การพัฒนาโปรแกรมให้รองรับ Universal Acceptance

พจนันท์ รัตนไชยพันธ์ และ สมเกียรติ ปุ่ยสูงเนิน

Thailand UA Local Initiative
21 สิงหาคม 2564

Training/Workshop ด้านการยอมรับชื่อโดเมนและอีเมลภาษาไทย

- 7 มีนาคม 2564 การปรับระบบอีเมลให้รองรับการรับส่งชื่ออีเมลภาษาไทย (systems)
- 12 มิถุนายน 2564 การพัฒนาเว็บให้รองรับ UA (web developers)
- 21 สิงหาคม 2564 การพัฒนาโปรแกรมให้รองรับ UA ด้วย Java (Programmer, DevOps)

ติดตามรายละเอียด https://academy.thnic.or.th/webinars-on-eai/

หัวข้อ

- Introduction to Universal Acceptance (20 mins)
- Fundamentals for IDNs and Email Address Internationalization (EAI) (20 mins)
  - Break (5 mins)
- Programming for UA (60 mins)
- Conclusion (5 mins)
- Q&A (10 mins)

Introduction to Universal Acceptance

Universal Acceptance of Domain Names and Email

Goal

All domain names and email addresses work in all software applications.

ชื่อโดเมนและอีเมลทุกชื่อสามารถใช้งานได้กับทุกซอฟต์แวร์แอบพลิเคชัน

Impact

Promote consumer choice, improve competition, and provide broader access to end users.

เพิ่มทางเลือก เกิดการแข่งขัน และเปิดการเข้าถึง [ซอฟต์แวร์แอบพลิเคชัน] ให้แก่ผู้ใช้งาน

Categories of Domain Names and Email Addresses

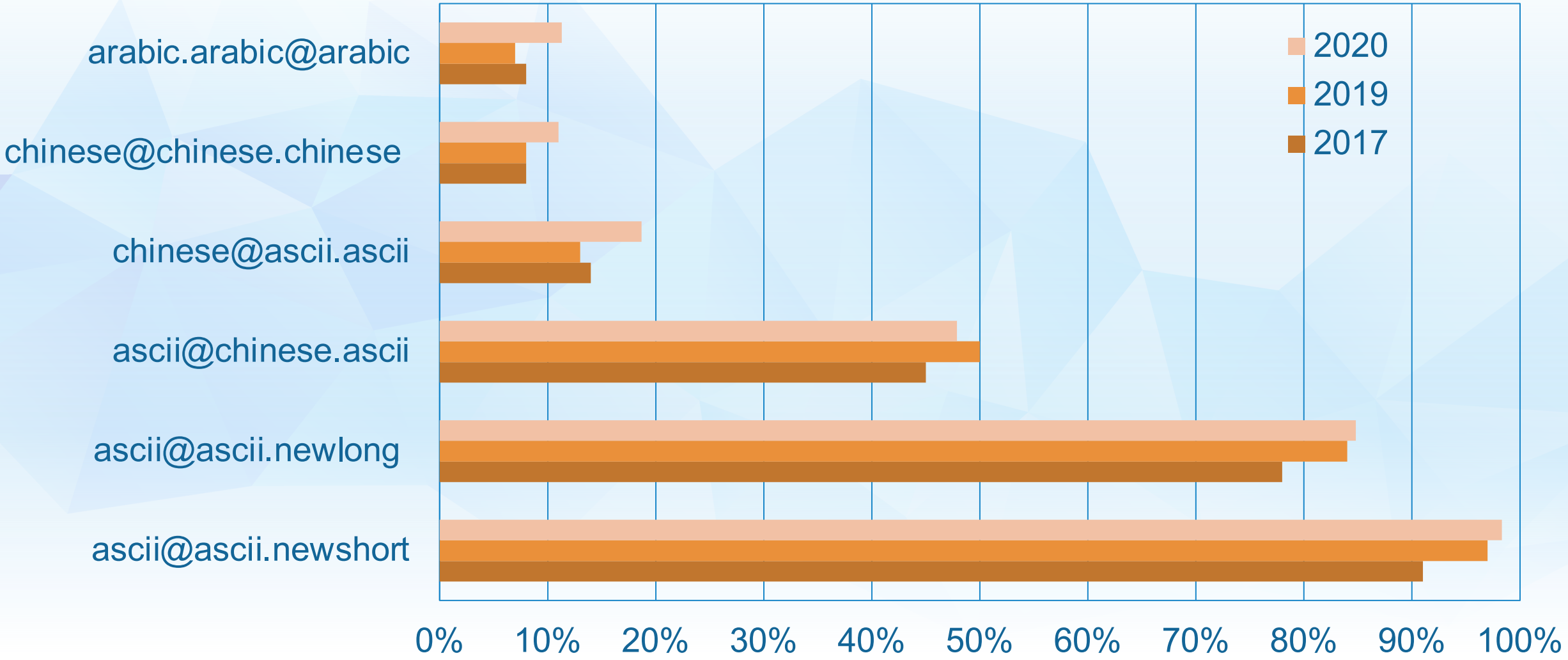
- It's now possible to have domain names and email addresses in local languages using UTF8.
  - Internationalized Domain Names (IDNs)
  - Email Address Internationalization (EAI)

- Domain names
  - **Newer** top-level domain names: example.sky
  - **Longer** top-level domain names: example.abudhabi
  - **Internationalized** Domain Names: รู้จัก.ไทย

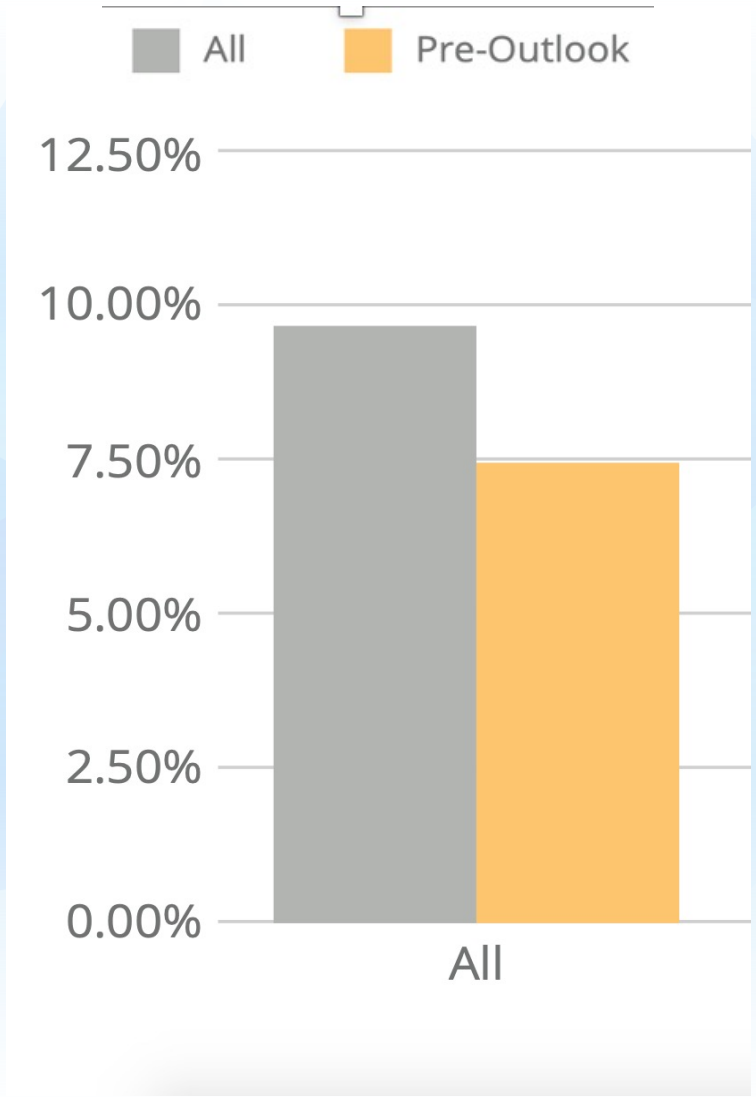
- Internationalized email addresses (EAI)
  - ASCII@IDN: marc@société.org
  - UTF8@ASCII: ईमेल@example.com
  - UTF8@IDN: พจนันท์@คน.ไทย
  - UTF@IDN; right-to-left scripts: ای-میل@مثال.موقع

Acceptance of Email Addresses by Websites Globally

For details, see UASG027



Estimated Support of EAI in Email Systems Under All TLDs



Only 9.7% of the domains sampled were EAI ready in 2019.

This is based on mail servers found through MX records in zones of all top-level domains (TLDs).

For details on methodology, see UASG021D: EAI Readiness in TLDs

Acceptance of Thai Email Addresses by Thai Popular Websites

ปี	จำนวนเว็บที่ทดสอบ	จำนวนเว็บที่รับชื่ออีเมล (A)	จำนวนเว็บที่รับชื่ออีเมลภาษาไทย (B)	จำนวนเว็บที่ตอบกลับชื่ออีเมลภาษาไทย (C)	อัตราการรับชื่ออีเมลภาษาไทย (B)/(A)	อัตราการตอบกลับชื่ออีเมลภาษาไทย (C)/(A)
2561	100	44	4	1	9%	2%
2564	100	43	2	1	5%	2%

(A) จำนวนเว็บที่รับสมัครสมาชิก รับติดต่อผ่านฟอร์ม หรือ ให้สมัคร mailing list ด้วยชื่ออีเมล

(B) จำนวนเว็บที่รับสมัครสมาชิก รับติดต่อผ่านฟอร์ม หรือ ให้สมัคร mailing list ด้วยชื่ออีเมลภาษาไทย

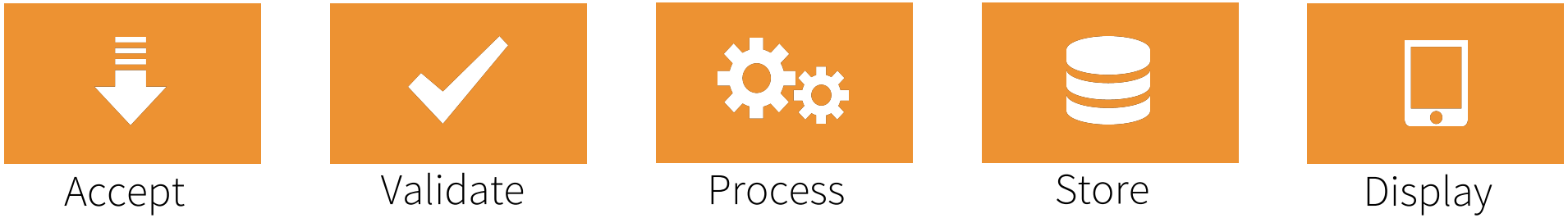
ทดสอบเว็บไซต์ที่เป็นที่นิยมในไทยจากการจัดอันดับเว็บไซต์ของ Alexa.com และ TrueHits

Universal Acceptance of Domain Names and Emails

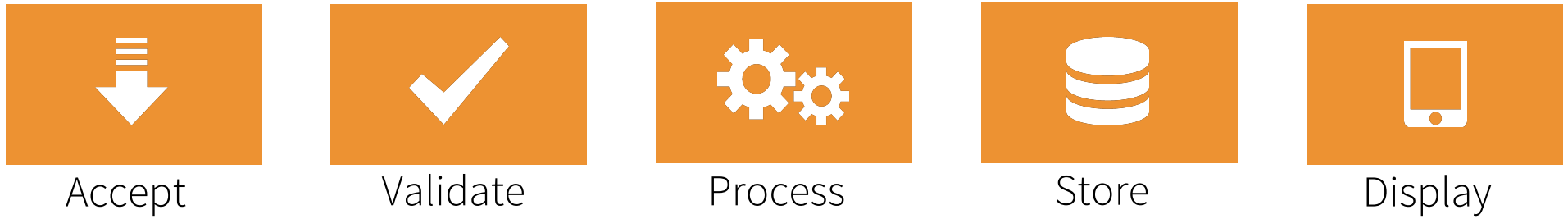
การที่ชื่อโดเมนและชื่ออีเมลทุกชื่อที่ถูกต้องตามกฎเกณฑ์ สามารถใช้งานได้โดยซอฟต์แวร์ประยุกต์ อุปกรณ์ ที่เชื่อมต่อกับอินเทอร์เน็ต ได้อย่างเท่าเทียมกัน

Scope of UA Readiness for Programmers

1. Support All Domain Names



2. Support All Email Addresses



Technology Stack for UA Consideration

Universal Acceptance

Applications and Websites

- Wikipedia.org, ICANN.org, Amazon.com, custom websites globally
- PowerPoint, Google-Docs, Safari, Acrobat, custom apps

Social Media and Search Engines

- Chrome, Bing, Safari, Firefox, local (e.g., Chinese) browsers
- Line, Facebook, Instagram, Twitter, Skype, WeChat, WhatsApp

Programming Languages and Frameworks

- JavaScript, Java, Swift, C#, PHP, Python
- Angular, Spring, .NET core, J2EE, WordPress, SAP, Oracle

Platforms, Operating Systems and Sytem Tools

- iOS, Windows, Linux, Android, App Stores
- Active Directory, OpenLDAP, OpenSSL, Ping, Telnet

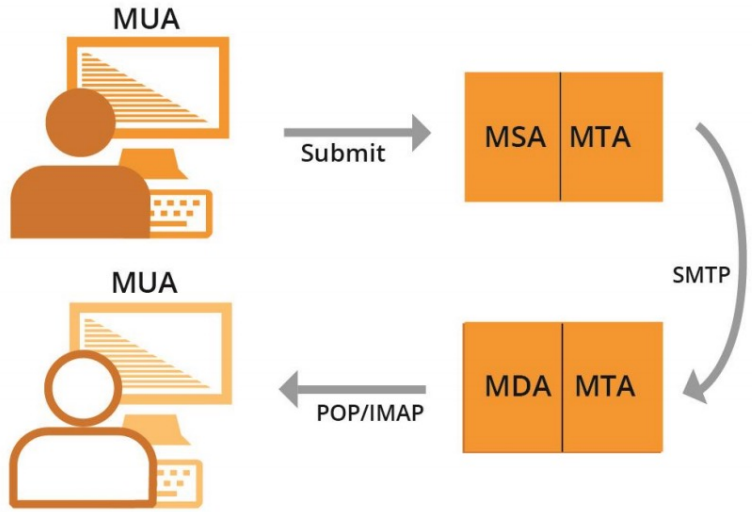
Standards and Best Practices

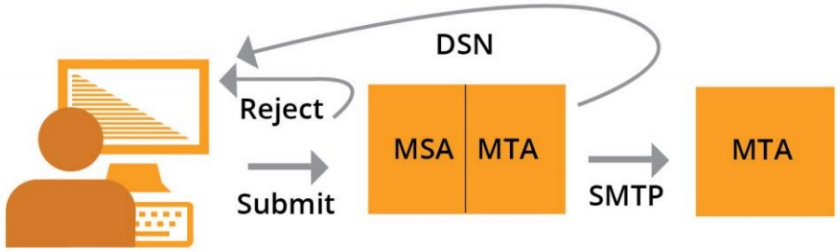
- IETF RFCs, W3C HTML, Unicode CLDR, WHATWG
- Industry-based standards (health, aviation, ...)

Accept, validate, process, store and display all domain names and email addresses.

Email Systems and EAI Support

- All email agents must be configured to send and receive internationalized email addresses. See EAI: A Technical Overview for details.
  - MUA – Mail User Agent: A client program that a person uses to send, receive, and manage mail.
  - MSA – Mail Submission Agent: A server program that receives mail from a MUA and prepares it for transmission and delivery.
  - MTA – Mail Transmission Agent: A server program that sends and receives mail to and from other Internet hosts. An MTA may receive mail from an MSA and/or deliver mail to an MDA.
  - MDA – Mail Delivery Agent: A server program that handles incoming mail and typically stores it in a mailbox or folder.





Quiz

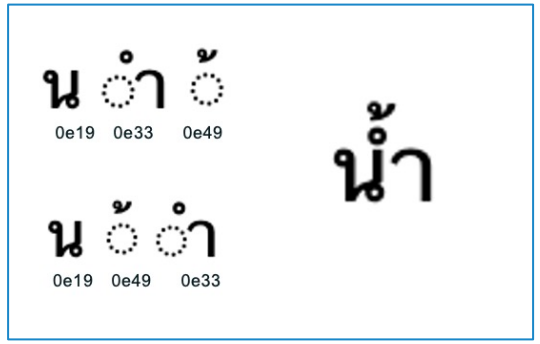
Quiz 1 Question

- To enhance systems to be Universal Acceptance (UA) ready, which of the following categories of domain names and email addresses are relevant?
  1. ASCII domain names.
  2. Internationalized Domain Names (IDNs).
  3. Internationalized email addresses (EAI).
  4. All the above.
  5. Only 2 and 3.

Fundamentals for Internationalized Domain Names and Email Addresses

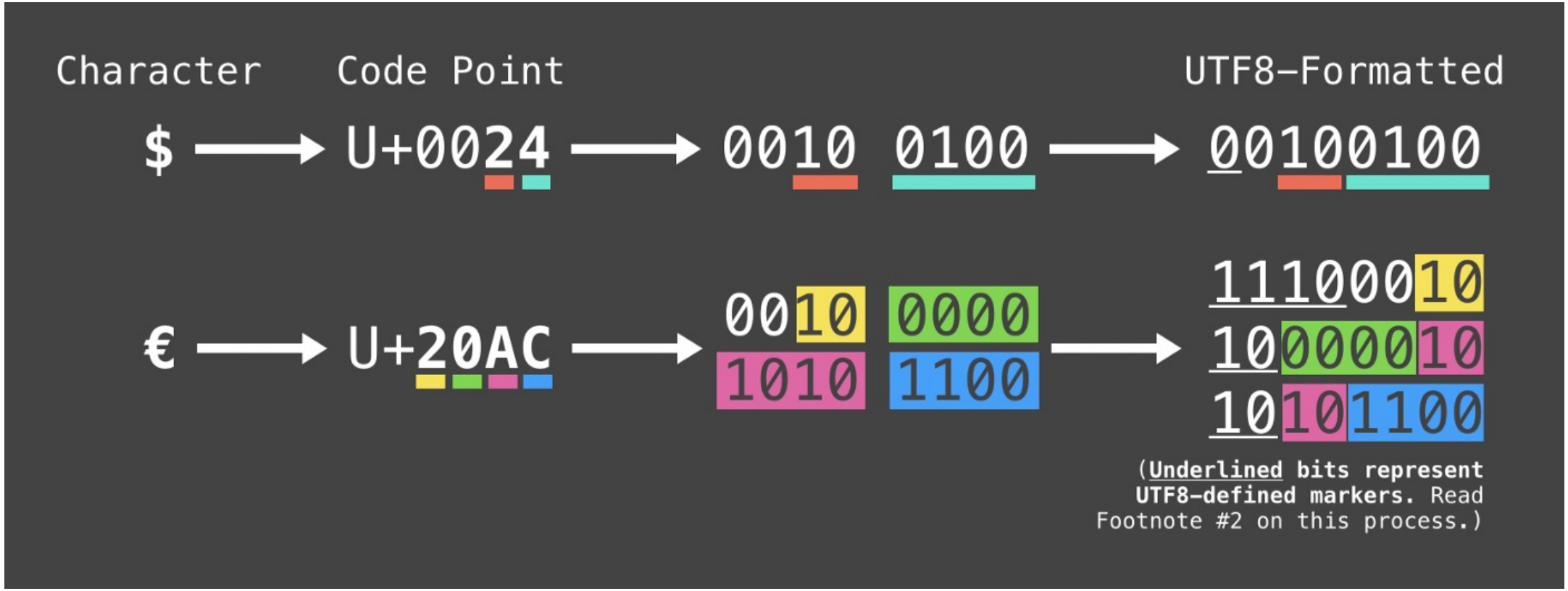
Unicode Encoding

- Unicode encodes glyphs into codepoints for different scripts of the world.
  - Codepoints shown in hex using the U+XXXX notation.
  - Unicode files typically in UTF8 format, using a variable number of bytes for a codepoint.
  - ASCII is used as is in Unicode: e = ASCII 65 = U+0065.
- There are multiple ways to encode certain glyphs in Unicode:
  - è = U+00E8
  - e + ` = è = U+0065 + U+0300



- Normalization ensures that the end representation is the same, even if users type differently.
  - IDN standards recommend using Normalization Form C (NFC).
  - Generates U+00E8 for both input versions above.

Unicode and UTF8



Domain Names and Internationalized Domain Names (IDNs)

- A domain name is an ordered set of labels or strings: example.co.th
  - The top-level domain (TLD) is the rightmost label: ”th”
  - Initially, TLDs were only two or three characters long (e.g., .th, .com).
  - Now TLDs can be longer strings (e.g., .info, .google, .engineering).
  - TLDs delegated in the root zone can change over time, so a fixed list can get outdated.
- Domain names can also be internationalized when one of the labels contains at least one non-ASCII character.
  - For example: exâmple.com or รู้จัก.ไทย
- Use the latest IDN standard called IDNA2008 for IDNs.
  - Do not use libraries for the outdated IDNA2003 version.

IDNs and EAI

- There are two equivalent forms of IDN domain labels: U-label and A-label.
  - Human users use the IDN version called U-label (using UTF-8 format): exâmple
  - Applications or systems internally use an ASCII equivalent called A-label:
    1. Take user input and normalize and check against IDNA2008 to form IDN U-label.
    2. Convert U-label to punycode (using RFC3492).
    3. Add the “xn--” prefix is added to identify the ASCII string as an IDN A-label.
       - exâmple => exmple-xta => xn--exmple-xta
       - 普遍接受-测试 => --f38am99bqvcd5liy1cxsg => xn----f38am99bqvcd5liy1cxsg
       - รู้จัก => 12cn4frcvb5f => xn--12cn4frcvb5f

- Email address syntax: mailboxName@domainName
  - EAI has the mailboxName in Unicode (in UTF8 format).
  - The domainName can be ASCII or IDN.
    - For example: kévin@example.org or พจนันท์@คน.ไทย.

หลักเกณฑ์การตั้งชื่อโดเมน .th

3.1. ชื่อโดเมน .th แต่ละชื่อจะต้องประกอบด้วยตัวอักษรภาษาอังกฤษ (a-z) และ/หรือ ตัวเลขอารบิค (0-9) และชื่อโดเมนต้องมีความยาวอย่างน้อย 1 ตัว แต่ไม่เกิน 63 ตัว ทั้งนี้ ชื่อโดเมนที่มีความยาว 1 ตัว จะอยู่ภายใต้การจดทะเบียนที่มีข้อกำหนดและเงื่อนไขเฉพาะ ซึ่งมูลนิธิฯ สงวนสิทธิในการพิจารณาเป็นกรณีไป

3.2. ชื่อโดเมนสามารถประกอบด้วยเครื่องหมายยัติภังค์ “-” คั่นระหว่างตัวอักษรหรือตัวเลข ได้ แต่ต้องไม่ใช้เรียงติดกันมากกว่า 1 ตัว นอกจากนี้ ไม่อนุญาตให้ใช้เป็นตัวอักษรแรก หรือตัวอักษรสุดท้ายของชื่อโดเมน

ชื่อโดเมน .th และ .ไทย

หลักเกณฑ์การตั้งชื่อโดเมน .ไทย

4.1. ชื่อโดเมน .ไทย แต่ละชื่อจะต้องประกอบด้วยตัวอักษรภาษาไทย พยัญชนะ(ก-ฮ ฤ ฤๅ ฦ ฦๅ) สระ วรรณยุกต์ ยมก(ๆ) พินทุ(อ) นฤคหิต(อำ) ไปยาลน้อย(ฯ) การันต์(์) และ/หรือ เลขไทย(๐-๙) โดยชื่อโดเมนจะต้องมีความยาวหลังจากแปลงเป็นพิวนีโค้ด(Punycode) ไม่เกิน 63 ตัว

4.2. ชื่อโดเมน .ไทย ประกอบด้วยตัวเลขอารบิก(0-9) ได้ แต่จะต้องประกอบกับตัวอักษรภาษาไทยที่กล่าวไว้ข้างต้นอย่างน้อย 1 ตัว

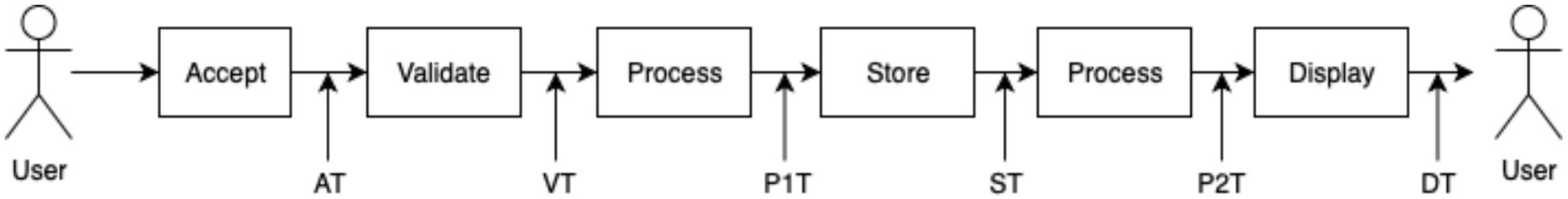
4.3. ชื่อโดเมนประกอบด้วยเครื่องหมายยัติภังค์ “-” คั่นระหว่างตัวอักษรหรือตัวเลขได้ แต่ต้องไม่ใช้เรียงติดกันมากกว่า 1 ตัว นอกจากนี้ ไม่อนุญาตให้ใช้เป็นตัวอักษรแรกหรือตัวอักษรสุดท้ายของชื่อ

Universal Acceptance Issues

- Some applications are still verifying domain names incorrectly by using one of the outdated methods:
  - Check for a fixed length of TLD between 2-4 characters (TLD can be up to 63 characters).
  - Check from a fixed set of TLDs, e.g., using static list of strings.
  - Check for only ASCII characters.
- Some applications do not cater to additional requirements for validating EAI:
  - Check mailbox name to be a valid string in UTF-8 format.
  - DomainName can be ASCII or IDN.

UA Readiness Testing Framework

- Based on UASG026, the application components can be generalized to put emphasis on the processing of internationalized identifiers.
- Each gate has its own set of requirements and processing.



- AT: Accept test
- VT: Validate test
- P1T: Process test on the input
- ST: Store test
- P2T: Process test on the output
- DT: Display test

Validating User Input

- Validating user input, or any input, is extremely useful for various reasons, some of which include: a better user experience, increased security, and avoiding irrelevant issues.
- Validating domain names and email addresses is useful.
- Some validation methods for domain names and email addresses:
  - Basic syntax checks: is the syntax of the string correct?
    - Does the domain name contain '.' ?
    - Does the email address contain '@' and a valid domain name part?
  - Functional checks: does the domain name or email address work?
    - Is the top-level domain (TLD) in use?
    - Is the whole domain name in use?
    - Is the email in use?

Validating Domain Name

- Validating syntax:
  - ASCII: RFC1035
    - Composed of letters, digits, and hyphen.
    - Max length is 255 octets with each label up to 63 octets.
  - IDN: IDNA2008 (RFCs 5890-5894)
    - Valid A-labels
    - Valid U-labels
- Validating function:
  - Is the top-level domain (TLD) in use?
    - Verify against the list of TLDs.
    - Verify using a DNS request.
  - Is the whole domain name in use?
    - Verify using a DNS request.

Resolving Domain Name

- After validation, a software would then use the domain name identifier as:
  - A domain name to be resolved in the DNS.
- Therefore, to be UA compliant, the software has to use proper methods that support UA.
  - For example, passing a U-Label to the traditional functions or methods may not succeed, as it is not expecting a UTF8 domain name.

Validating Email Address

- An email address is composed of: mailboxName@domainName
- Validating syntax:
  - For domainName, see earlier discussion.
  - For mailboxName:
    - ASCII
    - UTF8 (for EAI)
- Validating function:
  - Is the domain name set up to send and receive emails?
  - Is the mailbox name able to send and receive emails?

Sending Email

- After validation, a software would then use the email identifier as:
  - An email-address based user id.
  - To send an email.
- Therefore, to be UA compliant, the software must use proper methods that support UA.
  - For example, passing an UTF8 mailbox name email address to a mail sender may not succeed, as it is not expecting a UTF8 mailbox name in the email address.

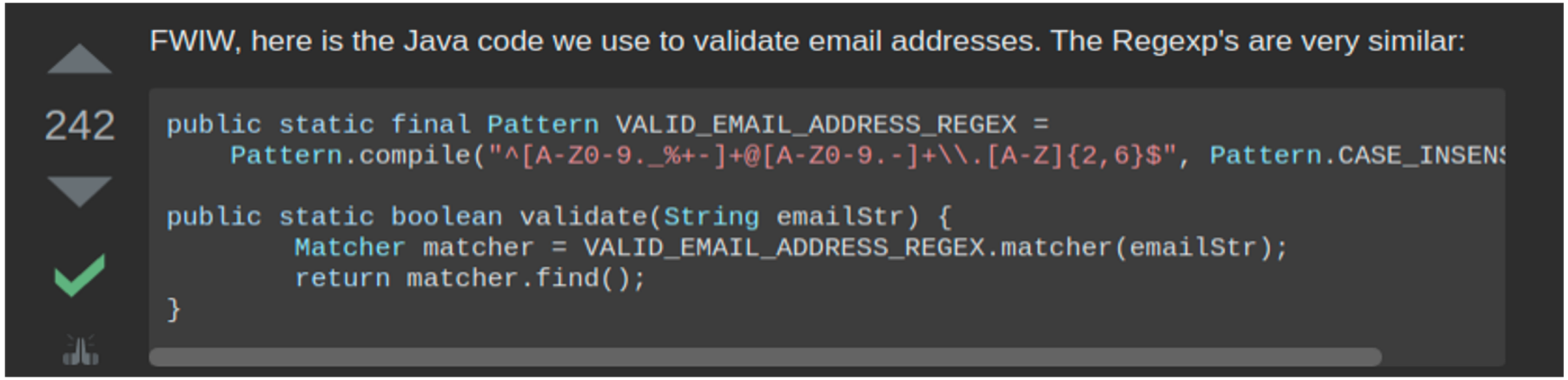
Sending Email

- A comprehensive list of UA test cases is documented in UASG004.
- Developers are strongly encouraged to use these test cases in its unit and system testing.

Quiz

Quiz 2: A Real Example

- A company built a website where international consumers can subscribe via their email. Since the subscription form is user input, developers validated the email address before trying to send the email.
  - Developers went to *Stackoverflow* and found a regular expression (regex) to perform the validation:



  - The regex limits mailbox to letters A-Z, digits 0-9, and some symbols, the domain labels to letters, digits and hyphen, and the top-level domain to letters only with length 2-6.

- Would this regex work for the company's website? Why or why not?

เอกสารที่เกี่ยวข้อง

- See https://uasg.tech for a complete list of reports.
  - Universal Acceptance Quick Guide: UASG005
  - Introduction to Universal Acceptance: UASG007
  - Quick Guide to EAI: UASG014
  - EAI – A Technical Overview: UASG012
  - EAI – Evaluation of Major Email Software and Services: UASG021B
  - Universal Acceptance Readiness Framework: UASG026
  - Considerations for Naming Internationalized Email Mailboxes: UASG028
  - Evaluation of EAI Support in Email Software and Services Report: UASG030

Programming for UA

Get Involved!

มีส่วนร่วมกับ Thailand UA Local Initiative และ UASG

- Visit Thai UA local initiative web page at รู้จัก.ไทย/ua/thailand-initiative
- Subscribe APAC EAI Implementers' Group for technical support (by THNIC)
- Access UA technical content by THNIC at: https://wiki.thnic.or.th
- Access all UASG documents and presentations at: https://uasg.tech
- Access details of ongoing work from wiki pages: https://community.icann.org/display/TUA
- Register to participate or listen in the UA discussion list at: https://uasg.tech/subscribe

Engage with THNIC

Webs site: thnic.or.th

facebook.com/THNIC.Foundation/

twitter@thnicf

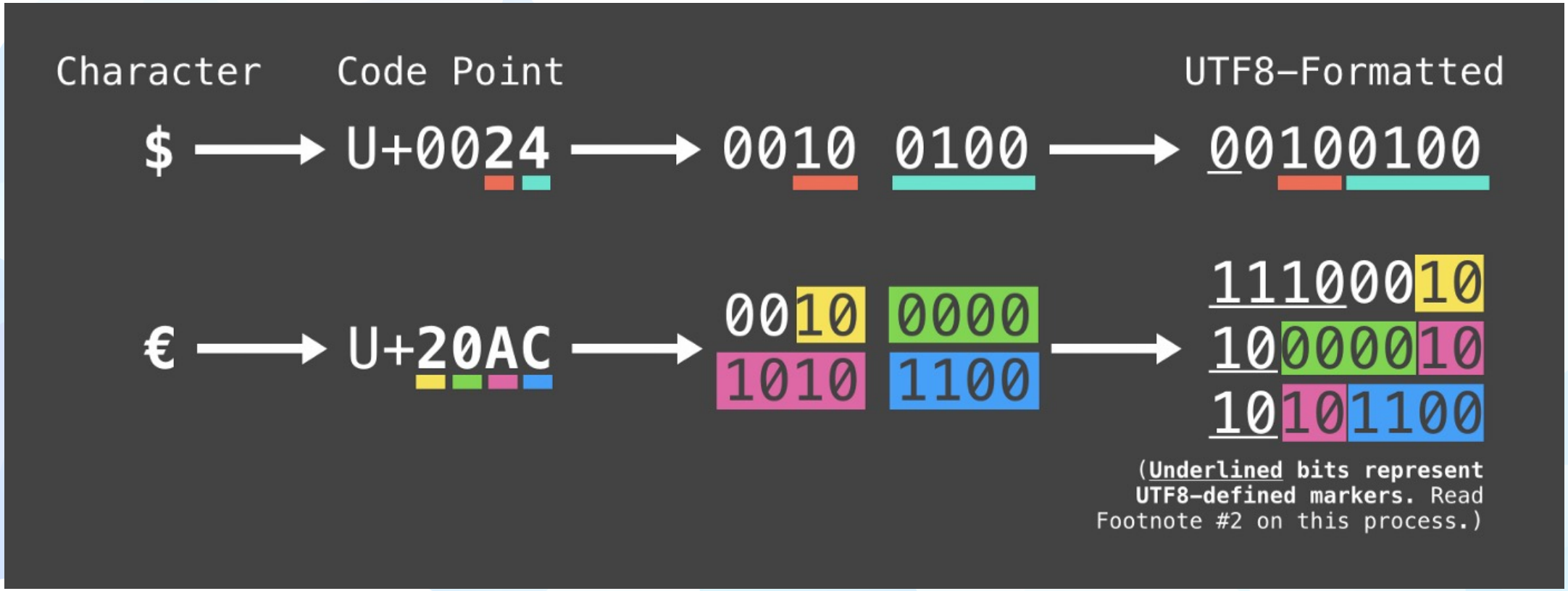
youtube.com/user/thnicf

ASCII

ASCII		Macintosh or Windows
Dec	Hex	Result
96	60	`
97	61	a
98	62	b
99	63	c
100	64	d
101	65	e
102	66	f
103	67	g
104	68	h
105	69	i
106	6A	j
107	6B	k
108	6C	l
109	6D	m
110	6E	n
111	6F	o

ASCII		Macintosh or Windows
Dec	Hex	Result
112	70	p
113	71	q
114	72	r
115	73	s
116	74	t
117	75	u
118	76	v
119	77	w
120	78	x
121	79	y
122	7A	z
123	7B	{
124	7C	|
125	7D	}
126	7E	~
127	7F	DEL

Unicode


Character
Code Point
UTF8-Formatted
$ ⟶ U+0024 ⟶ 0010 0100 ⟶ 00100100
€ ⟶ U+20AC ⟶ 0010 0000 1010 1100 ⟶ 11100010 10000010 10101100
(Underlined bits represent UTF8-defined markers. Read Footnote #2 on this process.)

Number of bytes	Bits for code point	First code point	Last code point	Byte 1	Byte 2	Byte 3	Byte 4
1	7	U+0000	U+007F	0xxxxxxx			
2	11	U+0080	U+07FF	110xxxxx	10xxxxxx		
3	16	U+0800	U+FFFF	1110xxxx	10xxxxxx	10xxxxxx	
4	21	U+10000	U+10FFFF	11110xxx	10xxxxxx	10xxxxxx	10xxxxxx

Thai Popular Web Sites

List	Rank	Site
AL50	1	google.com
AL50	2	youtube.com
AL50	3	google.co.th
AL50	4	pantip.com
AL50	5	lazada.co.th
AL50	6	line.me
AL50	7	facebook.com
AL50	8	blogspot.com
AL50	9	wikipedia.org
AL50	10	sanook.com
AL50	11	live.com
AL50	12	shopee.co.th
AL50	13	yahoo.com
AL50	14	kapook.com
AL50	15	imovie-hd.com
AL50	16	037hdd.com
AL50	17	roblox.com
AL50	18	wordpress.com
AL50	19	mthai.com

List	Rank	Site
AL50	21	netflix.com
AL50	22	movie2uhd.com
AL50	23	anime-sugoi.com
AL50	24	mgronline.com
AL50	25	trueid.net
AL50	26	hao123.com
AL50	27	doomovie-hd.com
AL50	28	trueplookpanya.com
AL50	29	dek-d.com
AL50	30	dltv.ac.th
AL50	31	y8.com
AL50	32	viu.com
AL50	33	siamsport.co.th
AL50	34	moph.go.th
AL50	35	bopp-obec.info
AL50	36	moe.go.th
AL50	37	pornhub.com
AL50	38	msn.com
AL50	39	aliexpress.com
AL50	40	khaosod.co.th

List	Rank	Site
TH100	1	sanook.com
TH100	2	thairath.co.th
TH100	3	khaosod.co.th
TH100	4	kapook.com
TH100	5	dek-d.com
TH100	6	mgronline.com
TH100	7	siamsport.co.th
TH100	8	matichon.co.th
TH100	9	dailynews.co.th
TH100	10	bugaboo.tv
TH100	11	priceza.com
TH100	12	ais.co.th
TH100	13	posttoday.com
TH100	14	wongnai.com
TH100	15	mello.me
TH100	16	longdo.com
TH100	17	ch3thailand.com
TH100	18	soccersuck.com
TH100	19	ch7.com
TH100	20	thaiware.com

Unicode Normalization Forms

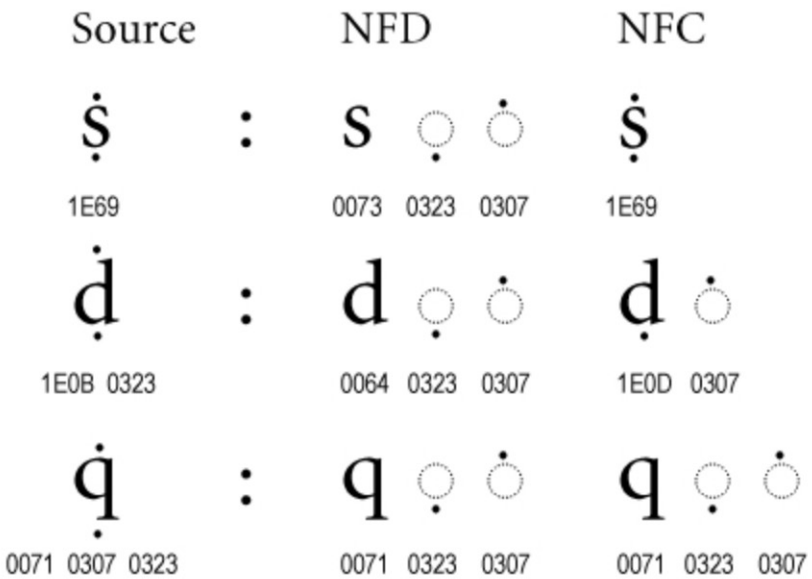
- Unicoe Standard Annex #15

Table 1. Normalization Forms

Form	Description
Normalization Form D (NFD)	Canonical Decomposition
Normalization Form C (NFC)	Canonical Decomposition, followed by Canonical Composition
Normalization Form KD (NFKD)	Compatibility Decomposition
Normalization Form KC (NFKC)	Compatibility Decomposition, followed by Canonical Composition

Figure 4. Canonical Composites

Figure 5. Multiple Combining Marks



0e19 0e33 0e49

0e19 0e49 0e33