Email Address Internationalization – Technical Perspective

Universal Acceptance

Warm-up Exercise

Each of the 3 groups below contain lists of Top Level Domains (TLDs) that are valid (approved and delegated by ICANN), except that each list contains one made-up or invalid TLD. ***Which TLD in each group is invalid?***

Group A

嘉里
ANALYTICS
BLOCKBUSTER
DIAMONDS
HOTELES
广东
MOVISTAR
இந்தியா
REALLY
政务

Group B

ABC
مصر
ATHLETA
இலங்கை
CANCERRESEARCH
CITIC
新加坡
ESURANCE
FAKE
ไทย

Group C

GMAIL
قطر
JOY
LIKE
հայ
ONYOURSIDE
OOO
فلسطين
SILLY
SUCKS

Warm-up Exercise

The 3 invalid TLDs are highlighted below in red. There are 1541 valid TLDs (as of January 2nd, 2018), and the list is continuing to grow.

Group A	Group B	Group C
嘉里	ABC	GMAIL
ANALYTICS	مصر	قطر
BLOCKBUSTER	ATHLETA	JOY
DIAMONDS	இலங்கை	LIKE
HOTELES	CANCERRESEARCH	հայ
广东	CITIC	ONYOURSIDE
MOVISTAR	新加坡	OOO
இந்தியா	ESURANCE	فلسطين
REALLY	FAKE	SILLY
政务	ไทย	SUCKS

Prepare to do business and communicate with a global user base

Continued expansion of the Internet is allowing access to an increasingly diverse user group

* There is a growing number of languages and scripts present on the internet, including non-Latin based, language-specific domain names in Arabic, Chinese and many other scripts.
* It is required that your internet-enabled applications, devices and systems accept, validate, store, process and display all domain names and email address appropriately.

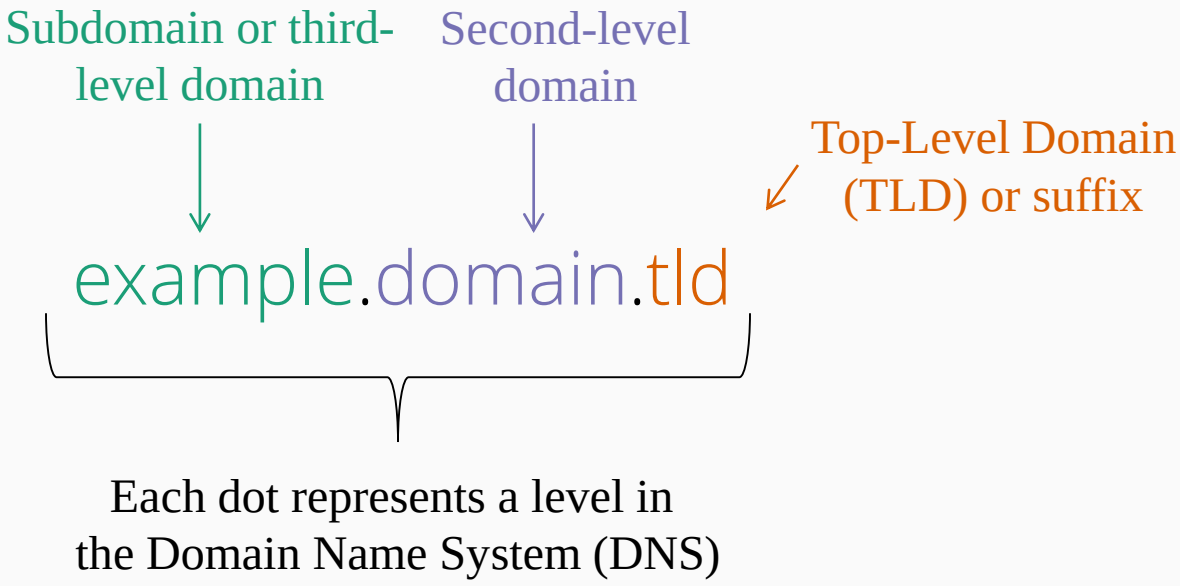
Goals for Today's Lecture

1. Understand character mapping, Punycode, right-to-left scripts, and the Bidi algorithm

2. 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 string used as a human-friendly technical identifier for computers and networks on the internet



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 IP addresses and human-readable domain names. Servers collectively providing a public DNS exist at well-known addresses on the Internet.

uasg.tech → 46.22.137.49

Discussion question: How does shared hosting fit into description above?

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

A domain name server table contains lists of domain name aliases for IP addresses. They use three different types of records as shown below.

Address records (A)

* These link a domain name to an IP address.

Mail Exchange records (MX)

* These are used to identify mail exchange servers.

CNAME records

* These allow domain names to be aliased to an IP address via a domain name already linked to an IP address.

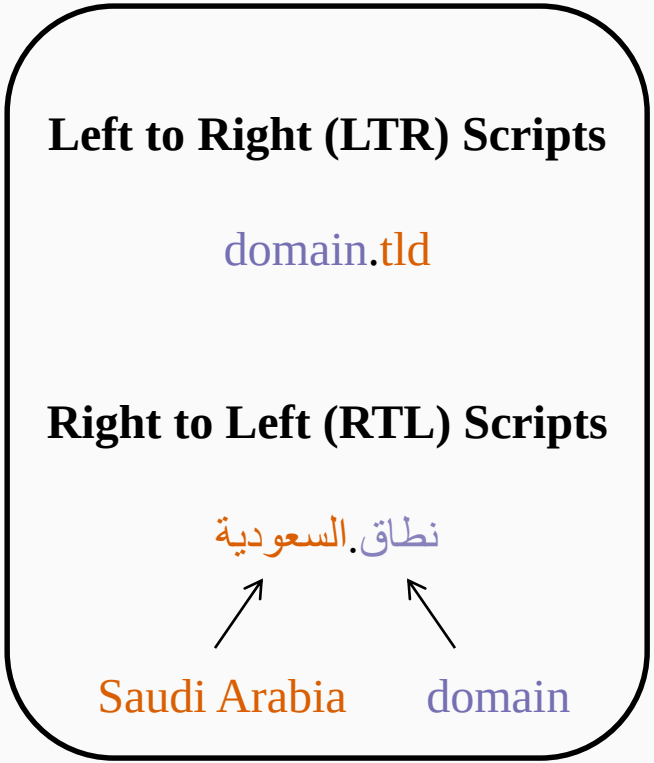
e.g.,

mail.made-this-up.com	IN	A	102.34.56.7
host.made-this-up.com	IN	MX	mail.made-this-up.com
www.made-this-up.com	IN	A	102.34.56.8
ftp.made-this-up.com		IN	CNAME www.made-this-up.com

Building Blocks: **Top Level Domains**

Human readable domain names are managed by organizations known as registries

* When a domain name is registered, it consists of multiple text strings representing multiple domain levels, each separated by a "." character
* In LTR scripts, the right-most domain level is the top-level domain (TLD)
* Some TLDs are delegated to specific countries or territories. These are called Country Code TLDs (ccTLDs)



Discussion question: When might RTL scripts cause problems?

Building Blocks: **Top Level Domains (cont.)**

* On January 2, 2018, there were 1541 valid top-level domains on the list maintained by the IANA.

Ever growing list of TLDs → ICU

* The list is updated from time to time and is available at http://data.iana.org/TLD/tlds-alpha-by-domain.txt

Building Blocks: **gTLDs**

* Starting in 2013, ICANN (the organization responsible for the creation and maintenance of TLD assignments) approved the creation of a large number of new TLDs. These new TLDs can represent brands, communities of interest, geographic communities (cities, regions) and more generic concepts. Collectively, all of these new TLDs are known as Generic Top Level Domains (gTLDs).

Original Seven gTLDs	Common cc TLDs	New gTLDs
.com	.de	.top
.org	.cn	.xyz
.gov	.uk	.loan
.edu	.nl	.club
.mil	.eu	.网址
.net	.ru	.信息
.int	.tk	.コム

Building Blocks: **Character Sets and Scripts**

Languages are written using writing systems.

* Most writing systems use one script, which is a set of graphic characters used for the written form of one or more languages.
* A small number of writing systems employ more than one script at the same time.

These characters or scripts can be recognized by humans. However, they are not useful to computers. Instead, a computer needs a script to be encoded in a way that it can process (for example, to resolve a web address). The mechanism for this is called a character mapping or coded character set (CCS), or a code page.

Building Blocks: **ASCII and Unicode**

A character mapping associates characters with specific numbers. Many different code pages have been created over time for different purposes, but for this topic we will focus on only two: ASCII and Unicode.

ASCII

Most of the text currently displayed on the internet is in the Latin character set. This character set is included in the American Standard Code for Information Interchange (ASCII, or US-ASCII) character-encoding scheme. ASCII is an older encoding scheme and was 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 Latin character set, alternate encodings have also been adopted. The most common form of Unicode is called Universal Coded Character Set Transform Format 8-bit (UTF-8).

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 128 characters, not all of which can be used in domain names.

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

Unicode

Unicode, also known as the Universal Coded Character Set (UCS), is capable of encoding more than 1 million characters.

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

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 𡁀
U+A841 # Phags_Pa LETTER KHA 𡁁
U+A842 # Phags_Pa LETTER GA 𡁂

Jiu - the Chinese word for 'alcoholic beverage'
U-label **酒**
Unicode code point is U+9152 (also referred to as: CJK UNIFIED IDEOGRAPH-9152);
A-label is xn--jj4

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). The internationalized portion of a domain name can be in any level – not just the TLD but also the other labels.
  * Email addresses that use non-ASCII characters are called Internationalized Email Addresses. The internationalized portion can be at the local or domain part of the address.

* Since the DNS itself previously only used ASCII, it was necessary to create an additional encoding to allow non-ASCII Unicode code points to be converted into ASCII strings, and vice versa.

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

* Most scripts display characters from left to right when text is presented in horizontal lines.
* However, there are also several scripts, such as Arabic or Hebrew, where the ordering of horizontal text in display is from right to left.
* The text can also be bidirectional (left to right – right to left) when a right-to-left script uses digits that are written from left to right or when it uses embedded words from English or other scripts.
* Challenges and ambiguities can occur when the horizontal direction of the text is not uniform. To solve this issue, there is an algorithm to determine the directionality for bidirectional Unicode text.

Building Blocks: **Bidi Algorithm**

* There is a set of rules that should be applied by the application to produce the correct order at the time of display which are described by the **Unicode Bidirectional Algorithm**. We generally refer to this as the "**Bidi algorithm**".

* To see the Bidi algorithm in detail, go to: http://unicode.org/reports/tr9/tr9-11.html

* Libraries and APIs exist which manage the directionality for the developer, e.g., java.text.Bidi

Building Blocks: **Bidi Algorithm (cont.)**

High-level Synopsis:

1. To know if a sequence is left-to-right or right-to-left, each character in Unicode has an associated directional property.
2. Spaces and punctuation are not strongly typed as either LTR or RTL in Unicode because they may be used in either type of script. They are therefore classified as neutral or weak characters.
3. Neutral characters between characters of the same LTR or RTL directionality will assume the same directionality. Even if there are several neutral characters between the two characters with the same directionality, all the neutral characters will be treated in the same way.
4. When a space or punctuation falls between two typed characters that have different directionality, the neutral character (or characters) will be treated as if they have the same directionality as the prevailing base direction.

Still, it is sometimes necessary to directly encode changes in base direction to ensure text displayed properly.

Building Blocks - **Punycode**

* The algorithm that implements this Unicode-to-ASCII encoding is called Punycode; the output strings are called A-Labels. A-Labels can be distinguished from an ordinary ASCII label because they always start with the following four characters:

  xn--

* These 4 characters are called the ACE prefix (ASCII Compatible Encoding)
* The Punycode transformation is reversible: it can transform from Unicode to an A-Label and also from an A-label back to Unicode (known as a U-Label)

The only RFC-defined use of the Punycode algorithm is for expressing internationalized domains. However, rather than implement Unicode, some developers choose to apply Punycode to other scenarios.

Building Blocks: **Punycode Examples**

* Examples of (imaginary) IDNs

example.**みんな** (Punycode = example.xn--q9jyb4c)
大坂.info (Punycode = xn--uesx7b.info)
みんな. **大坂** (Punycode = xn--q9jyb4c.xn--uesx7b)

Discussion question: Should you store the Punycode locally?

Building Blocks: **Punycode Implementation**

C#

* You can rely on the .NET built-in platform IDN/Punycode support, or use a 3rd party library such as Libidn.

JavaScript

* You can utilize a well-documented punycode.js library, which is bundled with node.js and io.js.

Email Address Internationalization: EAI

Email Address Internationalization (EAI)

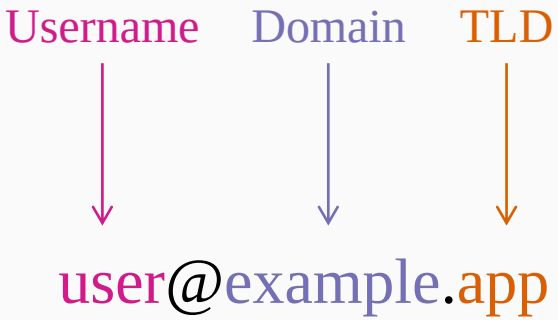
Email addresses contain two parts:

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

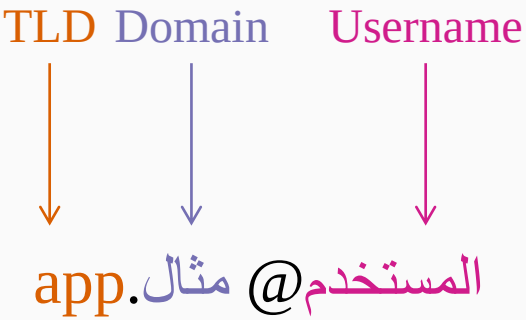
* The domain part can contain *any TLD*, including a new TLD.
* Both portions may be Unicode.

EAI, examples

Left to Right (LTR) Scripts



Right to Left (RTL) Scripts



More Examples of (imaginary) Email Addresses including IDNs

user@example.みんな	(Uses internationalized TLD)
user@大坂.info	(Uses internationalized 2nd level domain)
用戶@example.lawyer	(Uses internationalized user name and new gTLD)

Client Software (MUA – Mail User Agent)

* Display the domain name in Unicode.
* Pass the domain name to the MTA (Mail Transport Agent) in A-Label format (RFC 5890).
* Store and display the Mailbox name in Unicode.
* Follow good practice guides for Linkification within the body of the email (see UASG 010 – Quick Guide to Linkification).
* Follow good practice guides for validation of domain name (see UASG 007 – Introduction to Universal Acceptance).

Server Software (MTA - Mail Transport Agent)

* Confirm EAI-readiness (e.g. advertise SMTPUTF8 support) when making connection to another MTA.
  * If the SMTPUTF8 SMTP extension is not offered by the SMTP server, the SMTPUTF8-aware SMTP client must not transmit an internationalized email address and must not transmit a mail message containing internationalized mail headers as described in RFC 6532 at any level within its MIME structure RFC2045.

POP & IMAP Servers

* Post Office Protocol version 3 (POP3) supports international strings encoded in UTF-8 in usernames, passwords, mail addresses, message headers, and protocol-level text strings (see RFC 6856).
* Internet Message Access Protocol (IMAP) supports UTF-8 encoded international characters in user names, mail addresses, and message headers (see RFC 6855).

Items for Email Service Providers to Consider

* Don’t enforce case-sensitivity of local-part mailbox names.
* Allow the user to enter the email address in any combination of upper-and-lowercase characters so long as the script is correct.
* Don’t issue mailbox names which will duplicate other mailbox names which have the same characters but different cases (e.g. “user@example.TLD”) and “uSer@example.tld”).

Note non-ASCII case folding may not work because it is language specific.

Items for Email Service Providers to Consider

* Offer an all-ASCII mailbox name to the user when they are issued an EAI-compatible mailbox name.
  * If both names alias to the same mailbox (i.e. can be used interchangeably) users will find it easier to initially share addresses with other users who use a different script.
  * Once the ASCII address is initially shared, a user can decide whether to also add the EAI-compatible address to their address book.

Items for Email Service Providers to Consider

* Offer mailbox names which conform to the domain name Label Generation Rules (LGR) for the selected script.
  * Such names are guaranteed to be compatible with the Punycode algorithm.
  * These email addresses can easily be shared by users with their friends and colleagues who do not use their same writing method; the colleague or friend can address email to such an address, or create an address book entry, using the A-label format.
  * Upon use, the client MUA software should convert the A-Label to the appropriate U-Label, at which point the friend or colleague will possess the EAI formatted email address despite not having a keyboard or IME which supports the target script.

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.

- IDNs may display in their Punycoded (A-Label) form.
  - While undesirable, this should not stop messages from being delivered.

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 by applying Punycode to the Mailbox name.
  - Normalize mailbox names in non-ASCII scripts.

Summary

* The internet's technology, including its naming components, are under continual evolution and change. In recent years, a great number of new TLDs with ASCII characters and IDN top-level domains have been released by ICANN. Examples include .nyc, .hsbc, .eco, and .ストア. The languages used on the internet are increasingly non-Latin based. These changes affect the development and maintenance of all internet-enabled apps.
* It is required that your internet-enabled applications, devices and systems accept, validate, store, process and display all domain names and email address appropriately

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 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--”.

ACE prefix - ASCII Compatible Encoding Prefix.

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 requires the use of 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. It is open to any interested individual. The IETF develops Internet Standards and in particular the standards related to the Internet Protocol Suite (TCP/IP).

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

Punycode - It is an algorithm to represent Unicode with the limited character subset of ASCII supported by the Domain Name System. Punycode is intended for the encoding of labels in the Internationalized Domain Names in Applications (IDNA) framework.

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 formal document from the Internet Engineering Task Force (IETF) that is the result of committee drafting and subsequent review by interested parties.

Glossary, partial

Script - The collection of letters or characters used in writing, representing the sounds of 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. Conversions between U-labels and A-labels are performed according to the Punycode specification [RFC3492].

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.