COGNITE

Toveis tekst i Unicode

Også kjent som UAX #9

Eira Monstad

Revolverconf 2023.2

COGNITE

Toveis tekst: Unicode Bidirectional Algorithm

eller: Hvordan tegne hebraisk og arabisk tekst på skjermen

COGNITE

Først: Hva er tegnkoding?

COGNITE

Tegnkoding er en oppskrift for å gjøre en serie med tall om til kruseduller

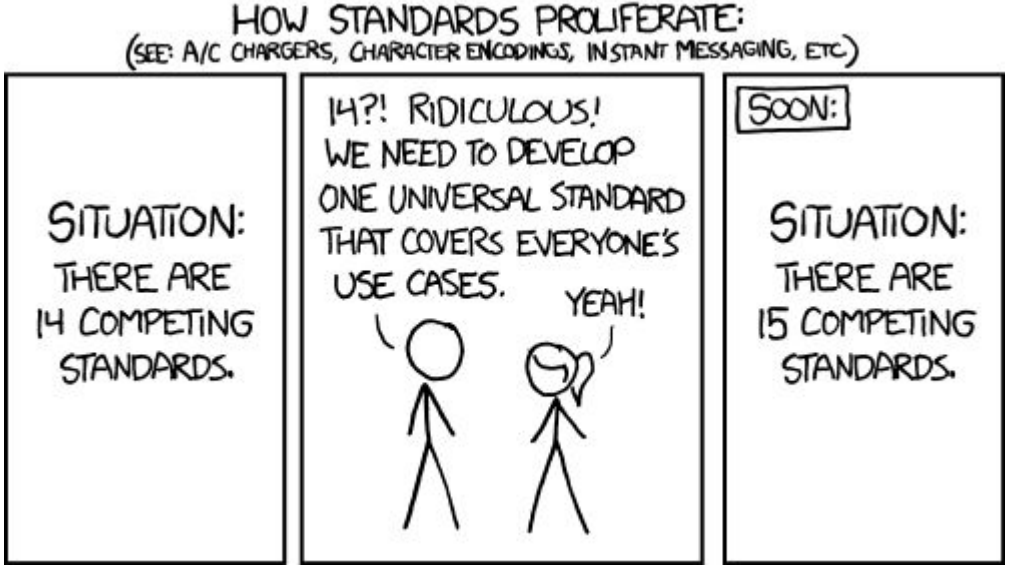
Hvor Vanskelig Kan Det Være?
En veldig, veldig forkortet historie om tegnkoding

1. ASCII.
2. Windows-1252 og ISO-8891-1
3. Shift-JIS, Big5, EUC-KR, …

Common character encodings [edit]

- ISO 646
  - ASCII
- EBCDIC
- ISO 8859:
  - ISO 8859-1 Western Europe
  - ISO 8859-2 Western and Central Europe
  - ISO 8859-3 Western Europe and South European (Turkish, Maltese plus Esperanto)
  - ISO 8859-4 Western Europe and Baltic countries (Lithuania, Estonia, Latvia and Lapp)
  - ISO 8859-5 Cyrillic alphabet
  - ISO 8859-6 Arabic
  - ISO 8859-7 Greek
  - ISO 8859-8 Hebrew
  - ISO 8859-9 Western Europe with amended Turkish character set
  - ISO 8859-10 Western Europe with rationalised character set for Nordic languages, including complete Icelandic set
  - ISO 8859-11 Thai
  - ISO 8859-13 Baltic languages plus Polish
  - ISO 8859-14 Celtic languages (Irish Gaelic, Scottish, Welsh)
  - ISO 8859-15 Added the Euro sign and other rationalisations to ISO 8859-1
  - ISO 8859-16 Central, Eastern and Southern European languages (Albanian, Bosnian, Croatian, Hungarian, Polish, Romanian, Serbian and Slovenian, but also French, German, Italian and Irish Gaelic)
- CP437, CP720, CP737, CP850, CP852, CP855, CP857, CP858, CP860, CP861, CP862, CP863, CP865, CP866, CP869, CP872
- MS-Windows character sets:
  - Windows-1250 for Central European languages that use Latin script, (Polish, Czech, Slovak, Hungarian, Slovene, Serbian, Croatian, Bosnian, Romanian and Albanian)
  - Windows-1251 for Cyrillic alphabets
  - Windows-1252 for Western languages
  - Windows-1253 for Greek
  - Windows-1254 for Turkish
  - Windows-1255 for Hebrew
  - Windows-1256 for Arabic
  - Windows-1257 for Baltic languages
  - Windows-1258 for Vietnamese
- Mac OS Roman
- KOI8-R, KOI8-U, KOI7
- MIK
- ISCII
- TSCII
- VISCII
- JIS X 0208 is a widely deployed standard for Japanese character encoding that has several encoding forms.
  - Shift JIS (Microsoft Code page 932 is a dialect of Shift_JIS)
  - EUC-JP
  - ISO-2022-JP
- JIS X 0213 is an extended version of JIS X 0208.
  - Shift_JIS-2004
  - EUC-JIS-2004
  - ISO-2022-JP-2004
- Chinese Guobiao
  - GB 2312
  - GBK (Microsoft Code page 936)
  - GB 18030
- Taiwan Big5 (a more famous variant is Microsoft Code page 950)
  - Hong Kong HKSCS
- Korean
  - KS X 1001 is a Korean double-byte character encoding standard
  - EUC-KR
  - ISO-2022-KR
- Unicode (and subsets thereof, such as the 16-bit 'Basic Multilingual Plane')
  - UTF-8
  - UTF-16
  - UTF-32
- ANSEL or ISO/IEC 6937

Velkommen: Unicode*!



* spesifikt: UTF-8. La oss ikke snakke om UTF-16. Og jo mindre vi sier om UTF-32, jo bedre.

Contents

Unicode® Standard Annex #9

UNICODE BIDIRECTIONAL ALGORITHM

COGNITE

Hva mener vi med Toveis?

* eller BiDi

COGNITE

Hvorfor er toveis tekst vanskelig?

Enkelt: Å tegne baklengs

Det du ser:

العربية

54321

Det datamaskinen ser:

1 2 3 4 5

norsk

12345

Det datamaskinen ser:

1 2 3 4 5

Litt vanskeligere: Å tegne baklengs og forlengs samtidig

Det du ser:

123العربية

67854321

Det datamaskinen ser:

1 2 3 4 5 6 7 8

norsk123

12345678

1 2 3 4 5 6 7 8

Enda vanskeligere: Tegnsetting

Tittel: “العربية!”

Hvordan vet datamaskinen om utropstegnet hører til setningen eller ordet?

Ekstra gøy: Speilvending av tegn

a < b < c

ع > ر > ل

‹ og › er samme tegn: Unicode U+003C Less-than sign

COGNITE

Hvordan løser vi dette?

Tegntyper: Sterke, svake, nøytrale, og formattering

Category	Type	Description	General Scope
Strong	L	Left-to-Right	LRM, most alphabetic, syllabic, Han ideographs, non-European or non-Arabic digits, ...
	R	Right-to-Left	RLM, Hebrew alphabet, and related punctuation
	AL	Right-to-Left Arabic	ALM, Arabic, Thaana, and Syriac alphabets, most punctuation specific to those scripts, ...
Weak	EN	European Number	European digits, Eastern Arabic-Indic digits, ...
	ES	European Number Separator	PLUS SIGN, MINUS SIGN
	ET	European Number Terminator	DEGREE SIGN, currency symbols, ...
	AN	Arabic Number	Arabic-Indic digits, Arabic decimal and thousands separators, ...
	CS	Common Number Separator	COLON, COMMA, FULL STOP, NO-BREAK SPACE, ...
	NSM	Nonspacing Mark	Characters with the General_Category values: Mn (Nonspacing_Mark) and Me (Enclosing_Mark)
	BN	Boundary Neutral	Default ignorables, non-characters, and control characters, other than those explicitly given other types.
Neutral	B	Paragraph Separator	PARAGRAPH SEPARATOR, appropriate Newline Functions, higher-level protocol paragraph determination
	S	Segment Separator	*Tab*
	WS	Whitespace	SPACE, FIGURE SPACE, LINE SEPARATOR, FORM FEED, General Punctuation spaces, ...
	ON	Other Neutrals	All other characters, including OBJECT REPLACEMENT CHARACTER
Explicit Formatting	LRE	Left-to-Right Embedding	LRE
	LRO	Left-to-Right Override	LRO
	RLE	Right-to-Left Embedding	RLE
	RLO	Right-to-Left Override	RLO
	PDF	Pop Directional Format	PDF
	LRI	Left-to-Right Isolate	LRI
	RLI	Right-to-Left Isolate	RLI
	FSI	First Strong Isolate	FSI
	PDI	Pop Directional Isolate	PDI

Avsnitt, baseretning og retningsløp

Algoritmen opererer på avsnitt.

Et avsnitt har en baseretning:
Høyre-til-venstre (RTL)
Venstre-til høyre (LTR)

Baseretningen settes av det første sterke tegnet i avsnittet.

Baseretningen er et retningsløp. Mer om det senere.

Dette avsnittet begynner med א
!en sterk høyre

A Dette avsnittet begynner med
en sterk venstre!

Sterke og nøytrale tegn

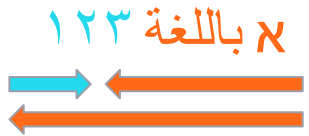
Sterke tegn følger alltid sin egen retning: RTL eller LTR.

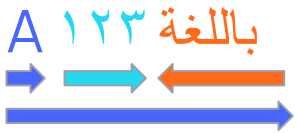
Nøytrale tegn følger alltid retningsløpet.

Svake tegn

Svake tegn har en *intern* retning RTL eller LTR, men følger retningsløpet.

Gøy: Arabiske tall skrives fra venstre til høyre!





Husker du denne?

Tittel: “العربية!”

Hvordan vet datamaskinen om utropstegnet hører til setningen eller ordet?

Og så har me juksa lite granne

Tittel: “العربية!”

Merker: Sterke, men usynlige
LRM U+200E Lat som om det er et sterkt venstre-til-høyre-tegn her
RLM U+200F Lat som om det er et sterkt høyre-til-venstre-tegn her

Flere juksekoder: Innbygging og overstyring

Innbygging: Lag et innbygget retningsløp

LRE U+202A: Start et retningsløp fra venstre mot høyre
RLE U+202B: Start et retningsløp fra høyre mot venstre

Overstyring: Lat som alle tegnene i dette løpet er sterke i en gitt retning

LRO U+202D: Start et løp der alle tegn vises venstre-til-høyre
RLO U+202E: Start et løp der alle tegn vises høyre-til-venstre

Slutt med det: Avslutt løpet og hopp tilbake til der du slapp

PDF U+202C: Avslutt nærmeste LRE, RLE, LRO eller RLO

Del 2

Neste gang: Vi bruker selve algoritmen på et eksempel

BD9. The *matching PDI* for a given isolate initiator is the one determined by the following algorithm:

- Initialize a counter to one.
- Scan the text following the isolate initiator to the end of the paragraph while incrementing the counter at every isolate initiator, and decrementing it at every PDI.
- Stop at the first PDI, if any, for which the counter is decremented to zero.
- If such a PDI was found, it is the matching PDI for the given isolate initiator. Otherwise, there is no matching PDI for it.

COGNITE

Spørsmål?

Kilder & nostalgi

- https://www.unicode.org/reports/tr9/
- https://en.wikipedia.org/wiki/Character_encoding

W3C home > Mailing lists > Public > public-css-testsuite@w3.org > July 2008

CSS2.1 i18n and bidi tests for review

This message: Message body | Respond | More options
Related messages: Next message | Next in thread | Replies

From: Eira Monstad <eiram@opera.com>
Date: Tue, 01 Jul 2008 10:20:49 +0200
To: public-css-testsuite@w3.org
Message-ID: <op.udltgz00nwv5is@voodoobaby>

Cheers,

I've been working on some bidi related tests for the 2.1 testsuite, as well as converting a few of Richard Ishida's tests for language dependent styling to match the testsuite template. The tests are ready for review at http://people.opera.com/eiram/test/css21/review/

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.1//EN" "http://www.w3.org/TR/xhtml11
<html xmlns="http://www.w3.org/1999/xhtml">
  <head>
    <title>CSS Test: bidirection box model - borders on inline in bidi-override
    <link rel="author" title="Eira Monstad, Opera Software ASA" href="mailto:pu
    <link rel="reviewer" title="Gérard Talbot" href="http://www.gtalbot.org/Bro
    <link rel="help" href="http://www.w3.org/TR/css-writing-modes-3/#bidi-box-m
    <link rel="help" href="http://www.w3.org/TR/CSS21/box.html#bidi-box-model"/
    <link rel="match" href="bidi-box-model-001-ref.xht" />

    <meta name="assert" content="Border sides should be unaffected by direction
    <style type="text/css"><![CDATA[
      span {
      border: 5px solid gray;
      border-color: orange purple teal yellow;
      }

      .rtol {
      direction: rtl;
      unicode-bidi: bidi-override;
      }

      p {text-align: left;}
    ]]></style>
  </head>

  <body>
    <p>Test passes if the 2 lines are <strong>identical</strong>.</p>
    <p>
      First <span>Second</span>
    </p>
    <p class="rtol">
      <span>dnoceS</span> tsriF
    </p>
  </body>
</html>
```