokyo
Make mail work for global users

Who is using EAI mail?

* Countries with literate populations, non-Latin alphabets
* Rajasthan (India) provides Hindi addresses to citizens
* Thailand has many Thai-only Internet users
* Important for in-country communication
  * To people you can’t reach with European languages
Goals for Today’s Lecture

1. Understand the basics of Internet SMTP mail

2. Understand Unicode and Internationalized Domain Names (IDNs)

3. Understand threats and challenges related to Email Address Internationalization (EAI)
Building Blocks: Domain Name System

* Internet resources have numeric Internet Protocol (IP) addresses. They are hard to remember, so the Domain Name System (DNS) provides readable names for IP addresses or other resources.

uasg.tech → 46.22.137.49 (IPv4)
uasg.tech → 2a01:a8:dc0:2002::100 (IPv6)
Building blocks: **Internet Mail**

Sender MTA

Receiver MTA

Sender MUA

Receiver MUA
Building blocks: **SMTP**

![Diagram of SMTP process]

- **User PC MUA**
  - SUBMIT or webmail to **MSA Sender MTA**
  - POP / IMAP or webmail to **Recipient MTA**

**SMTP**
Building blocks: **SMTP COMMANDS (1)**

R: 220 mail1.example.org ESMTP
S: EHLO mailout.example.com
R: 250- mail1.example.org
R: 250 8BITMIME
R: 250 STARTTLS
S: MAIL FROM: <boris@example.com>
R: 250 2.1.0 Sender ok.
S: RCPT TO: <ines@example.org>
R: 250 2.1.5 Recipient ok.
... to be continued ...
... continued from above ...

S: DATA
R: 354 Send your message.
S: ... message header and body ...
S: .
R: 250 2.6.0 Accepted.
S: QUIT
R: 221 2.0.0 Good bye.
Building Blocks: **Character Sets and Scripts**

Languages are written using writing systems.

* Most writing systems use a single script, a set of graphic characters (glyphs).
* Some, e.g. Japanese use several scripts.

People can read scripts. But computers need numeric values that they can process. The mechanism for this is called an *encoding*.
Building Blocks: ASCII and Unicode

A character mapping associates characters with specific numbers. Many different mappings have been created over time for different purposes, two are now by far the most widely used: ASCII and Unicode.

**ASCII**: unaccented Latin letters, digits, punctuation

**Unicode**: everything else
Building Blocks: ASCII and Unicode (cont.)

**ASCII**

Domain names limited to the characters A-Z, the numbers 0-9, and hyphen “-“.

**Unicode**

Over 1 million characters, intended to represent every written language. Each Unicode glyph is assigned a number called a code point.
Unicode Code Points Examples

U+041A Cyrillic letter Ka  
U+3069 Hiragana letter Do  
U+0636  Arabic letter Dad  
U+00E1 Small A with acute  
U+0062 Small letter a  
U+0201 Acute accent  

**U+xxxx** means the Unicode code point with hex value xxxx.
## Building Blocks: Unicode and UTF-8

<table>
<thead>
<tr>
<th>Unicode</th>
<th>UTF-8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Code points 0x0-0x7F are the same as ASCII. The highest code point is 0x10FFFF.</td>
<td>UTF-8 uses 1-4 bytes per Unicode code point.</td>
</tr>
<tr>
<td>Non-ASCII code points do not fit in a one 8-bit byte.</td>
<td>0x0-0x7F are the same as ASCII.</td>
</tr>
</tbody>
</table>
Building Blocks – Internationalized Domain Names and Email Addresses

* Unicode enables domain names and email addresses to contain non-ASCII characters.

* Domain names with non-ASCII characters are *Internationalized Domain Names* (IDNs). An IDN can be all non-ASCII or a mix of ASCII and non-ASCII labels.

* Email addresses with non-ASCII characters are called *Internationalized Email Addresses*. 
Building Blocks – Internationalized Domain Names and Email Addresses

* Non-ASCII labels use a new encoding in the DNS.
* Unicode labels are called U-labels. The ASCII-translated versions are A-labels, which start with xn--.
* For example, 普遍接受-测试.世界 becomes xn----f38am99bqvcd5liy1cxsg.xn--rhqv96g
* A-labels are not meaningful to human users, so whenever possible display the U-label instead.
Email addresses contain two parts:

1. **Local part** (the part before the “@” character)
2. **Domain** (after the “@” character)

* Both parts may be Unicode.
* A Unicode domain is an IDN
Email Address Internationalization: EAI

ASCII sender

Bob@example.com

ASCII recipient

EAI sender

猫王@普遍接受-测试.世界

EAI recipient
For MUA and MTA: Changes to SMTP

* New SMTP feature SMTPUTF8
* UTF-8 in addresses

R: 220 receive.net ESMTP
S: EHLO sender.org
R: 250 8BITMIME
R: 250 STARTTLS
R: 250 SMTPUTF8
S: MAIL FROM: <猫王@普遍接受-测试.世界> SMTPUTF8
R: 250 Sender accepted
Server Software (MTA - Mail Transport Agent)

* Servers advertise the SMTPUTF8 feature
* Clients check server for the SMTPUTF8 feature, use the SMTPUTF8 option when sending
* Don’t send EAI mail to servers that do not support it
  * Provide readable error reports when users try to do so
* Accept both U-label and A-label versions of domain names in e-mail addresses
* Do “fuzzy” matching in incoming addresses, variations such as upper/lower case or missing accents
Items for Email Service Providers to Consider

* Avoid addresses that can confuse users, offer Unicode mailbox names that conform to best practices
  * Unicode consortium and IETF provide guidance

* Avoid mailboxes with easily confused local parts
  * Don’t assign bob and bób and bøb
Items for Email Mailbox Providers to Consider

* Do “fuzzy” matching on local parts of incoming mail
  * Allow variations such as upper/lower case, wrong accents, or variant characters
  * Handled locally in MTA, remote MTAs and users don’t do anything special
  * Fuzzy matching is not new, that’s why upper/lower case in addresses doesn’t matter
* Offer ASCII mailbox aliases along with EAI mailbox names.
* Both names deliver to the same mailbox, so users can give addresses to both EAI and non-EAI correspondents.
Message downgrading

* You can’t downgrade an EAI message to an ASCII message without losing information.
  * One cannot turn an EAI address into an ASCII address.
* In general, spend effort making software EAI-capable rather than trying to invent non-EAI workarounds.
Security challenges

- Homographs and near homographs
- Variants
Homographs

* They look the same but are not the same
* Also near-homographs like ₁ ₁
* Forbid names in combined scripts
Two-way text

* Most languages written left to right (L2R)
* Arabic, Hebrew, some others written right to left (R2L)
* Mixed text is hard to type or display, confusing to users

www.مثال..sa

مثال.السعودية
Mixed scripts

* IDNs can use thousands of different characters
  * In different scripts
* Chinese, Arabic, Cyrillic, ..., mixtures allowed in IDNs
* Script rules: limit each label to one script or compatibles
* Rules only apply at 2\textsuperscript{nd} level
Variant characters

* Different appearance, same meaning
* Allow one in names, forbid the rest?
* Allow all, map to the same place?
* Something else?
* A decade long ICANN swamp
Mail address challenges

- Longer, unexpected domain names
  someone@home.sandvikcoromant
- Several ways to write the same character
  – Is it á or ´+ a ?
- Punctuation possible in local parts
- Way too many emojis 😣 😣 🥑
• A-labels are usually unreadable
  xn--onqrps50a3m1a8owtum7fb.xn--fiqs8s
  or 难以阅读的例子.中国
• Tools to convert can help
EAI software can be tricky to debug fully. Some problems may only be apparent when using some scripts, e.g. LTR and RTL scripts.

• Ensuring reliable EAI mail
  – Send and receive test messages using different scripts
  – Exchange test messages with many different other EAI-capable mail systems
Message forensics

* Most message headers unchanged other than UTF-8 text
* Received headers log EAI message transfers
* Messages more likely to be in languages your staff doesn’t read
Message forensics: Received headers

Received: from m227-25.mailgun.net (m227-25.mailgun.net
[159.135.227.25]) by mail1.iecc.com ([64.57.183.56])
with UTF8SMTP via TCP port 49209/25 id 601829735;
 06 Sep 2018 13:02:03 -0000

Received: from production.m3aawg.org
(ec2-34-214-179-220.us-west-2.compute.amazonaws.com
[34.214.179.220])
by mxa.mailgun.org with ESMTP id 5b912548.7f134c2f2930-smtp-out-n03; Thu, 06 Sep 2018
13:02:00 -0000 (UTC)
Summary

* EAI mail is on its way
* Likely to be popular for non-technical users
* Get your tools ready
Tools & Resources for Developers

Authoritative Tables:

* http://data.iana.org/TLD/tlds-alpha-by-domain.txt
* See also SAC070: https://tinyurl.com/sac070

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
Email Address Internationalization

CSA 2019 | Cologne | April 2019

John Levine, john.levine@standcore.com