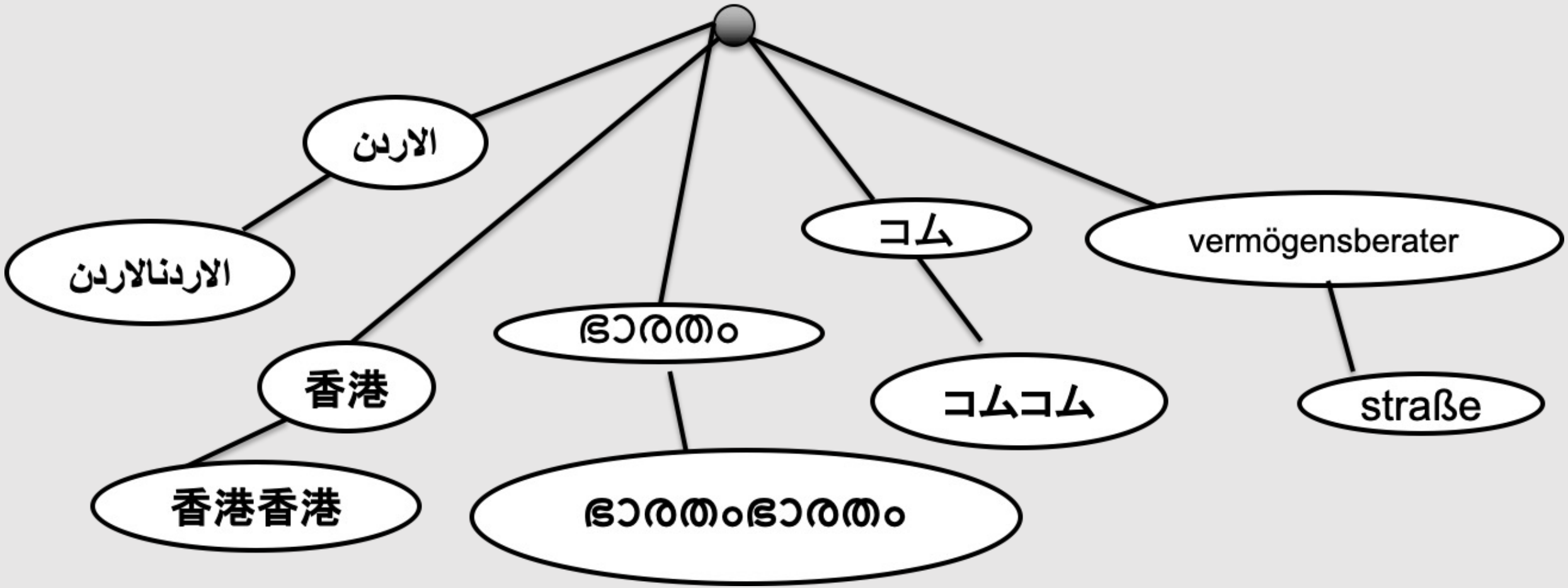

الاردن
الاردنالاردن
香港
香港香港
ᒪᑐᓂᐠ
ᒪᑐᓂᐠᒪᑐᓂᐠ
コム
コムコム
vermögensberater
straße

INTERNET STIFTELSEN

Internets domännamnssystem*

Föreläsning FL12, VT 2024

Mats Dufberg

* Se “Internets domännamnssystem”

Innehåll

▶Tecken i klassiska domännamn

[Innehåll]

Tecken i domännamn

DNS-paketen har oktetter (koder om 8 bitar) i namnsträngarna, men som tecken är det bara ASCII (7 bitar).

Teckenjämförelsen är skiftlägesokänslig (”case insensitive”), d.v.s.

”a” == ”A”

”b” == ”B”

o.s.v.

Denna bild och några bilder till kommer från FL02, något modifierade.

Teckenuppsättningen ASCII (decimalt)

0	nul	1	soh	2	stx	3	etx	4	eot	5	enq	6	ack	7	bel
8	bs	9	ht	10	nl	11	vt	12	np	13	cr	14	so	15	si
16	dle	17	dc1	18	dc2	19	dc3	20	dc4	21	nak	22	syn	23	etb
24	can	25	em	26	sub	27	esc	28	fs	29	gs	30	rs	31	us
32	sp	33	!	34	"	35	#	36	$	37	%	38	&	39	'
40	(	41	)	42	*	43	+	44	,	45	-	46	.	47	/
48	0	49	1	50	2	51	3	52	4	53	5	54	6	55	7
56	8	57	9	58	:	59	;	60	<	61	=	62	>	63	?
64	@	65	A	66	B	67	C	68	D	69	E	70	F	71	G
72	H	73	I	74	J	75	K	76	L	77	M	78	N	79	O
80	P	81	Q	82	R	83	S	84	T	85	U	86	V	87	W
88	X	89	Y	90	Z	91	[	92	\	93	]	94	^	95	_
96	`	97	a	98	b	99	c	100	d	101	e	102	f	103	g
104	h	105	i	106	j	107	k	108	l	109	m	110	n	111	o
112	p	113	q	114	r	115	s	116	t	117	u	118	v	119	w
120	x	121	y	122	z	123	{	124	|	125	}	126	~	127	del

0-31: Styrkoder (inga skrivbara tecken)
32: Mellanslag (skrivbart tecken)
33-126: Skrivbara tecken

127: Styrkod (inget skrivbart tecken)
128-255: Odefinierade (åttonde biten satt) – ej ASCII

Tecken i domännamn

”Hostname” är ytterligare begränsat till (gäller label):

a-z
A-Z
0-9
-

där ”-” inte får stå först eller sist.

-www.kth.se – ej OK
www-.kth.se – ej OK
w-w-w.kth.se – OK

Tecken i domännamn

Domännamn som ***inte*** är hostname kan ha understreck ”_” i domännamnet.

För vissa posttyper eller användningar så är det regel att ha inledande ”_” på speciella ”labels”:

```
_sip._udp.kth.se.           SRV    0 0 5060 sip1.sys.kth.se.
_25._tcp.mx1.iis.se.        TLSA   3 1 1 (
                                   3EC9DC5D031807738EF1CCF91D5990AA9BC110D9F250
                                   2DBCC44FCA8D80E18426 )
_acme-challenge.namn.se.    TXT    "FgQqzwRV2XNPxQPqrPF8zK4bmmyL83Pdf40aZDWbw-w”
```

Punkt i domännamn

Punkten ”.” i domännamn markerar (normalt) gränsen mellan ”labels”. I DNS-paketet så finns den inte med, utan ”labels” är definierade var för sig (RFC 1035).

Punkt i domännamn

I undantagsfall så kan en ”label” innehålla en punkt. I RNAME i SOA-posten (som representerar en mailadress) så kan första ”label” innehålla en eller flera punkter (se FL02).

Kanske enda spridda undantaget när det gäller punkt i label.

▶Nya domännamn och IDN

[Innehåll]

Krav på utökad teckenuppsättning

Det har länge funnits krav och önskemål om att kunna använda andra tecken i domännamn, speciellt från talare av språk vars skrivsystem inte är baserat på det latinska alfabetet. Bokstäverna i ASCII kommer från det latinska alfabetet.

Arbetet med att ta fram en lösning började i mitten av 1990-talet.

Det har funnits flera förslag på utökning, men för att kunna ha ett sammanhållet DNS-träd så behövs en gemensam teckenuppsättning med alla använda skrivtecken för olika språk.

Nya domännamn

Det dyker upp domännamn som faktiskt bryter mot domännamnsreglerna som vi har hittills beskrivit dem.

räksmörgås.se *www.språkförsvaret.se* *råttgift.se*

*villaägarna.se** *lägenhetsbyte.se** *www.lidingöloppet.se**

*www.länsförsäkringar.se** *malmö.se** *umeå.se**

*) Webbservern skriver om till domän utan ÅÄÖ.

IDN

Utökningen som ger de nya domännamnen heter IDN, *Internationalized domain name*

▶Teckenuppsättningar och Unicode

[Innehåll]

Äldre teckenuppsättningar

ASCII är en 7-bitars teckenuppsättning som användes tidigt. Inom ASCII rymdes bara bokstäverna A-Z (plus lite andra tecken).

Det fanns modifieringar av ASCII t.ex. för svenska datorer där []{}|\ byttes ut mot åääöÅÄÖ.

Sådana modifierade teckenuppsättningar kunde då inte användas för olika språk som använde andra bokstäver.

Det blev inte entydigt eftersom det var samma koder som tillskrevs olika tecken.

Äldre teckenuppsättningar

Senare kom 8-bitars teckenuppsättningar som då fick plats för dubbelt så många tecken (teoretiskt 256). Normalt var den nedre halvan identisk med ASCII.

Inte ens för alfabeten som är baserade på det latinska alfabetet (som svenska, tyska, spanska m.fl. alfabeten) så räcker det med ***en*** sådan uppsättning, utan det finns flera.

Det fanns heller inte något entydigt sätt att veta vilken som hade använts.

Äldre teckenuppsättningar

ISO 8859-1 var den teckenuppsättning som användes mest i Sverige tills UTF-8 kom. ISO 8859-1 är en 8-bitars teckenuppsättning. Windows har en variant med något fler tecken, men fortfarande 8 bitar.

ISO 8859-1 innehåller bl.a. ÅÄÆÖØ, men t.ex. inte Ł som används för polska.

För polska kan man använda ISO 8859-2, som t.ex. inte innehåller ÅÆØ.

Äldre teckenuppsättningar

När vi lägger till andra alfabeten, t.ex. grekiska och kyrilliska så krävs det ännu fler 8-bitars teckenuppsättningar.

Äldre teckenuppsättningar

Det finns tusentals kinesiska tecknen så det är bortom vad som ryms med 8 bitar.

Unicode

Unicode är en teckenuppsättning som innehåller de flesta tecken för skriven text. Första versionen av Unicode kom i början av 1990-talet men det tog 10-15 år innan den slog igenom.

Målsättningen är att just att täcka allt som kan tänkas tryckas i löpande text, inte bara bokstäver.

Istället för att ha många små teckenuppsättningar, så är Unicode *en* samlad teckenuppsättning.

Unicode

Unicode tog alla befintliga teckenuppsättningar och inkorporerade dem på ett strukturerat sätt.

ASCII är speciellt eftersom alla datorsystem sedan många år klarar av att hantera ASCII. I starten av Unicode ligger ASCII inkopierat precis som den är.

Unicode

Unicode består f.n. av ca 130.000 tecken, men utökas årligen med nya tecken.

Istället för 8 bitar med rum för teoretiskt 256 tecken så är ryms det upp till drygt 1 miljon tecken i Unicodes teckenrymd (drygt 20 bitar).

Koder: 0x0—0x10FFFF (hexadecimalt)

https://www.unicode.org/

https://codepoints.net/

Unicode

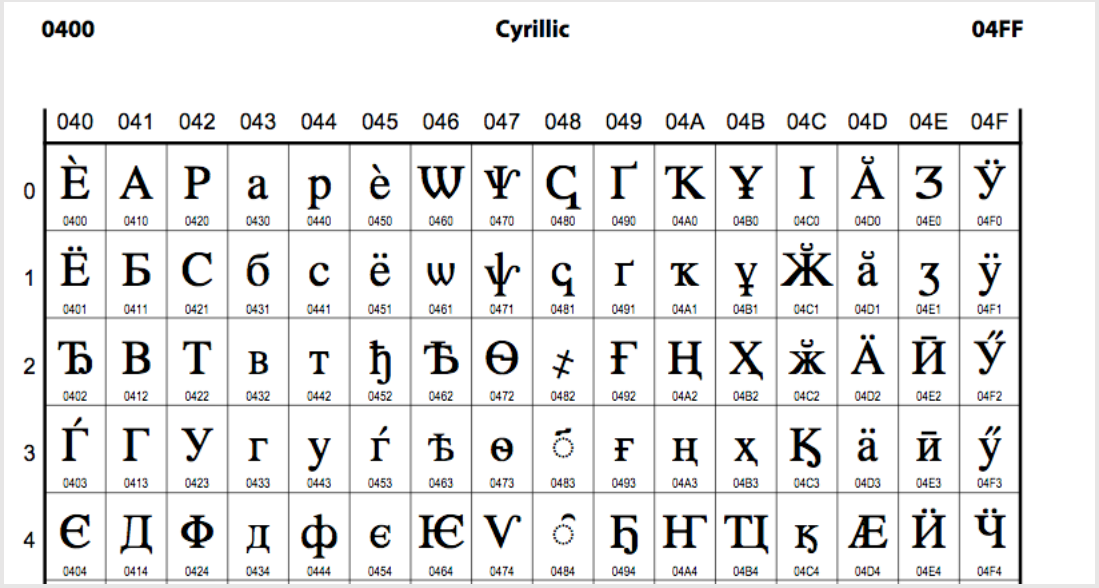
Unicode innehåller olika tecken som kan förekomma i olika typer av texter.

Unicode är indelad olika uppsättningar, ”scripts”.

De flesta ”scripts” motsvarar ”alfabeten” i vid bemärkelse och används för att skriva språk.

Andra innehåller istället matematiska tecken, emoji eller annat.

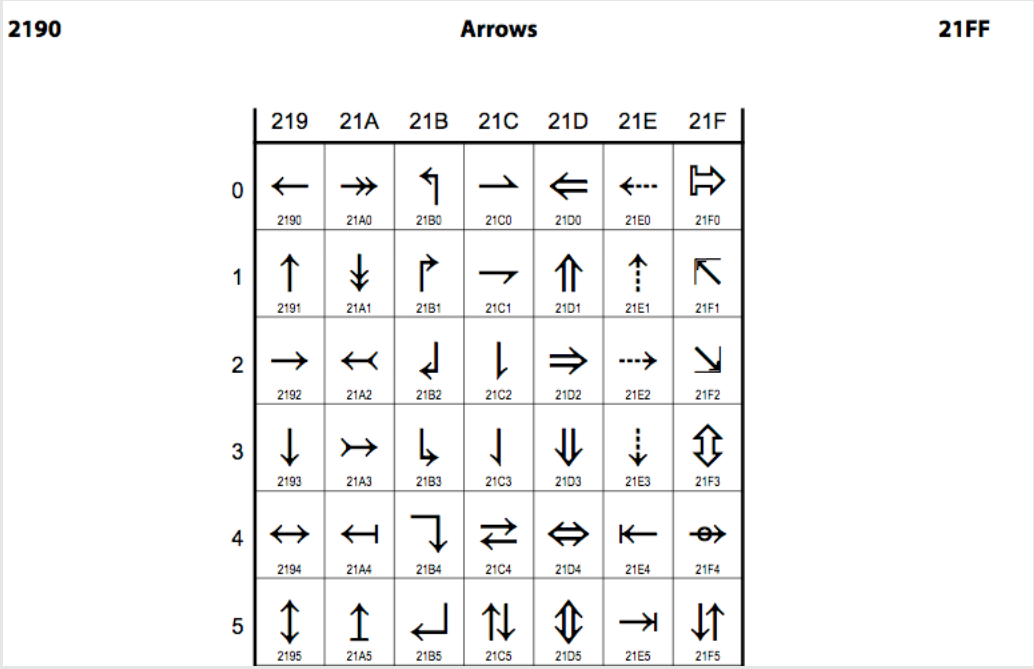
Kyrilliska bokstäver



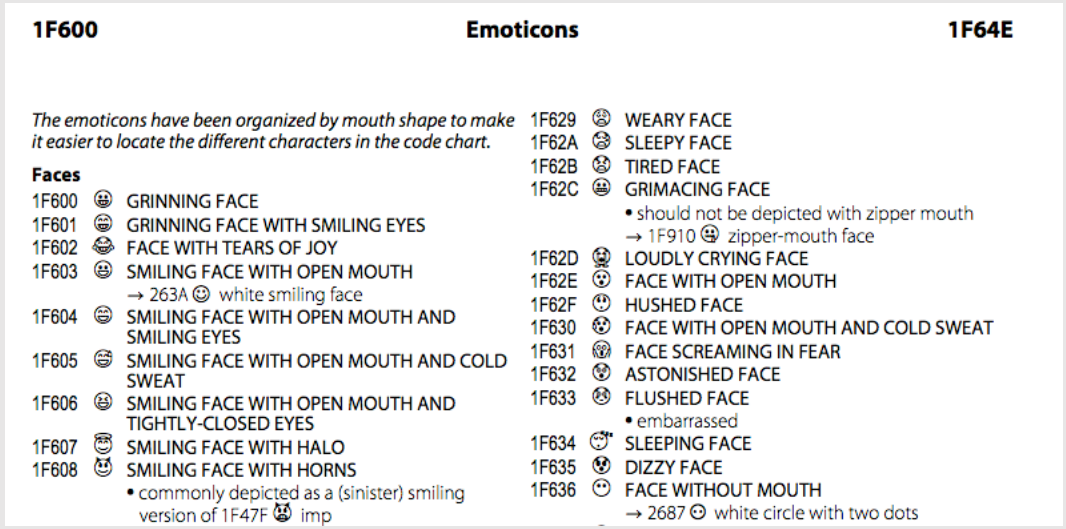
0400 **Cyrillic** **04FF**

	040	041	042	043	044	045	046	047	048	049	04A	04B	04C	04D	04E	04F
0	Ѐ 0400	А 0410	Р 0420	а 0430	р 0440	ѐ 0450	Ѡ 0460	Ѱ 0470	Ҁ 0480	Ґ 0490	Қ 04A0	Ұ 04B0	Ӏ 04C0	Ӑ 04D0	Ӡ 04E0	Ӱ 04F0
1	Ё 0401	Б 0411	С 0421	б 0431	с 0441	ё 0451	ѡ 0461	ѱ 0471	ҁ 0481	ґ 0491	қ 04A1	ұ 04B1	Ӂ 04C1	ӑ 04D1	ӡ 04E1	ӱ 04F1
2	Ђ 0402	В 0412	Т 0422	в 0432	т 0442	ђ 0452	Ѣ 0462	Ѳ 0472	≠ 0482	Ғ 0492	Ҋ 04A2	Ҳ 04B2	ӂ 04C2	Ӓ 04D2	Ӣ 04E2	Ӳ 04F2
3	Ѓ 0403	Г 0413	У 0423	г 0433	у 0443	ѓ 0453	ѣ 0463	ѳ 0473	◌҃ 0483	ғ 0493	ҋ 04A3	ҳ 04B3	Ҍ 04C3	ӓ 04D3	ӣ 04E3	ӳ 04F3
4	Є 0404	Д 0414	Ф 0424	д 0434	ф 0444	є 0454	Ѥ 0464	Ѵ 0474	◌҄ 0484	Ҕ 0494	Ҍ 04A4	Ҵ 04B4	ҍ 04C4	Ӕ 04D4	Ӥ 04E4	Ӵ 04F4

Pilar


2190
Arrows
21FF
219
21A
21B
21C
21D
21E
21F
0
1
2
3
4
5
2190
21A0
21B0
21C0
21D0
21E0
21F0
2191
21A1
21B1
21C1
21D1
21E1
21F1
2192
21A2
21B2
21C2
21D2
21E2
21F2
2193
21A3
21B3
21C3
21D3
21E3
21F3
2194
21A4
21B4
21C4
21D4
21E4
21F4
2195
21A5
21B5
21C5
21D5
21E5
21F5

Emoji



1F600 **Emoticons** **1F64E**

The emoticons have been organized by mouth shape to make it easier to locate the different characters in the code chart.

Faces

1F600 GRINNING FACE
1F601 GRINNING FACE WITH SMILING EYES
1F602 FACE WITH TEARS OF JOY
1F603 SMILING FACE WITH OPEN MOUTH
→ 263A white smiling face
1F604 SMILING FACE WITH OPEN MOUTH AND SMILING EYES
1F605 SMILING FACE WITH OPEN MOUTH AND COLD SWEAT
1F606 SMILING FACE WITH OPEN MOUTH AND TIGHTLY-CLOSED EYES
1F607 SMILING FACE WITH HALO
1F608 SMILING FACE WITH HORNS
• commonly depicted as a (sinister) smiling version of 1F47F imp
1F609 WINKING FACE
1F60A SMILING FACE WITH SMILING EYES
1F60B FACE SAVOURING DELICIOUS FOOD
1F60C RELIEVED FACE
1F60D SMILING FACE WITH HEART-SHAPED EYES
1F60E SMILING FACE WITH SUNGLASSES
1F60F SMIRKING FACE
1F610 NEUTRAL FACE
1F611 EXPRESSIONLESS FACE
1F612 UNAMUSED FACE
1F613 FACE WITH COLD SWEAT
1F614 PENSIVE FACE
1F615 CONFUSED FACE
1F616 CONFOUNDED FACE
1F617 KISSING FACE
1F618 FACE THROWING A KISS
1F619 KISSING FACE WITH SMILING EYES
1F61A KISSING FACE WITH CLOSED EYES
1F61B FACE WITH STUCK-OUT TONGUE
1F61C FACE WITH STUCK-OUT TONGUE AND WINKING EYE
1F61D FACE WITH STUCK-OUT TONGUE AND TIGHTLY-CLOSED EYES
1F61E DISAPPOINTED FACE
1F61F WORRIED FACE
1F620 ANGRY FACE
1F621 POUTING FACE
1F622 CRYING FACE
1F623 PERSEVERING FACE
1F624 FACE WITH LOOK OF TRIUMPH
1F625 DISAPPOINTED BUT RELIEVED FACE
1F626 FROWNING FACE WITH OPEN MOUTH
1F627 ANGUISHED FACE
1F628 FEARFUL FACE
1F629 WEARY FACE
1F62A SLEEPY FACE
1F62B TIRED FACE
1F62C GRIMACING FACE
• should not be depicted with zipper mouth
→ 1F910 zipper-mouth face
1F62D LOUDLY CRYING FACE
1F62E FACE WITH OPEN MOUTH
1F62F HUSHED FACE
1F630 FACE WITH OPEN MOUTH AND COLD SWEAT
1F631 FACE SCREAMING IN FEAR
1F632 ASTONISHED FACE
1F633 FLUSHED FACE
• embarrassed
1F634 SLEEPING FACE
1F635 DIZZY FACE
1F636 FACE WITHOUT MOUTH
→ 2687 white circle with two dots

Unicode

Exempel på ”script” i Unicode som används för IDN:

Kyrilliska (bl.a. ryska) bokstäver:	ц ф ы
Grekiska bokstäver:	η κ ξ
Romerska (latinska) bokstäver:	a ə ǧ
Arabiska bokstäver (höger-till-vänster):	ڒ ټ ڤ
Devanagari (bl.a. hindi):	स छ य़
Kinesiska tecken (bl.a. kinesiska):	伎傈孋

Unicode

Varje tecken i Unicode har ett hexadecimalt värde som oftast skrivs som ”U+” och 4-6 hexadecimala siffor. T.ex.

ф	η	a	-	स	㑵
U+0444	U+03B7	U+0061	U+002D	U+0938	U+3455

"a" och "-" har värdena hex 61 resp 2D i ASCII.

LATIN SMALL LETTER A

Unicode är mer än bara en lista över tecken som ingår i teckenuppsättningen. Det är en databas där det för varje kod, t.ex. U+0061 (”a”), finns en massa information, t.ex.:

Unicode name	LATIN SMALL LETTER A
Block	ASCII
General Category	Lowercase Letter
Script	Latin
Bidirectional Category	Left To Right
Uppercase Mapping	0041 A

Unikt namn på tecknet

Vänster-till-höger-skrift

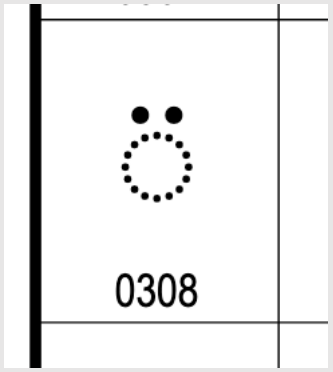
Den versala motsvarigheten

Kombinationstecken

Vissa tecken är kombinationstecken som hamnar över, under eller runt föregående tecken.

COMBINING DIAERESIS

U+0308



Unicode name	COMBINING DIAERESIS
Block	Diacriticals
General Category	Nonspacing Mark
Script	Inherited
Bidirectional Category	Nonspacing Mark

Kombinerade tecken

U+0308 är COMBINING DIAERESIS, ”prickar över”.

U+006F är ”LATIN SMALL LETTER O”, d.v.s. ”o”.

Kombinerat: ”U+006F U+0308”: ö

- I detta fall så finns det ett färdigt tecken med samma utseende, men med annan kod, och i domännamn måste man använda den:

U+00F6 LATIN SMALL LETTER O WITH DIAERESIS: ö

Kombinerade tecken

U+0308 är COMBINING DIAERESIS, ”prickar över”.

U+006E är ”LATIN SMALL LETTER N”, d.v.s. ”n”.

Kombinerat: ”U+006E U+0308”: n̈

- I detta fall finns inget färdigt tecken, utan man får använda kombinationen av två kodpunkter, eller tecken.

▶UTF-8 – teckenkodning av Unicode

[Innehåll]

Unicode och UTF-8

Unicode är en **teckenuppsättning**.

UTF-8 är en **teckenkodning** av Unicode och UTF-8 kan rymma upp till 1.112.064 tecken.

UTF-8 använder 1-4 byte per tecken beroende på Unicode-kod.

https://en.wikipedia.org/wiki/UTF-8

UTF-16 och UTF-32

Det finns även UTF-16 där minsta enheten är 16 bitar. För varje Unicode tecken används då 2 eller 4 byte.

UTF-32 använder alltid 4 byte, 32 bitar.

För internetprotokoll så är UTF-8 huvudalternativet. En anledning är att det fungerar väl med ASCII.

ASCII och UTF-8

UTF-8 är definierad så att ASCII blir en delmängd av UTF-8. Tecken U+0000 – U+007F kodas som 1 byte, vilket betyder att en fil kodad i ASCII också är kodad i UTF-8.

Detta gör UTF-8 mycket attraktivt för diverse protokoll.

▶Utökade tecken i IDN-domännamn

[Innehåll]

Tecken utöver ASCII

När mängden av tillåtna tecken i domännamn skulle utökas så ställdes följande krav:

- Alla tecken ur Unicode som används för att skriva språk ska inkluderas.
- Utökningen ska vara kompatibel med nuvarande DNS.

IDN

IDN = Internationalized Domain Name

Standard för att utöka domännamn med relevanta tecken ur Unicode.
Två versioner:

- IDNA2003 (ska inte användas)
- IDNA2008 (aktuell standard)

https://tools.ietf.org/html/rfc5890 m.fl.

Alla tecken ur Unicode

Inte bokstavligen, många tecken är uteslutna.

Följande typer av tecken kan användas i domännamn:

- Bokstäver och liknande tecken
- Siffror
- Vissa skiljetecken

Uteslutet är bl.a.:

- Nästan alla skiljetecken
- Emoji
- Matematiska tecken och andra symboler

Ordtecken

Domännamn är slags ord. De tecken som ska tillåtas är tecken som används för att skapa ord.

Domännamn behöver inte vara riktiga ord eller möjliga ord i något språk. Det kan vara en förkortning eller något nonsens.

Det är tecknen som ska vara ”ordtecken”.

Siffror är också inkluderade.

Inte versaler

Vissa skrivsystem, t.ex. det baserade på det latinska alfabetet, skiljer mellan versaler (”stora bokstäver”) och gemener (”små bokstäver”), t.ex.:

Gemen	Versal	Gemen	Versal
a	A	ö	Ö
æ	Æ	þ	Þ
đ	Đ	ď	Ď
ŋ	Ŋ	ə	Ǝ

Inte versaler

I traditionell DNS så är versaler A-Z samma sak som gemener a-z, men i IDN så får man inte använda versaler.

Exkluderingen av versaler i *IDN-namn* (***IDN labels***) gäller även A-Z. Inom IDN görs ingen ommappning från versaler till gemena.

Utanför IDN, i applikationer som hanterar IDN, så kan en mappning göras (skifta versal mot motsvarande gemen). Det är då en förprocess innan IDN.

Unicode-tecken får inte plats utan omkodning

DNS kan hantera oktetter, men det räcker inte för Unicode.

Kodning av tecken krävs för att kunna hantera så många olika tecken när domännamnet skickas i ett DNS-paket.

Mer om omkodningen i kommande bilder nedan.

▶Bakåtkompatibilitet

[Innehåll]

Kompatibel med DNS och domännamn

DNS och domännamn används i många applikationer. DNS klarar bara oktetter, och många system förväntar sig ”hostname”.

För att få maximal kompatibilitet så ska ***den omkodade domänen*** – eller snarare ***den omkodade ”label”*** – bara bara bestå av tecken som fungerar väl i DNS och som hostname.

Mera detaljer i kommande biler.

IDN hanterar ”label” inte domännamn

När vi ska avgöra om ett namn är ett IDN-namn eller ett traditionell namn så måste vi titta på varje ”label”. Reglerna för IDN gäller per ”label”.

När vi säger att ”räksmörgås.se” är ett IDN-namn så menar vi egentligen att någon ”label” är en IDN-label, den första i det här fallet. Den andra ”label” är en vanlig ASCII-label.

▶Två IDN-fomer: U-label och A-label

[Innehåll]

räksmörgås.se

Samma IDN-domän i två former

- Med U-label: räksmörgås.se
- Med A-label: xn--rksmrgs-5wao1o.se

- U-label = Unicode label
- A-label = ASCII compatible label

räksmörgås.se

Mera korrekt eftersom IDN handlar om ”label” inte domännamn:

U-label:	**ASCII-label:**
räksmörgås	se

A-label:	**ASCII-label:**
xn--rksmrgs-5wao1o	se

U-label

U-label är domännamnet (eller egenligen ”label”) med de avsedda tecknen som det är tänkt att användas när vanliga användare möter dem:

räksmörgås
香港
الاردن
ഭാരതം
コム

A-label

A-label är domännamnet (eller egentligen ”label”) omkodat för att användas i DNS-paket och zonfiler:

xn--rksmrgs-5wao1o
xn--j6w193g
xn--mgbayh7gpa
xn--rvc1e0am3e
xn--tckwe

A-label blir också ett reservalternativ att ta till ifall något system inte klarar en U-label, t.ex. inte kan visa den korrekt.

IDN-label

U-label	A-label
räksmörgås	xn--rksmrgs-5wao1o
香港	xn--j6w193g
الاردن	xn--mgbayh7gpa
ഭാരതം	xn--rvc1e0am3e
コム	xn--tckwe

A-label

A-label består av:

- Prefix ”xn--”
- omkodad U-label (punycode)

Ibland kallas hela A-label för ”punycode”, men det är oegentligt, för det är bara strängen efter "xn--" som är ”punycode”. A-label är mer än bara "punycode".

U-label till A-label

U-label	コム	bröd
Prefix	xn--	xn--
punycode-kodad U-label	tckwe	brd-tna
A-label	xn--tckwe	xn--brd-tna

Punycode

Punycode är en algoritm för att konvertera en Unicode-sträng till en ASCII-sträng, eller åt andra hållet. Konverteringen är utan förlust (förändring) om ASCII-strängen är taggad för att vara "punycode".

I den omkodade strängen används bara

- a-z (och *inte* A-Z)
- 0-9
- "-"

Fällor med A-label

Notera att det måste vara ”xn--”, inte ”xn—” eller något annat. ”xn--” förvanskas ofta i t.ex. mail eller worddokument.

I zonfilen så kommer det att bli fel om det inte är ”xn--”. Samma sak gäller verktyg som kan konvertera. Inklusive ”dig” och DNS-uppslagningar.

Punycode-kodning

abcå	defå	defä	defåä	åä
abc-wla	def-wla	def-sla	def-slae	4cab

ääö	å-äö	å-ä-ö	åaä-ö	
4cab6c	--0fab4e	---viac2g	a--viac2g	

Punycode-kodningen hanterar ASCII-tecken och icke-ASCII-tecken olika.

1. Om det finns ASCII-tecken:
    1. Behåll dessa i ordning och placera dess först. Även ”-” är ASCII.
    2. Lägg till ”-” efter ASCII-tecknena.
2. Skapa kodning av icke-ASCII och lägg på sist. Bara kodningen om inga ASCII-tecken.
3. Före punycode-koden ska vi sedan sätta prefixet ”xn--” för att skapa A-label.

U-label

Ett U-label får innehålla följande ASCII-tecken (förutom andra icke-ASCII-tecken):

a-z
0-9
”-”

Inga andra ASCII-tecken, inte ”_”, inte ens A-Z. Även restriktioner på bindestreck ”-”. Punkt är fortsatt avgränsaren ***mellan*** ”labels”.

Restriktionen gäller per ”label”, inte per domännamn.

U-label

_abc.räksmörgås.se – OK

_ääö.räksmörgås.se – Ej OK

Restriktion per ”label” inte per domännamn

Konvertering per label

Det är inte domännamnet som konverteras utan ”label” för ”label”.

bröd.smörgås.se → xn--brd-tna.xn--smrgs-pra0j.se

▶Verktyg för IDN-konvertering

[Innehåll]

”dig” och IDN

Nyare versioner av ”dig” kan hantera domännamn i U-labelformat. Om man inte ger någon parameter till den ”dig” som finns i labbmiljön så kommer ”dig”

- Hantera den givna domänen enligt IDN-standarden
- Presentera IDN-domäner i ”response” med U-label

”dig”-fråga med U-label

IDN-namn i U-labelformat till ”dig”

```
$ dig räksmörgås.se AAAA +noedns

; <<>> DiG 9.16.24-Ubuntu <<>> räksmörgås.se AAAA +noedns
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 2017
;; flags: qr rd ra ad; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 0

;; QUESTION SECTION:
;räksmörgås.se. IN AAAA

;; ANSWER SECTION:
räksmörgås.se. 60 IN AAAA 2a02:750:7::817
```

”dig” visar domännamnet med U-label, men i paketet från servern så var det ändå A-label. Det är ”dig” som har konverterat från A-label till U-label.

Fråga med A-label – ”dig” konverterar ändå

IDN-namn i A-labelformat till ”dig”

```
$ dig xn--rksmrgs-5wao1o.se AAAA +noedns

; <<>> DiG 9.16.24-Ubuntu <<>> xn--rksmrgs-5wao1o.se AAAA +noedns
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 2017
;; flags: qr rd ra ad; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 0

;; QUESTION SECTION:
;räksmörgås.se. IN AAAA

;; ANSWER SECTION:
räksmörgås.se. 60 IN AAAA 2a02:750:7::817
```

”dig” visar domännamnet med U-label, men i paketet från servern så var det ändå A-label. Det är ”dig” som har konverterat från A-label till U-label.

”dig” – konvertering i svaret avstängt

```
$ dig räksmörgås.se AAAA +noedns +noidnout

; <<>> DiG 9.16.24-Ubuntu <<>> räksmörgås.se AAAA +noedns +noidnout
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NOERROR, id: 24914
;; flags: qr rd ra ad; QUERY: 1, ANSWER: 1, AUTHORITY: 0, ADDITIONAL: 0

;; QUESTION SECTION:
;xn--rksmrgs-5wao1o.se. IN AAAA

;; ANSWER SECTION:
xn--rksmrgs-5wao1o.se. 60 IN AAAA 2a02:750:7::817
```

IDN-namn i U-labelformat till ”dig”

Konvertera inte utdata.

A-label, precis som det kommer från servern.

”dig” – konvertering av indata avstängt

```
$ dig räksmörgås.se AAAA +noedns +noidnin

; <<>> DiG 9.16.24-Ubuntu <<>> räksmörgås.se AAAA +noedns +noidnin
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NXDOMAIN, id: 34472
;; flags: qr rd ra ad; QUERY: 1, ANSWER: 0, AUTHORITY: 1, ADDITIONAL: 0

;; QUESTION SECTION:
;r\195\164ksm\195\182rg\195\165s.se. IN AAAA
```

IDN-namn i U-labelformat till ”dig”

Konvertera inte indata.

Oktetter från UTF-8 i indata.

”dig” och U-label

Med äldre ”dig” eller ”dig” som har kompilerats utan stöd för IDN så går det inte att fråga efter U-label:

```
$ dig räksmörgås.se +noedns
; <<>> DiG 9.10.6 <<>> räksmörgås.se +noedns
;; global options: +cmd
;; Got answer:
;; ->>HEADER<<- opcode: QUERY, status: NXDOMAIN, id: 46080
;; flags: qr rd ra ad; QUERY: 1, ANSWER: 0, AUTHORITY: 1, ADDITIONAL: 0

;; QUESTION SECTION:
;r\195\164ksm\195\182rg\195\165s.se. IN A
```

Inget stöd för U-label i denna ”dig”.

```
$ dig räksmörgås.se +idnin
Invalid option: +idnin
Usage:  dig [@global-server] [domain] [q-type] [q-class] {q-opt}
            {global-d-opt} host [@local-server] {local-d-opt}
            [ host [@local-server] {local-d-opt} [...]]

Use "dig -h" (or "dig -h | more") for complete list of options
```

”dig” och IDN-konvertering

Olika versioner av ”dig” har olika defaulter. Äldre version har inget IDN-stöd. Stäng av IDN-konvertering vid felsökning och för att få ”rå” data.

- +noidnin – stäng av konvertering av indata
- +idnin – slå på konvertering av indata
- +noidnout – stäng av konvertering av utdata
- +idnout – slå på konvertering av utdata

”dig” och IDN-konvertering

Stäng av konvertering vid uppslag av ogiltigt IDN-namn.

```
dig xx--smrgs-pra0j.dufberg.se txt +idnin +idnout

dig xx--smrgs-pra0j.dufberg.se txt +noidnin +noidnout

dig xx--smrgs-pra0j.dufberg.se txt +idnin +noidnout

dig xx--smrgs-pra0j.dufberg.se txt +noidnin +idnout
```

En label får inte börja på ”xx--”.

”dig”

”dig” kan används för att konvertera mellan U-label- och A-label-format.

- U-label- till A-label-format: ”dig +idnin +noidnout”
- A-label- till U-label-format: ”dig +idnin +idnout”

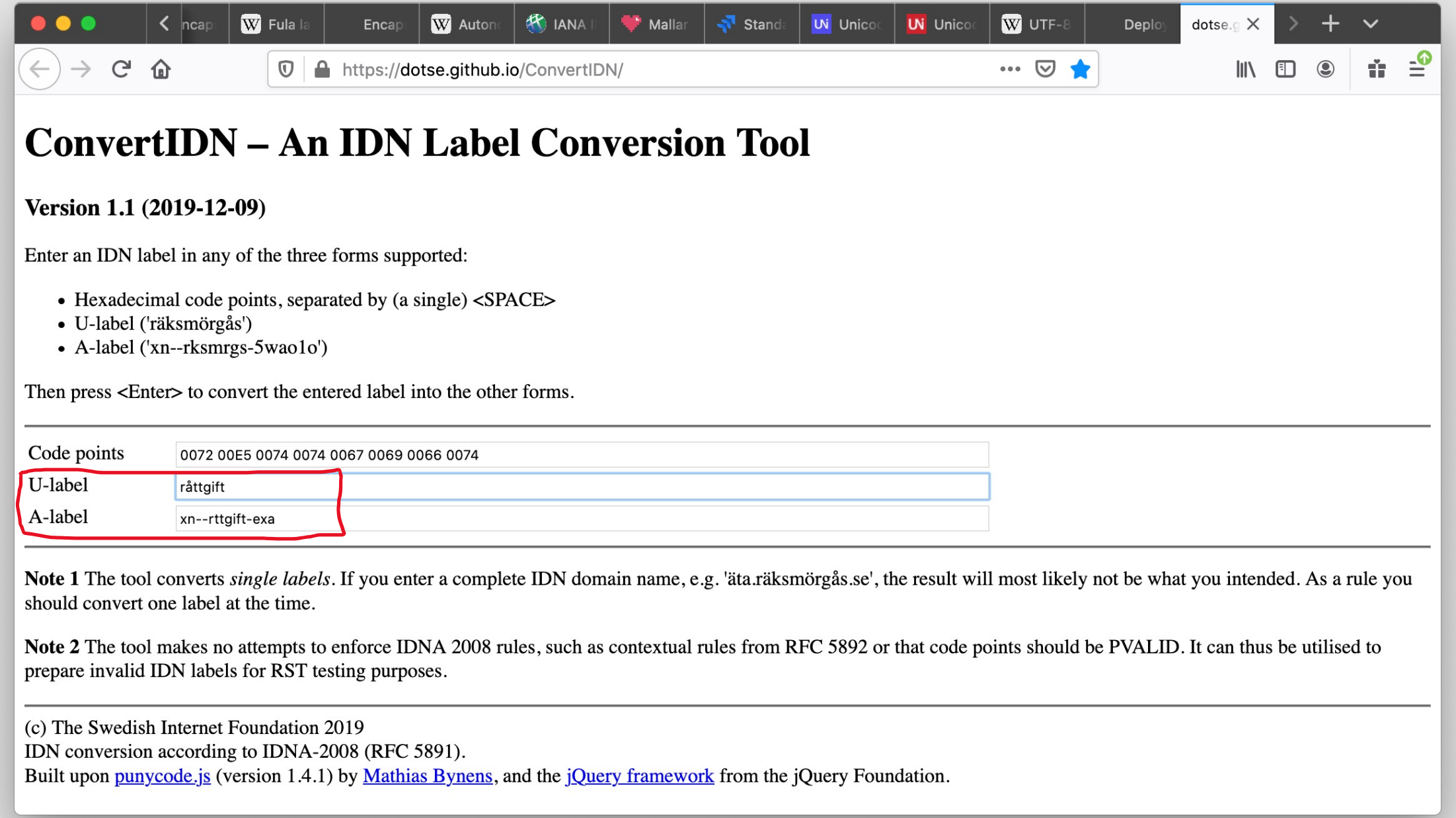
”dig” är hjälpsamt och byter ut ev. versaler mot motsvarande gemena (förprocessning innan IDN): RÅTTGIFT.se → råttgift.se

Prova i labbmiljön.

ConvertIDN

https://dotse.github.io/ConvertIDN/

- Kan bara konvertera ”label”, t.ex. räksmörgås, och inte hela domännamn, t.ex. räksmörgås.se
- Accepterar även ogiltiga IDN-labels (ska vara så i verktyget).
- Versaler och gemena hanteras olika (ska vara så i verktyget).

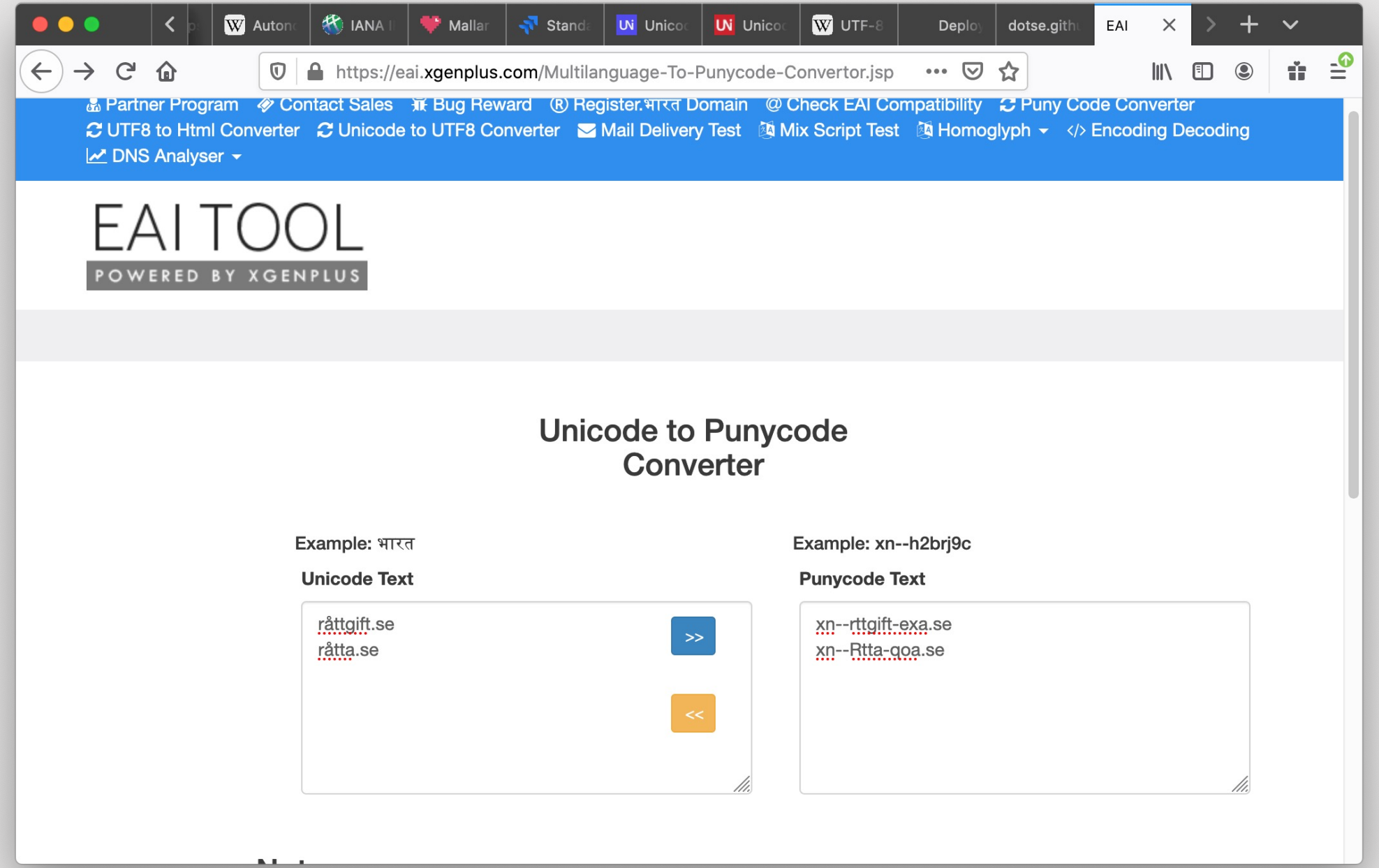

https://dotse.github.io/ConvertIDN/
ConvertIDN – An IDN Label Conversion Tool
Version 1.1 (2019-12-09)
Enter an IDN label in any of the three forms supported:
Hexadecimal code points, separated by (a single) <SPACE>
U-label ('räksmörgås')
A-label ('xn--rksmrgs-5wao1o')
Then press <Enter> to convert the entered label into the other forms.
Code points
0072 00E5 0074 0074 0067 0069 0066 0074
U-label
råttgift
A-label
xn--rttgift-exa
Note 1 The tool converts single labels. If you enter a complete IDN domain name, e.g. 'äta.räksmörgås.se', the result will most likely not be what you intended. As a rule you should convert one label at the time.
Note 2 The tool makes no attempts to enforce IDNA 2008 rules, such as contextual rules from RFC 5892 or that code points should be PVALID. It can thus be utilised to prepare invalid IDN labels for RST testing purposes.
(c) The Swedish Internet Foundation 2019
IDN conversion according to IDNA-2008 (RFC 5891).
Built upon punycode.js (version 1.4.1) by Mathias Bynens, and the jQuery framework from the jQuery Foundation.

Xgenplus Unicode to Punycode Converter

https://eai.xgenplus.com/Multilanguage-To-Punycode-Convertor.jsp

Konverterar domännamn med U-label till domännamn med A-label (oegentligt kallat punycode här).

- Kan hantera hela domännamn.
- Gör om versaler till gemena innan konvertering. Hjälpsamt.


EAI
https://eai.xgenplus.com/Multilanguage-To-Punycode-Convertor.jsp
Partner Program
Contact Sales
Bug Reward
Register.भारत Domain
Check EAI Compatibility
Puny Code Converter
UTF8 to Html Converter
Unicode to UTF8 Converter
Mail Delivery Test
Mix Script Test
Homoglyph
Encoding Decoding
DNS Analyser
EAI TOOL
POWERED BY XGENPLUS
Unicode to Punycode Converter
Example: भारत
Unicode Text
råttgift.se
råtta.se
>>
<<
Example: xn--h2brj9c
Punycode Text
xn--rttgift-exa.se
xn--Rtta-qoa.se

▶Restriktioner i U-label resp A-label

[Innehåll]

U-label

En U-label är en sträng av Unicode-tecken som

1. innehåller minst ett icke-ASCII-tecken,
2. endast innehåller tecken som är tillåtna för IDN,
3. uppfyller andra restriktioner för IDN.

En U-label har ingen direkt maximal längd, utan det är dess A-label som har en maximal längd.

A-label

En ASCII-sträng är en A-label om

1. den börjar på ”xn--”,
2. dess punycode-kod kan konverteras till en giltig U-label,
3. den är maximalt 63 tecken lång.

63 tecken är maxlängden på alla ”label” i DNS, oavsett om det är en A-label eller vanlig ASCII-label.

Ogiltig A-label

En sträng som ser ut som en A-label, men inte är det, är en ogiltig sträng.

Om punycode-delen har förvrängts så att det inte går att skapa en U-label så är det en ogiltig A-label.

Inga ASCII-strängar ska börja två tecken plus ”--” om de inte är A-label. T.ex. ”xx--” är otillåtet. Endast "xn--" är tillåtet, och endast om det är en giltig A-label.

U-label

Det finns en lång rad tecken som inte är tillåtna i IDN, t.ex.

1. versaler av bokstäver som ÅÄÖ,
2. många skiljetecken som ”!”,
3. symboler som emoji.

En applikation kan välja att som pre-process göra om versaler till gemener så att det blir en giltig U-label.

U-label

Det finns en rad restriktioner på hur tecken för kombineras eller placeras. T.ex.

1. vänster-till-höger-bokstäver (t.ex. romerska bokstäver) får inte kombineras med höger-till-vänster-bokstäver (t.ex. hebreiska bokstäver) i *samma label*,
2. en höger-till-vänster ”label” (t.ex. arabiska eller hebreiska bokstäver) får inte börja på en siffra.

IDN under .se

Under .se så kan en label innehålla vissa romerska bokstäver som "åäö" eller många hebreiska bokstäver, men inte i samma "label".

U-label

En U-label får inte

1. börja på ett ”-”,
2. sluta på ett ”-”,
3. ha formen ”xx--”, d.v.s. bindestreck i position tre och fyra

”xn--” är för A-label, inte för U-label.

Ej OK:

-abå	abå-	ab--abcåä	xn--abcåä

▶Använda U-label eller A-label?

[Innehåll]

Var används A-label?

U-label ryms inte i DNS-paketet. Där måste vi alltid använda A-label. I alla fält i DNS-paketet där det ska vara domännamn så måste det vara A-label.

I en normal zonfil använder vi A-label.

```
www.xn--rksmrgs-5wao1o.se.      A       79.99.1.121
                                AAAA    2a02:750:7::817
```

Var används A-label?

Även i konfigurationen en webbserver så används förmodligen A-label, men det finns inget tekniskt som hindrar att en webbserver har stöd för U-label i konfigurationen.

Var används U-label?

U-label är tänkt att vara den form som ska användas av användarna i sina klienter.

Webbläsare accepterar normalt U-label.

http://räksmörgås.se/

Stödet hos mailklienter är svagare, men det är på gång.

https://räksmörgås.se

räksmörgås.se

Detta är en test av IDN, *Internationalized Domain Names*. Det finns mer att läsa om IDN på Internetstiftelsens hemsida om IDN . Denna domän heter även xn--rksmrgs-5wao1o.se.

Testlänkar:

- www.räksmörgås.se
- www.xn--rksmrgs-5wao1o.se

Var används U-label?

Vanliga användare och domännamnsägare ska inte behöva hantera A-label. Det är DNS-teknikern som ska kunna göra konverteringen och se till att det blir rätt.

Det är OK att skicka rader som följande i en beställning, bara vi lägger in rätt format i zonfilen (använd verktyg, t.ex. modern dig, för att konvertera):

```
www.räksmörgås.se.     A      79.99.1.121
                       AAAA   2a02:750:7::817
```

Klienter och U-label

För att en klient eller applikation ska kunna hantera U-label så måste den har stöd för att kunna konvertera U-label till A-label (och tvärtom).

Alla DNS-uppslagningar måste göras med A-label, vilket betyder att alla klienter eller applikationer som stödjer U-label måste kunna konvertera till A-label.

Alla stora webbläsare kan hantera U-label, men kan också göra ”fallback” till A-label.

▶Utmaningar med IDN

[Innehåll]

Kan inte visa tecken i U-label

Om klienten inte kan visa tecken i U-label för att det inte finns någon font med dessa tecken så bör klienten göra ”fallback” till att visa A-label.

Användaren kan inte skriva av U-label

Om användaren får URL elektroniskt så kan den kopiera den till webbläsaren, men om det är en bild eller i tryckt form och användaren inte vet hur den ska skriva den? Tecknen finns kanske inte på tangentbordet.

Om A-label också finns så kan användaren göra ”fallback” till A-label och använda den istället.

A-label visas för användaren

Om stödet för IDN är begränsat i klienten (t.ex. mailprogrammet) så kan A-label visas istället för U-label.

Användaren kan kanske inte koppla ihop A-label med U-label, utan ser det som olika saker, där A-label är obegriplig.

Klienten stödjer inte U-label

Om klienten (t.ex. mailprogram) inte stödjer U-label så kan användaren inte skicka mail till en mailadress som bygger på U-label.

Om användaren inte har kunskap att göra konverteringen så blir det stopp.

Klienten slår upp U-label i DNS

Om klienten slår upp U-label i DNS utan att först konvertera den till A-label så kommer uppslagningen att misslyckas.

Användaren kan då blockeras från tjänsten.

Olika tecken med samma form

U-label	Unicode	A-label
acpë.xa	0430 0441 0440 0451	xn--80a6ac9d.xa
acpë.xa	0061 0063 0070 00EB	xn--acp-lma.xa

Den första raden är kyrilliska bokstäver (t.ex. ryska), den andra är romerska bokstäver (t.ex. franska).

- Samma form, samma sak för ”läsaren”.
- Olika Unicode-koder.
- Olika A-label, olika DNS-uppslagningar.
- I tryck så är det omöjligt att skilja dem åt.

U-label är inte entydig

Om U-label kommer i tryckt form eller i en bild så kan det vara så att teckna inte är entydiga. Se förra bilden.

Olika tecken med liknande form

Även inom ASCII så finns det tecken som är lika, t.ex. ”1”, ”l” och ”I” (beroende på teckensnitt).

Med IDN så har antalet föväxlingsbara teckenpar mångdubblats. Förväxlingar finns både inom samma ”script” och mellan olika ”script”.

Förväxlingar i ont syfte

Förväxlingsdomäner kan registreras för att försöka lura användare till att gå till fel webbplats.

Restriktioner

I viss utsträckning så finns det restriktioner som minskar risken för förväxlingar.

De flesta toppdomänerna tillåter bara en begränsad mängd tecken, en bråkdel av vad IDN tillåter.

De flera toppdomänerna tillåter inte att olika ”script” blandas i domännamnet man registrerar.

Två nivåer

Med IDN så har vi två nivåer av domännamn. En nivå som används i DNS-paketen och en annan nivå som ska användas i andra applikationer mot användaren.

A-label är också IDN

Ofta beskrivs IDN som att det bara är U-label.

Det är fel.

A-label är samma IDN-namn, bara omkodat till ASCII-tecken.

Två sidor av samma namn

A-label och U-label är samma IDN-namn (IDN-label), bara olika format.

Om man inte kan konvertera A-label till U-label så är det inget A-label.

Och U-label måste uppfylla restriktioner om tecken och kombinationer.

▶Om presentationen

[Innehåll]

Internets domännamnssystem

Denna presentation är framtagen 2019–2024 av Mats Dufberg (mats.dufberg@internetstiftelsen.se) på Internetstiftelsen (https://internetstiftelsen.se/). Den är en del av undervisningsmaterialet för kursen ”Internets domännamnssystem” vid Kungliga tekniska högskolan, KTH (kurskod HI1037) resp. Karlstads universitet, KAU (kurskod DVGC28).

Licens

Dokumenthistorik

- Rev A: Ursprunglig version VT 2024
- Rev B: Mindre korr i bild 2.

Slut.