Email Address Internationalization – Technical Overview

Universal Acceptance

A very short history of e-mail

In three acts

Internet mail, classic edition

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

How about 1 PM at the cafe?
```

All text is ASCII

Internet mail, MIME edition

```
From: Борис <boris@example.com>
To: Iñes <ines@example.org>
Subject: Когда будет ланч?

How about 1 PM at the café?
```

Non-ASCII in most headers
Non-ASCII bodies

Internet mail, now with EAI

```
From: Борис <Борис@пример.com>
To: Iñes <iñes@example.org>
Subject: Когда будет ланч?

How about 1 PM at the café?
```

- UTF-8 everywhere
- In all visible headers and bodies

Work and communicate with a global users

The Internet now provides access to an increasingly diverse user group

- Many languages and scripts are used on the Internet, including non-Latin based Arabic, Chinese and many other scripts.
- Users want domain names and email addresses in their own scripts.
- Internet-enabled applications, devices and systems need to accept, validate, store, process and display all domain names and email address appropriately.

Goals for Today’s Lecture

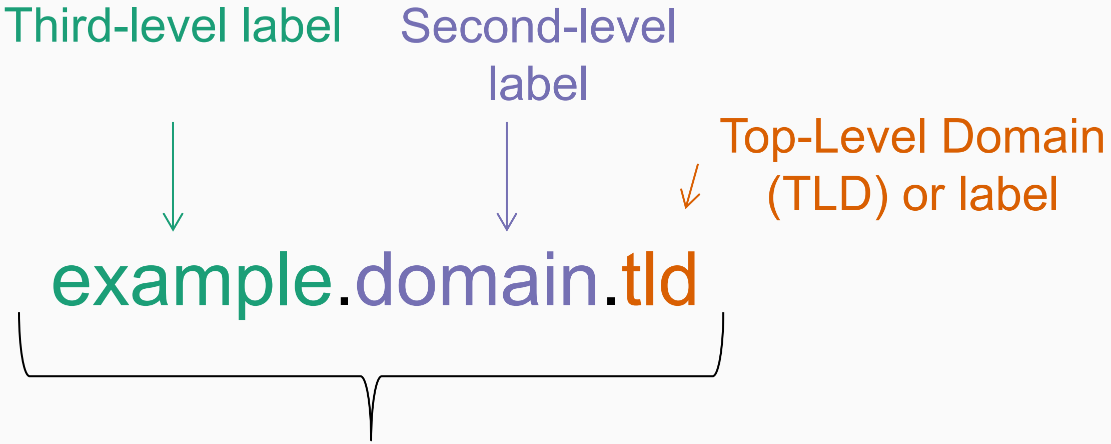
1 Understand the basics of Internet SMTP mail

2 Understand Unicode and Internationalized Domain Names (IDNs)

3 Understand and be able to implement good practices related to supporting Email Address Internationalization (EAI)

Building Blocks: **Domain Names**

A domain name is dotted text strings used as a human-friendly technical identifier for computers and networks on the Internet



Third-level label

Second-level label

Top-Level Domain (TLD) or label

example.domain.tld

Each dot represents a level in the Domain Name System (DNS)

Building Blocks: **Domain Name System**

- Each resource on the Internet is assigned an address to be used by the Internet Protocol (IP). Since IP addresses are difficult to remember, the Domain Name System (DNS) provides a mapping between human-readable names and IP addresses or other resources.

uasg.tech → 46.22.137.49 (IPv4)

uasg.tech → 2a01:a8:dc0:2002::100 (IPv6)

Building Blocks: **DNS (cont.)**

A domain name server handles domain name lookups, and return many different record types. Some of the most common are as shown below.

Address records (A and AAA)

- These link a domain name to an IPv4 or IPv6 address.

Mail Exchange records (MX)

- These to identify mail servers.

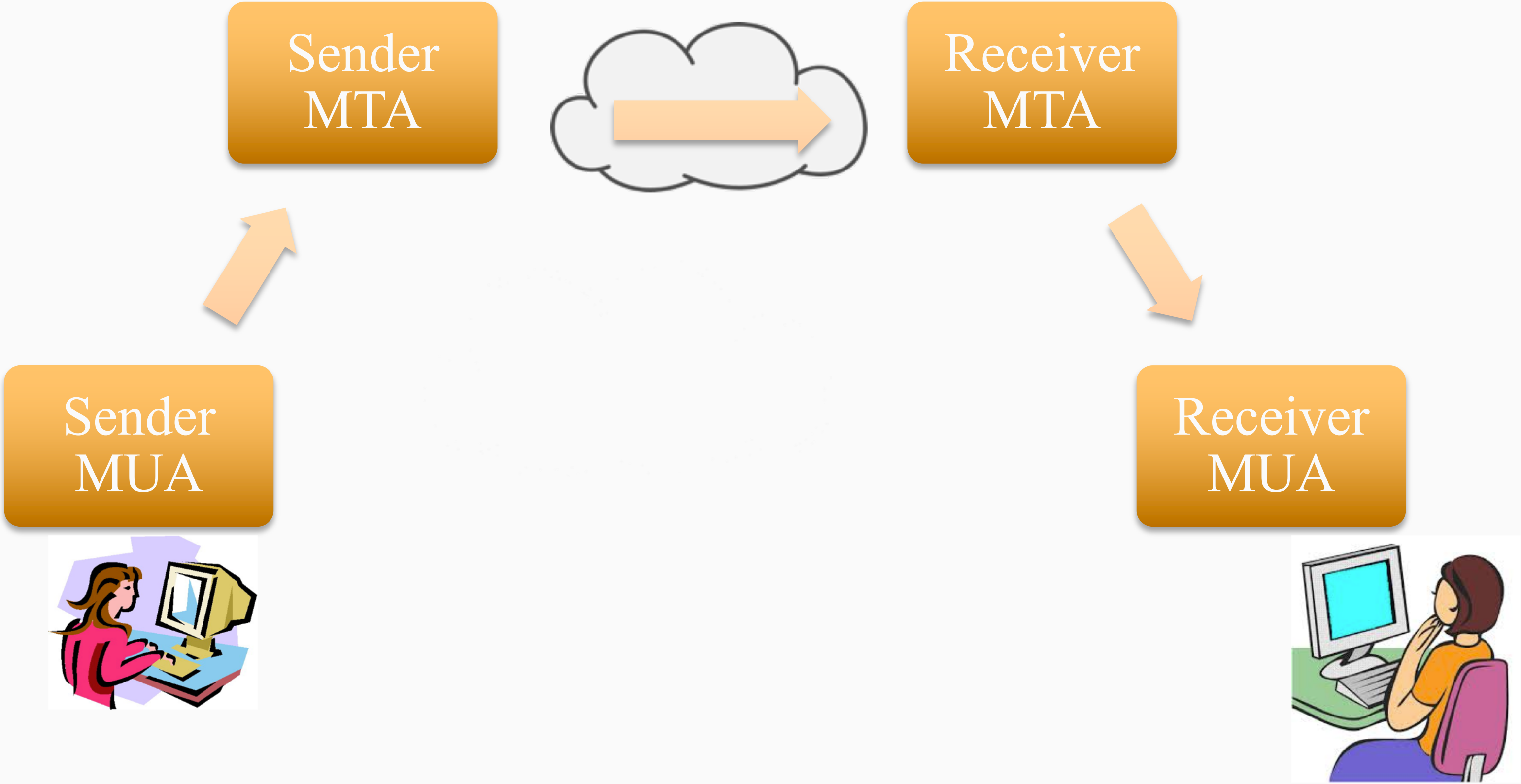
CNAME records

- These allow domain names to be aliased to another domain name.

e.g.,

```
mail.example.com. IN    A      102.34.56.7
mail.example.com. IN    AAAA   2620:0:2d0:201::1:71
example.com.      IN    MX     mail.example.com.
www.example.com.  IN    A      102.34.56.8
ftp.example.com.  IN    CNAME       www.example.com.
```

Building blocks: **Internet Mail**



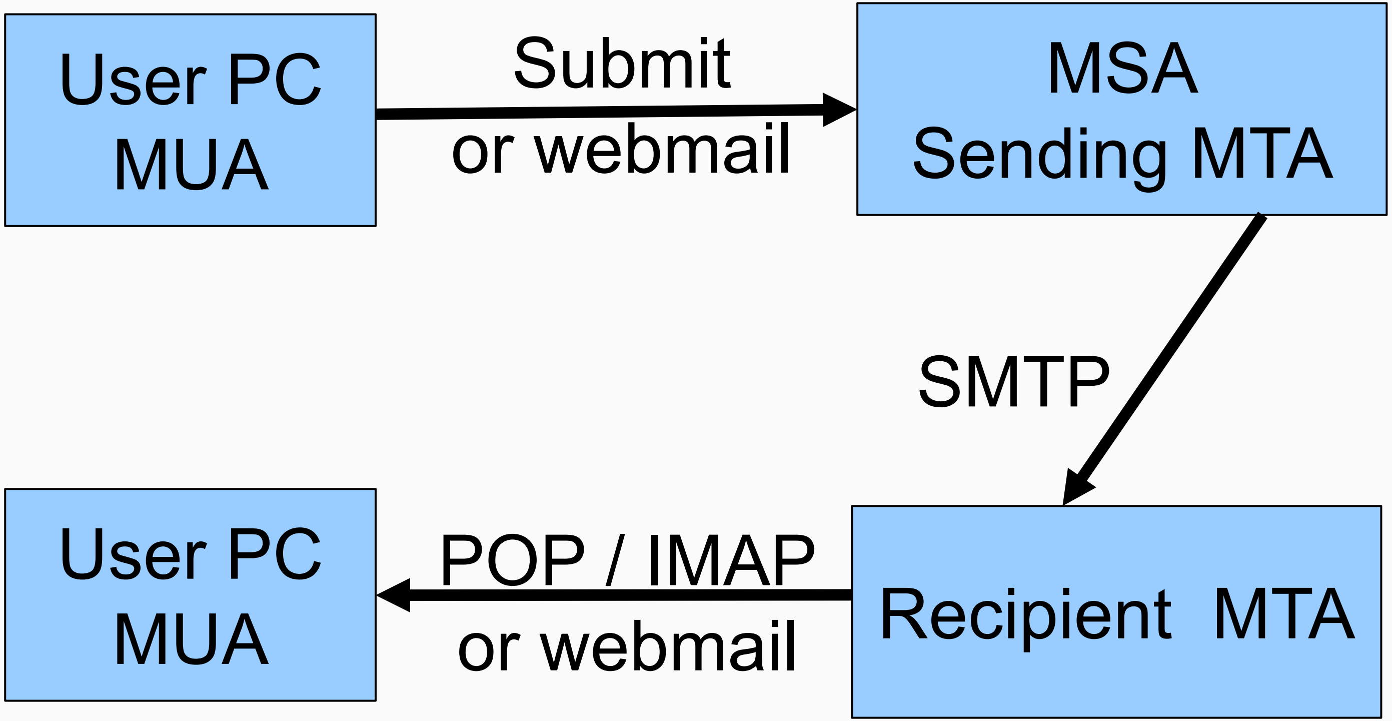
Sender MTA

Receiver MTA

Sender MUA

Receiver MUA

Building blocks: **SMTP**



Building blocks: **SMTP COMMANDS**

```
R: 220 mail1.example.org ESMTP
S: EHLO mailout.example.com
R: 250-mail1.example.org
R: 250 8BITMIME
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.
S: DATA
R: 354 Send your message.
S: … message contents …
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, which is a set of graphic characters (glyphs) used for the written form of that language.
- Some writing systems such as Japanese use more than one script in a single document.

These scripts can be read by humans. But they are not useful to computers, which need a script encoded in a way that they can process (for example, to resolve a web address). The mechanism for this is called a character mapping .

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

Most of the text currently displayed on the Internet is in the unaccented Latin character set. The American Standard Code for Information Interchange (ASCII) character-encoding scheme includes upper and lower case Latin letters. ASCII was designed in the 1960s, and is based on the English language. For historical reasons, it became the standard character encoding scheme on the Internet.

Unicode

Because most writing systems do not use the unaccented Latin character set, expanded encodings have been created.

To see all Unicode character code charts, go to: http://unicode.org/charts

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

ASCII

ASCII uses only 7 bits per character, which limits the set to 127 characters, not all of which can be used in domain names.

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

Unicode

Unicode encodes more than 1 million characters. Unicode intends to represent every character in every written language.

Each of these Unicode characters is assigned a number called a code point.

Unicode Code Points Examples

U+041A # Cyrillic LETTER KA к
U+041B # Cyrillic LETTER EL л
U+041C # Cyrillic LETTER EM м

U+A840 # Phags_Pa LETTER KA ꡀ
U+A841 # Phags_Pa LETTER KHA ꡁ
U+A842 # Phags_Pa LETTER GA ꡂ

Jiu - the Chinese word for 'alcoholic beverage’
Unicode character 酒
Unicode code point is U+9152 (also referred to as: CJK UNIFIED IDEOGRAPH-9152);

Building Blocks: **Unicode and UTF-8**

Unicode

Unicode assigns each character a numeric code point. Points 0x0-0x7F are the same characters as ASCII. The highest code point is 0x10FFFF.

Non-ASCII code points do not fit in a one 8-bit byte.
UTF-32 stores each code point in a 32-bit word, convenient for processing but bulky in storage.

UTF-8

UTF-8 is a variable length encoding of Unicode.
Points 0x0-0x7F are encoded in a single byte, backward compatible with ASCII.

Other points are encoded as two, three, or four bytes. The number of bytes depends on the code point.

Building Blocks – **Internationalized Domain Names and Email Addresses**

- The use of Unicode enables domain names and email addresses to contain non-ASCII characters.
- Domain names that use non-ASCII characters are called Internationalized Domain Names (IDNs). An IDN can be all non-ASCII or a mix of ASCII and non-ASCII labels.
- Email addresses that use non-ASCII characters are called Internationalized Email Addresses.

Building Blocks – **Internationalized Domain Names and Email Addresses**

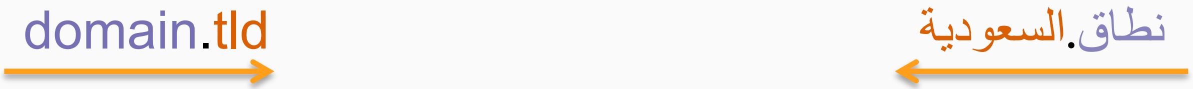
- The DNS previously only used ASCII, so non-ASCII labels use a new encoding in the DNS.
- Unicode labels are called U-labels. The ASCII-translated equivalent are A-labels, which start with xn--.
- For example,

  普遍接受-测试.世界

  becomes

  xn----f38am99bqvcd5liy1cxsg.xn--rhqv96g
- The ASCII encoding after the xn-- is sometimes called Punycode.
- A-labels are not meaningful to human users, so whenever possible display the meaningful U-label instead.

Building Blocks: **LTR, RTL, and bi-directional**



- Most scripts display characters from left to right when text is presented in horizontal lines, but some, such as Arabic or Hebrew, display from right to left.
- Text can also be bidirectional (left to right – right to left) when a right-to-left script uses digits written from left to right or includes words from English or other scripts.
- Bidirectional text can occur in any text fields such as subject lines and message bodies, as well as in IDNs and e-mail addresses.
- Displaying bi-directional text is complex and often confusing to users.

Email Address Internationalization: EAI

Email Address Internationalization (EAI)

Email addresses (also called Mailbox names) contain two parts:

1. **Local part** (the username, before the “@” character)
2. **Domain** (after the “@” character, which in this case includes the **TLD** part)

- * The domain part can contain ***any TLD,*** including IDN TLDs.
- * Both portions may be Unicode.

Advice for Client Software (MUA)

- Display headings and prompts in the user’s language
- Store and display the Mailbox name in Unicode
  - Accept Unicode mailbox names everywhere mailboxes are used, e.g., message composition, reply, address book
- Follow good practice guides for Linkification within the body of the email (see UASG 010 – Quick Guide to Linkification).
- Follow good practice for validation of domain names (see UASG 007 – Introduction to Universal Acceptance).
- Identify EAI messages when submitting to MSA/MTA
  - Be prepared for submission to fail with a non-EAI MSA

Server Software (MTA - Mail Transport Agent)

- When receiving mail, advertise the SMTPUTF8 feature
- When sending mail, check for the SMTPUTF8 feature on the remote mail server, use the SMTPUTF8 option when sending mail
- Do not send EAI mail to remote servers that do not support it
  - Provide readable error reports when users attempt to do so
- Accept both U-label and A-label versions of domain names in e-mail addresses
- Do “fuzzy” matching of local parts in incoming addresses
  - Allow minor variations such as upper/lower case or missing accents

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 SMTPUTF8
S: MAIL FROM:<猫王@普遍接受-测试.世界> SMTPUTF8
R: 250 Sender accepted
```

POP & IMAP Servers

- Post Office Protocol (POP3) has UTF8 option to allow UTF-8 in usernames, passwords, and protocol-level text strings (see RFC 6856).
- Internet Message Access Protocol (IMAP4) has UTF-8 option for UTF-8 in user names, passwords, folder names, and search strings (see RFC 6855).
- Both can optionally downgrade received messages so non-EAI clients can see an approximation of received EAI messages

Items for Email Service Providers to Consider

- To avoid addresses that can confuse users, offer Unicode mailbox names that conform to best practices.
  - Follow the domain name label generation rules for the selected script, or
  - Follow the identifier rules in Unicode UTS#39, or
  - Use the the IETF PRECIS UsernameCaseMapped identifier profile.
- Don’t create different mailboxes with easily confused local parts
  - Avoid easily confused characters, e.g., homographs or names that differ only in accents or other modifiers
- Do “fuzzy” matching on local parts of incoming mail
  - Allow minor variations such as upper/lower case or missing accents

Items for Email Service Providers to Consider

- Consider offering an ASCII mailbox name to users in addition to an EAI mailbox name.
- If both names deliver to the same mailbox, users will find it easier to share addresses with other users whose mail systems don't support EAI.

Message downgrading

- In general it is not possible to downgrade an EAI message to an ASCII message without losing information.
  - There is no way to turn an EAI address into an ASCII address.
- If a sending system knows ASCII aliases for all EAI addresses in a message (To:, From:, etc.) it may be possible for that system to downgrade and resend it.
- In general, spend effort making software EAI-capable rather than trying to invent non-EAI workarounds.

Challenges during transition

Until all the email software deployed is EAI-ready, there will be some challenging situations that arise in the sending and receiving of emails.

- Ensuring delivery to non-EAI-ready mail systems:
  - Create aliases for mailbox names in non-ASCII scripts.
  - Send mail to ASCII addresses as ASCII mail

Challenges during transition

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

Summary

- The Internet’s technology, including its naming components, are evolving and changing. The languages used on the Internet are increasingly non-Latin based. Non-Latin domain names (IDN) and email addresses (EAI) are increasingly popular.
- These changes affect the development and maintenance of all Internet-enabled apps.
- Your Internet-enabled applications, devices and systems must accept, validate, store, process and display all domain names and email address appropriately

Tools & Resources for Developers

Authoritative Tables:

- http://data.iana.org/TLD/tlds-alpha-by-domain.txt
- See also SAC070: https://tinyurl.com/sac070
- Repository of IDN Practices: https://www.iana.org/domains/idn-tables/

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

Glossary, partial

A-label - The ASCII-compatible encoded (ACE) representation of an internationalized domain name, e.g. how it is transmitted internally within the DNS protocol. A-labels always commence with the prefix “xn--”.

ASCII Characters - American Standard Code for Information Interchange. These are characters from the basic Latin alphabet together with the European-Arabic digits. These are also included in the broader range of "Unicode characters" that provides the basis for IDNs.

API - An Application Programming Interface (API) is a set of routines, protocols, and tools for building software and applications. An API may be for a web based system, operating system, or database system, and it provides facilities to develop applications for that system using a given programming language.

Codespace - Range that define the lower and upper bounds for an encoding.

Code Points - A code point or code position is any of the numerical values that make up the code space. They are used to distinguish both, the number from an encoding as a sequence of bits, and the abstract character from a particular graphical representation (glyph).

DNS Root Zone - The root zone is the central directory for the DNS, which is a key component in translating readable host names into numeric IP addresses.

Glossary, partial

EAI - Email Address Internationalization is an email address that allows Unicode in all parts of the email address.

IANA - Internet Assigned Numbers Authority. Its functions include: (1) Maintenance of the registry of technical Internet protocol parameters, (2) Administration of certain responsibilities associated with Internet DNS root zone, and (3) Allocation of Internet numbering resources.

ICANN - The Internet Corporation for Assigned Names and Numbers (ICANN) is an internationally organized, non-profit corporation that has responsibility for Internet Protocol (IP) address space allocation, protocol identifier assignment, generic (gTLD) and country code (ccTLD) Top-Level Domain name system management, and root server system management functions.

IDN - Internationalized Domain Names. IDNs are domain names that include characters used in the local representation of languages that are not written with the twenty-six letters of the basic Latin alphabet “a-z”, the numbers 0-9, and the hyphen “-“.

IDNA - Internationalized Domain Names in Applications.

IDN ccTLD - Country Code Top-level Domain that includes characters used in the local representation of languages that are nor written with the twenty-six letters of the basic Latin alphabet “a-z”. Examples: .рф (Russia), . صر Egypt, and . السعودية Saudi Arabia.

Glossary, partial

IETF - The Internet Engineering Task Force (IETF) is a large open international community of network designers, operators, vendors, and researchers concerned with the evolution of the Internet architecture and the smooth operation of the Internet. The IETF develops Internet Standards and in particular the standards related to the Internet Protocol Suite (TCP/IP), e-mail (SMTP), and other protocols.

Language - The method of human communication, either spoken or written, consisting of the use of words in a structured and conventional way.

Registrar - An organization where domain names are registered by users. The registrar keeps records of the contact information and submits the technical information to a central directory known as the “registry”.

Registry - The authoritative, master database of all domain names registered in each Top Level Domain.

RFC - A Request for Comments (RFC) is a document from the Internet Engineering Task Force (IETF.)

Glossary, partial

Script - The collection of letters or characters used in writing, representing a language.

Second-level domain name - In the Domain Name System (DNS) hierarchy, a second-level domain (SLD or 2LD) is a domain that is directly below a top-level domain (TLD). For example, in example.com, example is the second-level domain of the .com TLD.

U-label - A "U-label" is an IDNA-valid string of Unicode characters including at least one non-ASCII character. Every U-label corresponds to a unique A-label, and vice-versa

UA-ready Software or UA-Readiness - Universal Acceptance Ready Software. It is a software that has the ability to Accept, Store, Process, Validate and Display all Top Level Domains equally and all IDNs, hyperlink and email addresses equally.

Unicode - A universal character encoding standard. It defines the way individual characters are represented in text files, web pages, and other types of documents. Unicode was designed to support characters from all languages around the world.