Unicode in subject DN attributes

CA Browser Forum – Validation SC
February 10, 2022

PKI Consortium

Introduction

- With a common understanding and rules, we can create a lint to **detect and prevent** validation or quality issues
- The report shows **encoding errors** in most of the subject DN attributes
  - The OU attribute looks to have the most errors
- Values of the **address** must come from an authentic source
  - A linter such as the region zlint lint of the PKI Consortium matches against authentic values in Europe and ISO 3166-2 and highlights common issues
    - currently isn't good enough to enforce
    - not detailed enough to cover the locality, street or other address fields

PKI Consortium

Unicode usage

in the subject DN

```
#### INPUT: "tést\u00a01\u00ad\u0084\u200e\ufeff\u200b \u2004\u3000\u2028\ufff9\ufffa\ufffb "
U+0074 't'        116 LATIN SMALL LETTER T
U+0065 'e'        101 LATIN SMALL LETTER E
U+0301 ''         769 COMBINING ACUTE ACCENT
U+0073 's'        115 LATIN SMALL LETTER S
U+0074 't'        116 LATIN SMALL LETTER T
U+00A0        160 NO-BREAK SPACE
U+0031 '1'         49 DIGIT ONE
U+00AD        173 SOFT HYPHEN
U+0084        132 <control>
U+200E       8206 LEFT-TO-RIGHT MARK
U+FEFF      65279 ZERO WIDTH NO-BREAK SPACE
U+200B       8203 ZERO WIDTH SPACE
U+0020 ' '     32 SPACE
U+2004       8196 THREE-PER-EM SPACE
U+3000      12288 IDEOGRAPHIC SPACE
U+2028       8232 LINE SEPARATOR
U+FFF9      65529 INTERLINEAR ANNOTATION ANCHOR
U+FFFA      65530 INTERLINEAR ANNOTATION SEPARATOR
U+FFFB      65531 INTERLINEAR ANNOTATION TERMINATOR
U+FFFC ' '  65532 OBJECT REPLACEMENT CHARACTER
U+FFFD '' 65533 REPLACEMENT CHARACTER
```

```
#### Cleaned: "tést 1" (input length 21 <> result length 6)
U+0074 't'        116 LATIN SMALL LETTER T
U+00E9 'é'        233 LATIN SMALL LETTER E WITH ACUTE
U+0073 's'        115 LATIN SMALL LETTER S
U+0074 't'        116 LATIN SMALL LETTER T
U+0020 ' '         32 SPACE
U+0031 '1'         49 DIGIT ONE
```

- The 21 runes on the left, could simply be encoded as 6 runes at the right

PKI Consortium

Errors

February 8

Error details	Count
invalid character: U+0084 (132) - <control>	6
invalid character: U+00A0 (160) - NO-BREAK SPACE	178
invalid character: U+00A1 '¡' (161) - INVERTED EXCLAMATION MARK	64
invalid character: U+00A2 '¢' (162) - CENT SIGN	2
invalid character: U+00A4 '¤' (164) - CURRENCY SIGN	21
invalid character: U+00A7 '§' (167) - SECTION SIGN	16
invalid character: U+00AD (173) - SOFT HYPHEN	52
invalid character: U+00B6 '¶' (182) - PILCROW SIGN	13
invalid character: U+00B7 '·' (183) - MIDDLE DOT	114
invalid character: U+060C '،' (1548) - ARABIC COMMA	1
invalid character: U+2004 (8196) - THREE-PER-EM SPACE	13
invalid character: U+200B (8203) - ZERO WIDTH SPACE	119
invalid character: U+200E (8206) - LEFT-TO-RIGHT MARK	33
invalid character: U+2020 '†' (8224) - DAGGER	4
invalid character: U+2022 '•' (8226) - BULLET	7
invalid character: U+2116 '№' (8470) - NUMERO SIGN	6
invalid character: U+25CB 'ㅇ' (9675) - WHITE CIRCLE	1
invalid character: U+2F08 '人' (12040) - KANGXI RADICAL MAN	1
invalid character: U+2F2D '山' (12077) - KANGXI RADICAL MOUNTAIN	3
invalid character: U+2F4C '止' (12108) - KANGXI RADICAL STOP	1
invalid character: U+2F8F '行' (12175) - KANGXI RADICAL WALK ENCLOSURE	1
invalid character: U+3000 (12288) - IDEOGRAPHIC SPACE	1176
invalid character: U+3001 '、' (12289) - IDEOGRAPHIC COMMA	313
invalid character: U+30FB '・' (12539) - KATAKANA MIDDLE DOT	563
invalid character: U+321C '㈜' (12828) - PARENTHESIZED HANGUL CIEUC U	95
invalid character: U+FEFF (65279) - ZERO WIDTH NO-BREAK SPACE	21
invalid character: U+FF06 '＆' (65286) - FULLWIDTH AMPERSAND	38
invalid character: U+FF0C '，' (65292) - FULLWIDTH COMMA	23
invalid character: U+FF0E '．' (65294) - FULLWIDTH FULL STOP	7
invalid character: U+FF0F '／' (65295) - FULLWIDTH SOLIDUS	1
invalid character: U+FF65 '･' (65381) - HALFWIDTH KATAKANA MIDDLE DOT	9
invalid character: U+FFFD '  ' (65533) - REPLACEMENT CHARACTER	8
value contains multiple scripts map[Arabic:true Latin:true]	1
value contains multiple scripts map[CJK:true Greek:true Latin:true]	1
value contains multiple scripts map[CJK:true Latin:true]	1565
value contains multiple scripts map[Cyrillic:true Latin:true]	105
value contains multiple scripts map[Greek:true Latin:true]	59
value contains multiple scripts map[Hangul:true Latin:true]	14
value contains multiple scripts map[Latin:true Malayalam:true]	6
value contains multiple scripts map[Latin:true Thai:true]	11
value is not in unicode normalization form C (NFC)	152
Grand Total	**4824**

U+FFFD '  ' (65533)

REPLACEMENT CHARACTER

Used to replace an unknown, unrecognized, or unrepresentable character

Subject
CN=ovweb.viaginterkom.de, OU=Formware GmbH Nu  dorf am Inn, O=Telefonica Germany GmbH un
CN=ovweb.viaginterkom.de, OU=Formware GmbH Nu  dorf am Inn, O=Telefonica Germany GmbH un
CN=*.pp.uczelnia.awf.krakow.pl, OU=Akademia Wychowania Fizycznego im. Bronis  awa Czecha, O=A
CN=autodiscover.stwn.de, OU=St  dtische Werke Nrnberg GmbH, O=Städtische Werke Nürnberg Gr
CN=*.monnet-seve.fr, O=MONNET SEVE SA, OU=IT, L=Maillat, ST=Auvergne-Rh  ne-Alpes, C=FR
CN=*.monnet-seve.fr, O=MONNET SEVE SA, OU=IT, L=Maillat, ST=Auvergne-Rh  ne-Alpes, C=FR
CN=www.ventanillaunicaabogados.org, O=CONSEJO GENERAL DE LA ABOGACIA ESPA  OLA, OU=ITCG,
CN=*.mairie-lyon.fr, O=COMMUNE DE LYON, L=Lyon, ST=Auvergne-Rh  ne-Alpes, C=FR

PKI Consortium

U+00A4 '¤' (164)

CURRENCY SIGN

Used to denote an unspecified currency

Subject
CN=work.gut.tuwien.ac.at, OU=E080 GebÃ¤ude und Technik, O=Technische Universität Wien, L=Wien,
CN=video.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte
CN=openjournals.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, L=Vienna, ST=
CN=shop.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte K
CN=hyytiala.helsinki.fi, OU=HyytiÃ¤lÃ¤ forestry field station, O=University of Helsinki, L=Helsinki, C=FI
CN=interfaz-sn.gsi.com.mx, OU=CompaÃ±ia Mexicana de Traslado de Valores S.A de C.V., O=Compania M
CN=sec-gw.gut.tuwien.ac.at, OU=GebÃ¤ude und Technik, O=Technische Universität Wien, L=Wien, C=A
CN=kriswiki.irt.uu.se, OU=SÃ¤kerhetsavdelningen, O=Uppsala universitet, L=Uppsala, C=SE
CN=http-zid-voip.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angev
CN=stream.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandt
CN=supa.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte K
CN=viewer.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandels
CN=goobi.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandelsp
CN=bibtools.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandel
CN=openjournals.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welth
CN=www.gyn24.de, OU=Apotheke an der UniversitÃ¤t\, Bielefeld, O=Apotheke an der Universität Inha
CN=trafiidp.trafi.fi, O=Liikenne- ja viestintÃ¤virasto, L=Helsinki, ST=Uusimaa, C=FI
CN=161.41.251.3, serialNumber=1.2.246.10.2163978.10.0.13.2, O=Pirkanmaan Sairaanhoitopiirin kun
CN=143.51.4.122, serialNumber=1.2.246.10.2164623.13.1, O=Keski-Pohjanmaan sosiaali- ja terveyspa
CN=www.hausamkarswald.sachsen.de, OU=SÃ¤chsische Staatskanzlei, O=Freistaat Sachsen, street=Ar
CN=kvarkkigw.psshp.fi, serialNumber=1.2.246.10.1714953.10.0.13.5, O=Pohjois-Savon sairaanhoitopii

PKI Consortium

U+00A1 '¡' (161)

INVERTED EXCLAMATION MARK

Used to begin interrogative and exclamatory sentences or clauses in Spanish and some languages which have cultural ties with Spain, such as the Galician, Asturian and Waray languages

Subject
CN=work.gut.tuwien.ac.at, OU=E080 GebÃ¤ude und Technik, O=Technische Universität Wien, L=Wien, C=AT
CN=video.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte Kunst Wien
CN=openjournals.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, L=Vienna, ST=Vienna, C=
CN=shop.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte Kunst Wien,
CN=hyytiala.helsinki.fi, OU=HyytiÃ¤lÃ¤ forestry field station, O=University of Helsinki, L=Helsinki, C=FI
CN=interfaz-sn.gsi.com.mx, OU=CompaÃ±ia Mexicana de Traslado de Valores S.A de C.V., O=Compania Mexicana de
CN=sec-gw.gut.tuwien.ac.at, OU=GebÃ¤ude und Technik, O=Technische Universität Wien, L=Wien, C=AT
CN=kriswiki.irt.uu.se, OU=SÃ¤kerhetsavdelningen, O=Uppsala universitet, L=Uppsala, C=SE
CN=http-zid-voip.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte Kuns
CN=stream.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte Kunst Wie
CN=supa.uni-ak.ac.at, OU=UniversitÃ¤t fÃ¼r angewandte Kunst Wien, O=Universität für angewandte Kunst Wien,
CN=viewer.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandelsplatz 1, L=W
CN=goobi.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandelsplatz 1, L=Wi
CN=bibtools.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandelsplatz 1, L=
CN=openjournals.wu.ac.at, OU=UniversitÃ¤tsbibliothek, O=Wirtschaftsuniversität Wien, street=Welthandelsplatz
CN=www.gyn24.de, OU=Apotheke an der UniversitÃ¤t\, Bielefeld, O=Apotheke an der Universität Inhaber Dr. Tho
CN=trafiidp.trafi.fi, O=Liikenne- ja viestintÃ¤virasto, L=Helsinki, ST=Uusimaa, C=FI
CN=161.41.251.3, serialNumber=1.2.246.10.8265978.10.0.1[illegible]2, O=Pirkanmaan Sairaanhoitopiirin kuntayhtymÃ¤,
CN=143.51.4.122, serialNumber=1.2.246.10.2164623.13.1, O=Keski-Pohjanmaan sosiaali- ja terveyspalvelukuntay
CN=www.hausamkarswald.sachsen.de, OU=SÃ¤chsische Staatskanzlei, O=Freistaat Sachsen, street=Archivstrasse
CN=kvarkkigw.psshp.fi, serialNumber=1.2.246.10.1714953.10.0.13.5, O=Pohjois-Savon sairaanhoitopiirin kuntayht

PKI Consortium

Whitespace normalization

v · t · e

Name	Code point		Width box	May break?
CHARACTER TABULATION	U+0009	9		Yes
LINE FEED	U+000A	10	Is a line-break	
LINE TABULATION	U+000B	11	Is a line-break	
FORM FEED	U+000C	12	Is a line-break	

v · t · e

Name	Code point		Width box	May break?
MONGOLIAN VOWEL SEPARATOR	U+180E	6158		Yes
ZERO WIDTH SPACE	U+200B	8203		Yes
ZERO WIDTH NON-JOINER	U+200C	8204		Yes
ZERO WIDTH JOINER	U+200D	8205		Yes
WORD JOINER	U+2060	8288		No
ZERO WIDTH NON-BREAKING SPACE	U+FEFF	65279		No

LINE SEPARATOR	U+2028	8232	Is a line-break	
PARAGRAPH SEPARATOR	U+2029	8233	Is a line-break	
NARROW NO-BREAK SPACE	U+202F	8239		No
MEDIUM MATHEMATICAL SPACE	U+205F	8287		Yes
IDEOGRAPHIC SPACE	U+3000	12288		Yes

PKI Consortium

NFC normalization

UAX #15

Canonical-equivalent strings have the same binary representation.

In NFC normalization the runes:

```
U+0065 'e'   (LATIN SMALL LETTER E) and
U+0301 '́'   (COMBINING ACUTE ACCENT)
```

that make the letter é, are combined into the single rune:

```
U+00E9 'é'   (LATIN SMALL LETTER E WITH
ACUTE).
```

Visually they look the same but in a byte for byte comparison they are not.

Unicode advised to use NFC normalization when comparing values

PKI Consortium

Do we need these characters?

Block	Range	Example
Domino Tiles	U+1F030 - U+1F09F	
Braille Patterns	U+2800 - U+28FF	
Geometric Shapes	U+25A0 - U+25FF	
Emoticons	U+1F600 - U+1F64F	
Enclosed Alphanr.	U+1F100 - U+1F1FF	
Arrows	U+2190 - U+21FF	
Block Elements	U+2580 - U+259F	
Musical Symbols	U+1D100 - U+1D1FF	
Playing Cards	U+1F0A0 - U+1F0FF	
Transport & Map …	U+1F680 - U+1F6FF	
Chess Symbols	U+1FA00 - U+1FA6F	
Egypt. Hieroglyphs	U+13000 - U+1342F	

PKI Consortium

Naming police

What is acceptable?

- The objective of the subject information is to **inform and protect** the relying party
  - Relying parties **should not be confused** with encoding errors, symbols, and other markup
- We **do not** want to police what a valid name looks like
  - Authentic sources have different rules
  - Where to draw the line, do we accept:
    - Names in **small form**, **full or half width** characters?
    - Name that consists solely of **lines**, **shades** and **symbols**?
    - Character replacements?

Figurative name	Code	Representing
▒▒▒▒▒▒▒▒▒▒▒	U+2587	Redacted
☢ Company	U+2622	Radioactive Company
Ｃｏｍｐａｎｙ	U+FF00 - U+FFEF	Company
⏎ Home	U+23CE	Return Home
⍰ Inc.	U+1F6D7	Elevator Inc.
⠠⠭⠉⠕⠍⠏⠁⠝⠽ ⠊⠝⠉	U+2800 - U+28FF	Company
🅱🅾🆇	U+1F100 - U+1F1FF	Box

PKI Consortium

Thanks

PKI
Consortium