Mengenal *Character Set*

Ida Bagus Adi Sudewa

deweu@yahoo.com

Bagi para pembaca yang sudah cukup lama berkecimpung di dunia komputer, pasti pernah bekerja dengan 'kode ASCII'. Dan bagi anda yang bekerja dengan mesin-mesin mainframe IBM, pasti pernah menjumpai 'kode EBCDIC' (dibaca: *eb-si-dik*). Di luar ASCII dan EBCDIC, besar kemungkinan anda paling tidak pernah mendengar istilah-istilah ajaib seperti berikut ini: ISO-8859-1, UCS-2, UTF-8, UTF-16, atau windows-1252. Kode-kode apakah itu? ASCII, EBCDIC, ISO-8859-*x*, UCS-2, UTF-*x*, dan windows-*x* merupakan sebagian dari kumpulan *character set* (set karakter) yang ada di dunia komputer. Makin bingung? Mari kita mulai!

Pada Awalnya: ASCII dan EBCDIC

ASCII (American Standard Code for Information Interchange) dan EBCDIC (Extended Binary Coded Decimal Interchange Code) merupakan cikal bakal dari set karakter lainnya. ASCII merupakan set karakter yang paling umum digunakan hingga sekarang. Set karakter ASCII terdiri dari 128 buah karakter[1] yang masing-masing memiliki lebar 7-bit atau tujuh angka 0 dan 1, dari 0000000 sampai dengan 1111111. Mengapa 7-bit? Karena komputer pada awalnya memiliki ukuran memori yang sangat terbatas, dan 128 karakter dianggap memadai untuk menampung semua huruf Latin dengan tanda bacanya, dan beberapa karakter kontrol.

ASCII terdiri dari huruf-huruf, angka-angka, dan tanda-tanda baca sebagai berikut:

```
  ! " # $ % & ' ( ) * + , - . /
0 1 2 3 4 5 6 7 8 9 : ; < = > ?
@ A B C D E F G H I J K L M N O
P Q R S T U V W X Y Z [ \ ] ^ _
` a b c d e f g h i j k l m n o
p q r s t u v w x y z { | } ~
```

ASCII juga terdiri dari karakter-karakter yang tidak kelihatan, seperti CR (*carriage return*), LF (*line feed*) dan ESC (*escape*).

[1] Ada 128 buah karakter dengan lebar 7-bit, $2^7 = 128$

Tabel 1 Contoh Karakter ASCII

Biner	***Heksadesimal***	***Nama***	***Bentuk (Glyph)***
000 0101	0x05	CARRIAGE RETURN	N/A
011 0000	0x30	DIGIT ZERO	0
100 0001	0x41	LATIN CAPITAL LETTER A	A
111 1101	0x7D	RIGHT CURLY BRACKET	}

ASCII telah dibakukan pada tahun oleh ANSI (American National Standards Institute) menjadi standar ANSI X3.4-1986.

EBCDIC merupakan set karakter yang merupakan ciptaan dari IBM. Salah satu penyebab IBM menggunakan set karakter di luar ASCII sebagai standar pada komputer ciptaan IBM adalah karena EBCDIC lebih mudah dikodekan pada *punch card* yang pada tahun 1960-an masih jamak digunakan. Penggunaan EBCDIC pada mainframe IBM masih terbawa hingga saat ini, walaupun *punch card* sudah tidak digunakan lagi.

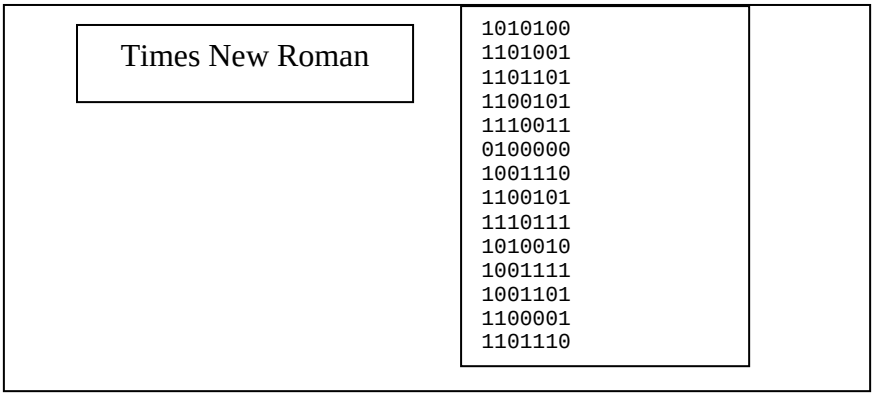
Seperti halnya ASCII, EBCDIC juga terdiri dari 128 karakter yang masing-masing berukuran 7-bit. Hampir semua karakter pada ASCII juga terdapat pada set karakter EBCDIC.

Kalau ASCII dan EBCDIC sudah mampu mengkodekan 128 karakter, lalu mengapa masih dianggap perlu untuk menciptakan set karakter baru? Lebar set karakter yang cuma 128 karakter tidak memungkinkan penulisan karakter-karakter di luar huruf Latin (*basic Latin)*, seperti misalnya huruf *ü* atau simbol-simbol matematika dalam huruf Yunani. Maka lahirlah keluarga set karakter ISO-8859-*x*.

Karakter dan *Font*

Sebelum membahas tentang ISO-8859-*x*, mungkin ada pembaca yang masih merasa bingung. Mengapa sih kita perlu set karakter? Lalu apa hubungannya dengan *font* pada Microsoft Windows?

Karakter merupakan representasi huruf pada memori komputer. *Font* adalah kumpulan *bentuk* huruf yang memetakan karakter menjadi komposisi *pixel* pada layar atau printer. ASCII, ISO-8859-1 atau Unicode adalah set karakter. “Times New Roman” adalah *font* yang mampu memetakan (*render*) semua karakter ISO-8859-1 pada layar, dengan bentuk-bentuk huruf “bergaya” Times New Roman.



Gambar 1 Kalimat "Times New Roman" dan Representasinya pada Memori Komputer dengan Set Karakter ASCII

ISO-8859-*x*

ISO (International Organization for Standardization) menetapkan keluarga set karakter ISO-8859-*x* sebagai penerus dari ASCII. Ada empat belas set karakter pada keluarga ISO-8859-*x*. Karakteristik dari keluarga ini adalah:

- Masing-masing terdiri dari 256 karakter dengan lebar masing-masing karakter adalah 8-bit
- Kompatibel ke belakang (*backward* compatible) dengan ASCII, dengan 128 karakter pertama sama persis dengan 128 karakter ASCII
- Sisa jumlah karakter sebanyak 128 karakter dipergunakan untuk mengkodekan karakter-karakter yang berbeda, seperti pada tabel berikut ini:

Tabel 2 Keluarga ISO-8859-*x*

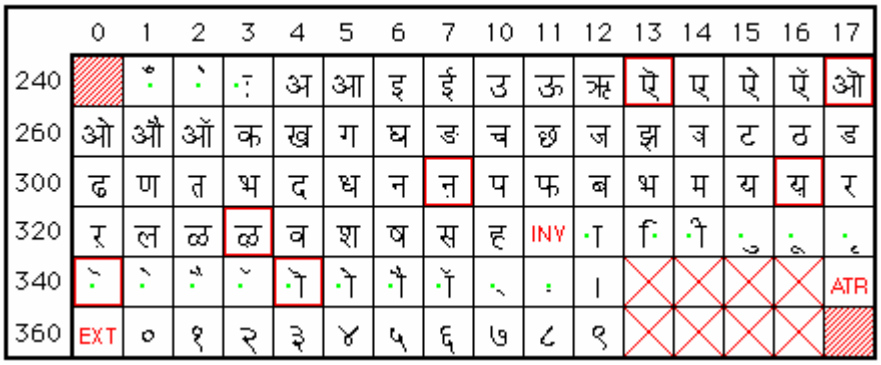
No	*Kode*	*Nama*	*Penjelasan*
1	ISO-8859-1	Latin-1 Western European	128 karakter untuk mengkodekan karakter-karakter dalam alfabet-alfabet Eropa Barat seperti Perancis dan Jerman (misalnya *é* dan *ö*).
2	ISO-8859-2	Latin-2 Central/East European	Mengkodekan alfabet negara-negara Eropa Timur kecuali Rusia
3	ISO-8859-3	Latin-3 South European	Mengkodekan alfabet negara-negara Eropa Selatan seperti Malta dan Esperanto
4	ISO-8859-4	Latin-4 North European	Mengkodekan alfabet negara-negara Eropa Utara seperti Swedia dan Denmark
5	ISO-8859-5	Latin/Cyrillic	Mengkodekan alfabet Cyrillic yang digunakan di Rusia
6	ISO-8859-6	Latin/Arabic	Mengkodekan alfabet Arab
7	ISO-8859-7	Latin/Greek	Mengkodekan alfabet Yunani, yang juga dipakai dalam simbol-simbol matematika
8	ISO-8859-8	Latin/Hebrew	Mengkodekan alfabet Ibrani yang digunakan di Israel
9	ISO-8859-9	Latin-5 Turkish	Mengkodekan alfabet Turki
10	ISO-8859-10	Latin-6 Nordic	Mengkodekan alfabet Nordic seperti alfabet Eslandia dan Eskimo
11	ISO-8859-11	Latin/Thai	Mengkodekan alfabet Thai. Perhatikan bahwa tidak ada ISO-8859-12
12	ISO-8859-13	Latin-7 Baltic/Rim	Mengkodekan alfabet Baltic yang digunakan di Lithuania, Estonia, dan Latvia
13	ISO-8859-14	Latin-8 Celtic	Mengkodekan alfabet Celtic/Gaelic yang masih dipergunakan di Skotlandia dan Irlandia
14	ISO-8859-15	Latin-9 with “euro”	Sama persis dengan Latin-1, kecuali perubahan pada 8 karakter, salah satunya adalah penambahan karakter simbol “euro”
15	ISO-8859-16	Latin-10 Collection of languages	Diciptakan untuk mengkodekan aksara Rumania

Semua set karakter tidak bisa dicampur dalam satu dokumen teks. Anda bisa menulis satu dokumen teks dalam Bahasa Inggris dan Jerman, akan tetapi anda tidak bisa menulis dalam Bahasa Jerman dan Bahasa Arab (dalam huruf Arab) dalam dokumen yang sama! Itulah salah satu keterbatasan dari ISO-8859-*x*. Keterbatasan lain adalah ketidakmampuan ISO-8859-*x* untuk mengkodekan alfabet yang memiliki jumlah huruf lebih banyak dari 256.

Set Karakter Non-Latin

Semua set karakter yang telah kita bahas sejauh ini hanya mampu mengkodekan alfabet Latin dan beberapa alfabet non-latin (Arab, Ibrani, dan Thai). Anda mungkin pernah melihat alfabet Devanagari (India) atau Kanji (Jepang) juga digunakan pada komputer. Set karakter apakah yang digunakan untuk mengkodekan alfabet-alfabet tersebut?

India memiliki sebelas alfabet resmi: Devanagari, Gujarati, Bengali, Punjabi, Oriya, Gurmukhi, Telugu, Tamil, Malayalam, Kannada, dan Assamese. Kesemuanya dirangkum dalam satu keluarga set karakter, yaitu ISCII (Indian Standard Code for Information Interchange). Konsep ISCII sama persis dengan konsep ISO-8859, yaitu 128 karakter pertama sama dengan ASCII, dan 128 karakter lain digunakan oleh salah satu dari alfabet-alfabet India tersebut di atas.



Gambar 2 ISCII untuk Devanagari[2]

Jepang, sebagai salah satu negara maju, memiliki set karakter sendiri untuk alfabetnya, yaitu JIS (Japan Industrials Standard). JIS merupakan sebuah set karakter dengan karakteristik "double-byte", dengan tiap karakternya memiliki lebar 16-bit. Dengan demikian, JIS mampu mengkodekan karakter sebanyak 2^{16} = 65.535 buah. Terdapat pula satu set karakter yang merupakan varian dari JIS, yang disebut dengan Shift-JIS.

Gambar 3 Contoh Huruf pada Alfabet Jepang[3]

Oleh karena tiap negara memiliki standar sendiri-sendiri (Jepang – JIS, India – ISCII , Thailand – TIS-620, Vietnam – VISCII, Taiwan – Traditional Chinese "Big5", Cina – Simplified Chinese GBK[4]) para vendor komputer mengalami kesulitan untuk mengimplemtasi set karakter pada produk-produk mereka. Apakah akan mendukung alfabet Thai dengan JIS atau Shift-JIS? Bagaimana dengan alfabet-alfabet lain

[2] Gambar ini diambil dari http://homepages.cwi.nl/~dik/english/codes/indic.html

[3] Kalau yang ini dari http://208.17.151.81/lafferty/Japanese_characters.htm

[4] Saya berharap suatu saat nanti Indonesia akan memiliki set karakter untuk huruf Jawa, Bali, Batak, dan Bugis

yang belum memiliki set karakter? Maka para vendor komputer [5] tersebut membentuk Unicode Consortium yang bertujuan untuk menciptakan satu set karakter untuk semua alfabet yang ada di dunia. Bagaimana caranya? Unicode akan dijelaskan secara ringkas pada artikel ini. Akan tetapi, kita akan membahas set karakter pada Microsoft® Windows™ terlebih dahulu.

Keluarga Set Karakter pada Sistem Operasi Windows

Karena alasan sejarah, Microsoft menggunakan istilah "codepage" sebagai pengganti "character set". Codepage 437 (juga dikenal dengan istilah "PC8") merupakan "ASCII", plus karakter-karakter untuk menggambar garis, yang digunakan oleh para programmer DOS.

Tabel 3 Keluarga Set Karakter pada Sistem Operasi Windows

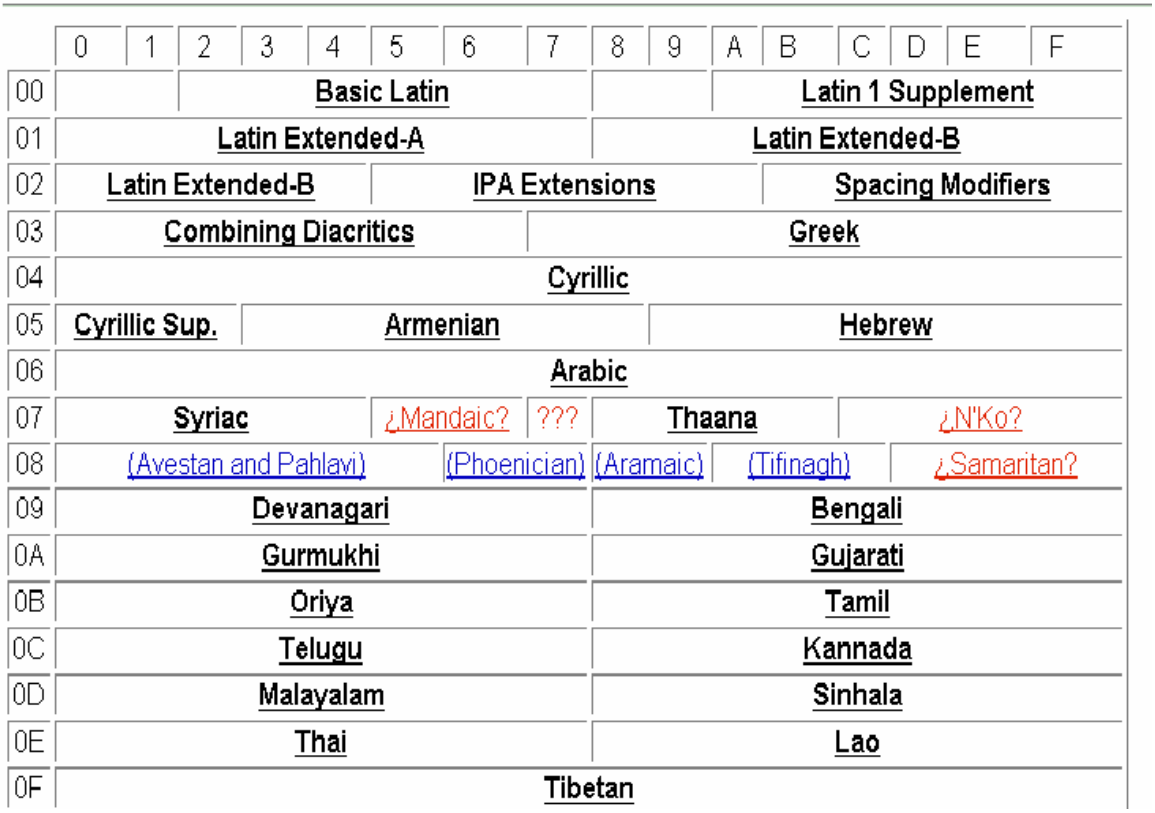
Codepage	***Ekivalen dengan***
windows-437	Set karakter yang merupakan superset dari ASCII, digunakan pada DOS dan Windows saja
windows-1250	ISO-8859-2 (Eropa Tengah dan Timur)
windows-1251	ISO-8859-5 (Rusia)
windows-1252	ISO-8859-1 (Eropa Barat)
windows-1253	ISO-8859-7 (Yunani)
windows-1254	ISO-8859-9 (Turki)
windows-1255	ISO-8859-8 (Ibrani)
windows-1256	ISO-8859-6 (Arab)
windows-1257	ISO-8859-13 (Baltic)
windows-1258	VISCII (Vietnam)
windows-874	TIS-620 (Thailand)
windows-932	Shift-JIS (Jepang)
windows-936	Simplified Chinese GBK (China)
windows-949	Korea
windows-950	Traditional Chinese Big5 (Taiwan)

Codepage Windows pada Tabel 2 tidak sama persis dengan set karakter yang ekivalen. Misalnya saja, windows-1252 memiliki sedikit perbedaan dengan ISO-8859-1.

Unicode

Set karakter Unicode mampu menampung lebih dari satu juta karakter (2^{20} = 1.048.576). Akan tetapi, saat ini hanya 65.535 karakter[6] yang pertama yang saat ini mampu direpresentasikan pada komputer. Karakter 0 s.d. 65.535 menampung karakter-karakter dari alfabet-alfabet yang belum punah (Latin, Kanji, Devanagari, dlsb.) sedangkan karakter 65.536 s.d. 1.048.575 menampung karakter-karakter dari alfabet-alfabet yang sudah punah (misalnya *hieroglyph* dan beberapa karakter Cina yang sangat jarang digunakan).

[5] Termasuk di antaranya IBM, Oracle, Microsoft, dan SAP
[6] 65.535 karakter pertama dari Unicode sering disebut dengan istilah UCS-2 (Universal Character Set-2)



Gambar 4 Sebagian Alokasi Alfabet pada Unicode

Seperti terlihat pada Gambar 4, set karakter Unicode dialokasikan untuk lebih dari satu alfabet. Bahkan, Unicode Consortium mentargetkan untuk mengkodekan seluruh alfabet yang ada di dunia.

Set karakter Unicode ini diharapkan dapat menjadi standar set karakter pada semua komputer di masa depan. Karena Unicode mampu merepresentasikan semua alfabet yang ada di dunia ini, maka secara teori seluruh set karakter lainnya tidak diperlukan lagi.

Unicode mampu mengkodekan berbagai karakteristik alfabet. Mulai dari alfabet Latin yang sederhana, alfabet Arab yang ditulis sambung-menyambung (*cursive*) dari kanan ke kiri, alfabet Cina yang ditulis dari atas ke bawah, dan alfabet India yang memiliki huruf vokal yang letaknya di atas-bawah-depan-belakang dari konsonan.

Oleh karena ke-superior-an dari Unicode, maka Unicode menjadi set karakter standar yang digunakan pada komunikasi data melalui Internet. Unicode memiliki lebar per karakter sebesar 20 bit. Akan menjadi sangat boros jika kita akan mengirim data Unicode yang berisikan teks huruf Latin saja menggunakan 20-bit per karakter. Oleh karena itu, maka Unicode perlu ditransformasikan terlebih dahulu menjadi UTF-8 atau UTF-16 (Unicode Transformation Format). Dengan UTF-8, maka karakter-karakter pada U+0000[7] s.d. U+007F (128 karakter pertama) dapat dikodekan menjadi satu byte saja; sedangkan karakter-karakter lainnya dikodekan dengan menggunakan antara 2 sampai 4 byte per karakter. UTF-8 dan UTF-16 digunakan hanya pada saat komunikasi data saja sedangkan Unicode dalam bentuk normal (20 bit) tetap digunakan pada representasi karakter di memori komputer.

[7] Notasi U+abcd digunakan untuk mengacu pada karakter bernomor *abcd* pada tabel Unicode. Sebagai contoh, U+0053 adalah LATIN CAPITAL LETTER S dan U+0584 adalah ARMENIAN SMALL LETTER KEH

Aplikasi Set Karakter

Sekarang anda sudah mengenal berbagai set karakter, berikut dengan karakteristiknya masing-masing. Apakah kegunaan langsung dari pengetahuan baru ini bagi anda? Orang Indonesia pada umumnya tidak menggunakan alfabet selain Latin pada komputer. Manfaat yang paling besar akan dirasakan oleh para praktisi teknologi informasi yang bergelut di bidang integrasi aplikasi. Biarpun sama-sama menggunakan aksara Latin, tranformasi antar set karakter seringkali masih harus dilakukan.

Anda juga akan bisa lebih memahami standar-standar yang ada di dunia jaringan komputer. RFC 1630 mensyaratkan ASCII untuk menyusun sebuah Internet URL, akan tetapi halaman web bisa dikodekan denga n UTF-8. Artinya, URL (sampai saat ini) tidak bisa ditulis dengan alfabet Cina atau Jepang, akan tetapi isi halaman web bisa ditulis dengan alfabet apa saja yang menggunakan Unicode.

Dan tentu saja, jika anda cinta pada budaya Indonesia, anda bisa memulai untuk mempelajari Unicode dan menyusun set karakter alfabet Jawa[8] atau alfabet daerah lainnya untuk dimasukkan ke Unicode.

Sampai jumpa pada artikel berikutnya. Tanggapan untuk artikel ini tolong dikirim ke alamat email deweu@yahoo.com. Terima kasih!

Referensi

Cukup banyak materi yang saya kutip dari halaman web ini: http://www.cs.tut.fi/~jkorpela/chars.html, dan dari http://www.unicode.org

Biografi

Lulus dari Teknik Informatika ITB tahun 2000. Sekarang bekerja untuk IBM Business Consulting Services dengan spesialisasi pada bidang EAI (Enterprise Application Integration) dengan *tools* IBM WebSphere Application Server, WebSphere MQ dan WebSphere MQ Integrator. Saat ini bertempat tinggal di Singapura.

[8] Alokasi untuk Aksara Jawa, Bali, Bugis dan Batak pada Unicode dapat dilihat pada website Unicode http://www.unicode.org/roadmaps/bmp/ .