unicode

Niels Möller

5 februari 2002

Designmål, enligt Unicode Consortium

- 16-bitars tecken
- Enkelt och effektivt att parsa och hantera
- Tecken, inte glyfer
- Representera ”plain text”
- Logisk ordning
- Unifiering
- Dynamisk komposition
- Konvertering till och från alla vanliga teckenuppsättningar

Unidata.txt

Code	Name	Cat.	CC	Decomp
0000	<control>	Cc	0	
0020	Space	Zs	0	
0041	Latin capital letter A	Lu	0	
00BC	Fraction one quarter	No	0	0031 2044 0034
00C5	A with ring above	Lu	0	0041 030A
030A	Combining ring above	Mn	230	
0627	Arabic letter Alef	Lo	0	
0E52	Thai digit two	Nd	0	
212B	Angstrom sign	Lu	0	00C5
2282	Subset of	Sm	0	

Code	Name	Bidi	N.	M.	Lower
0000	<control>	BN		N	
0020	Space	WS		N	
0041	Latin capital letter A	L		N	0061
00BC	Fraction one quarter	ON	1/4	N	
00C5	A with ring above	L		N	00E5
030A	Combining ring above	NSM		N	
0627	Arabic letter Alef	AL		N	
0E52	Thai digit two	L	2	N	
212B	Angstrom sign	L		N	
2282	Subset of	ON		Y	

UTF-8

ISO-10646	Byte 1	Byte 2	Byte 3
xxxxxxx	0xxxxxxx		
yyy yyxxxxxx	110yyyyy	10xxxxxx	
zzzzyyyy yyxxxxxx	1110zzzz	10yyyyyy	10xxxxxx

Sorteringsregler – default

```
+ "& TH, \u00DE &TH, \u00FE " // thorn
+ "<u,U"
+ "<v,V"
+ "<w,W"
+ "<x,X"
+ "<y,Y"
+ "<z,Z"
+ "&AE,\u00C6"  // ae & AE ligature
+ "&AE,\u00E6"
+ "&OE,\u0152"  // oe & OE ligature
+ "&OE,\u0153"
```

Sorteringsregler – spanska

```
     "& N < n\u0303, N\u0303 "
// + "& L < ll , Ll, LL"
```

Sorteringsregler – svenska

```
"& Z < \u00e6 , \u00c6 " + // Z < ae ligature
"< a\u030a , A\u030a , aa , aA , Aa , AA" +
"< a\u0308 , A\u0308 < o\u0308 , O\u0308 " +
"; u\u030b , U\u030b " + // u-double-acute
"< o\u030b , O\u030b ; \u00f8 , \u00d8 " +
"& V ; w , W" +
"& Y, u\u0308 , U\u0308"
```

Att representera tecken

```
struct Character
{
  unsigned ulength; wchar_t *chars;

  /* NULL if this character is already
   * in canonical representation */
  struct Character *canonical;

  /* non-NULL if character can be up-cased,
   * lower-cased or title-cased */
  struct Character *upper, *lower, *title;

  /* The syntactic class */
  enum character_syntax syntax;

  /* Additional information needed if we
   * want to draw text on screen */
  struct Glyph *glyph;

  enum { LEFT, RIGHT } direction;
};
```