Internationalization and Internet Engineering

IETF68 Technical Plenary

Prague

March 22, 2007

Agenda

- A bit of stage setting (Leslie Daigle)
- Protocol engineering & Languages -- the many pieces of the puzzle (Ted Hardie)
- The IETF experience (John Klensin)
- Open discussion with panel
  - Ted Hardie
  - John Klensin
  - 李曉東 (Xiaodong Lee)
  - Patrik Fältström,
  - Pete Resnick

Scope & Purpose

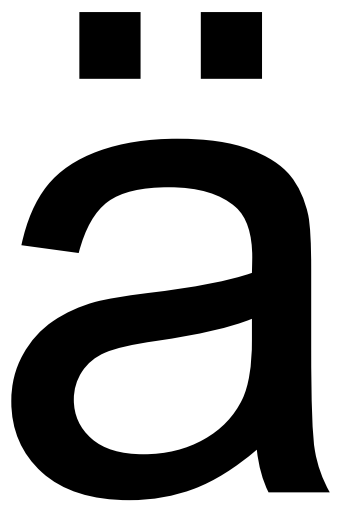
- Shine a light on the significant issues in internationalization and protocol design/Internet Engineering
- Provide broader picture than covered by the IAB IDN Next Steps document
  - RFC 4690
  - Reviewed at IETF66 plenary

Does your consciousness need raising?

- “No, I don’t do Applications”
- I do bytes on the wire
- Oh, I thought this was ROUTING and addressing

- How big an issue could it be? ...

加了分音符号的拉丁文小写字母A



Arial 250pt

编码

ASCII	*不存在*	
ISO-646-SE	0x7B	1字节，16进制
Unicode	U+00E4	码点
UTF-8	0xC3A4	2字节，16进制
UTF-16	0x00E4	2字节，16进制
UTF-32	0x000000E4	4字节，16进制
XHTML	ä	文本表示
XHTML	ä	文本表示

编码的混淆

同样的线上比特流，不同的编码解释

ISO-646-SE	0x7B	ä
ISO-8859-1	0x7B	{

同样的线上比特流，不同的编码解释

UTF-8	0xC3A4	ä
ISO-8859-1	0xC3 0xA4	Ã¤

一个punycode，不同的编码字符

Punycode	Ascii: xn--4ca	xn--4ca
Unicode	0x00E4	ä
Unicode	U+0061 U+0308	ä (a + ¨)
Unicode	U+04D3	ä

字形的混淆

	U+007B	U+04DE
Arial	ä	ä
Times	ä	ä
Zapf Dingbats	➤	ä

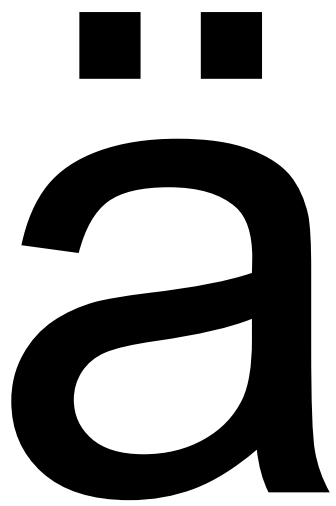
??

ASCII	???	
ISO-646-SE	0x7B	???, ???
Unicode	U+00E4	??
UTF-8	0xC3A4	???, ????
UTF-16	0x00E4	???, ????
UTF-32	0x000000E4	???, ????
XHTML	ä	????
XHTML	ä	????

Got it? ;-)

- Credit to Patrik for the slides, and Xiaodong for the translations
- English versions included in the online version of this deck
- Ted & John will give the whole picture
  - including more detail on the last few slides

LATIN SMALL LETTER A WITH DIAERESIS



Arial 250pt

Encoding

ASCII	*Does not exist*	
ISO-646-SE	0x7B	1 byte, hex rep
Unicode	U+00E4	Codepoint
UTF-8	0xC3A4	2 bytes, hex rep
UTF-16	0x00E4	2 bytes, hex rep
UTF-32	0x000000E4	4 bytes, hex rep
XHTML	ä	text representation
XHTML	ä	text representation

Confusion - Encoding

Same bits on the wire; different encoding interpretation

ISO-646-SE	0x7B	ä
ISO-8859-1	0x7B	{

Same bits on the wire; different encoding interpretation

UTF-8	0xC3A4	ä
ISO-8859-1	0xC3 0xA4	Ã¤

One punycode; different encoded characters

Punycode	Ascii: xn--4ca	xn--4ca
Unicode	0x00E4	ä
Unicode	U+0061 U+0308	ä (a + ¨)
Unicode	U+04D3	ä

Confusion - Font

	U+007B	U+04DE
Arial	ä	ä
Times	ä	ä
Zapf Dingbats	➤	ä