情報基礎

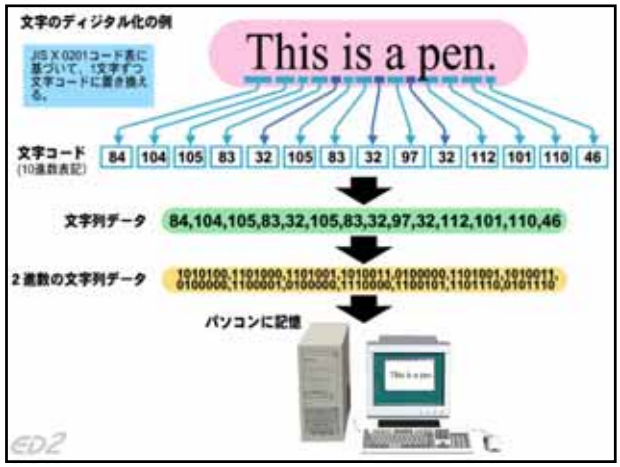
情報の符号化 (2)
文字コードとその周辺

今日やること

- コンピュータで文字情報をどのように扱うか
  - 文字コード
  - 各種エンコーディング
    - ISO-2022-JP, Shift_JIS, EUC-JP
  - 電子メールやWebと文字コードの関係

文字の扱いかた

- コンピュータで(数値だけでなく)文字情報も扱いたい!
  - コンピュータは数値しか扱えない
  - 文字をどうやって扱うか?



文字コード

- 文字コード: 文字に割り当てた数値
- “This is a pen.” ⇒ 84 104 105 83 …
  - 符号化 (エンコード, encode)
  - 何かを数値に置き換える(コード化する)こと
- 84 104 105 83 … ⇒ “This is a pen.”
  - 復号 (デコード, decode)
  - 数値から元に戻すこと

制御コード

- 文を表現するには,「改行」が必要

“Do you know Tom Riddle?”
“Yes”

改行!
改行!

- 「改行」のような見えないものにもコードを割り当てて表現する
  - 制御コード

文字コードの重要性

- 誰かと文字情報をやり取りするためには，お互いに同じ文字コードにする必要がある
  - みんなが勝手な文字コードを使うと困る
  - お互いに利用する文字コードに対する合意が必要
  - 違う文字コードを使うと文字化けが発生する

ASCIIコード

- American Standard Code for Information Interchange
  - アスキー
- 1963年ANSI (American National Standards Institution)が制定
- アメリカで必要な文字を集めて7ビットで表現
  - 7ビットなので128種類の文字を表現可能
  - 0x00〜0x7F を使用する
  - 8ビットで使用するときは，最上位ビット(MSB)を0にしておく
- コンピュータの最も基本となる文字コード

ASCIIコード表

下位の桁

各種制御コード

上位の桁	0	1	2	3	4	5	6	7	8	9	A	B	C	D	E	F
00	制御文字領域															
10																
20		!	"	#	$	%	&	'	(	)	*	+	,	-	.	/
30	0	1	2	3	4	5	6	7	8	9	:	;	<	=	>	?
40	@	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
50	P	Q	R	S	T	U	V	W	X	Y	Z	[	\	]	^	_
60	`	a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
70	p	q	r	s	t	u	v	w	x	y	z	{	|	}	~	

0x20はスペース(“ ”)

よく使われる制御コード

- 0x09 水平タブ (Horizontal Tabulation, HT)
  (水平方向に一定の桁まで移動.
  Tabキーで入力)
- 0x0a 改行 (Line Feed, LF)
  (一行紙を送る)
- 0x0d 復帰 (Carriage Return, CR)
  (紙の行頭へ戻す)
- 0x1b エスケープ (Escape, ESC)
  (後述)

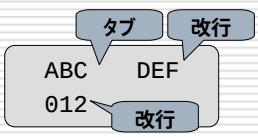
テレタイプで使用していた名残

改行コードの違い

- Windows
  - CR + LF (0x0d + 0x0a)　復帰+改行 (複改)
- UNIX系 (Linux, MacOS X など)
  - LF (0x0a)　改行

演習

- Linux上のエディタ(kwrite)で次のようなファイルを作成し適当なファイル名でセーブ



- 16進エディタ(khexedit)でファイルを覗いてみよ
  - 文字やアルファベットとASCIIコード表を照合せよ
  - タブや改行のコードを確認せよ

演習

- 次のバイト列をASCIIコード列とみなし，何が書いてあるのかを読み取れ(16進数)
  - ASCIIコード表を参照せよ
  - 78 3d 32 33 0a 79 3d 78 2b 35 39 0a
  - またこのバイト列はどの環境で作られたものか? (Windows, UNIX)

JIS X 0201

日本工業規格
Japanese Industrial Standard

- 「7ビット及び8ビットの情報交換用符号化文字集合」
- ASCIIを拡張
  - 0x5c バックスラッシュ(＼) を円記号(\)に変更
    - 英語版Windowsではディレクトリの区切りは ＼
    - 日本語版Windowsでは \
  - 0x7e チルダ(～)をオーバースコア(̄)に変更
- いわゆる半角カナを追加
  - 濁点(゛)や半濁点(゜)つきの文字は2文字で表す
    例: ｱﾎﾞｶﾞﾄﾞ

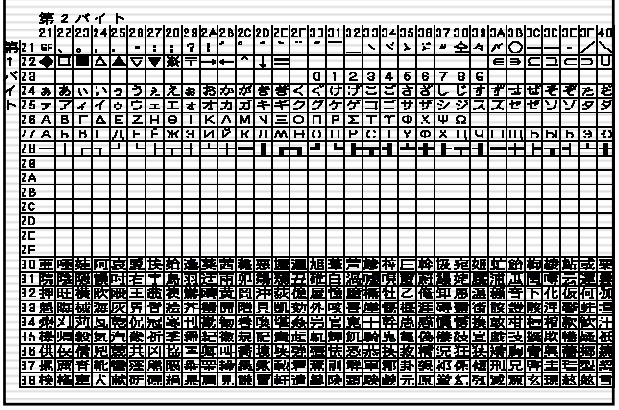
JIS漢字コード(JIS X 0208) (1)

- 「7ビット及び8ビットの2バイト情報交換用符号化漢字集合」
- いわゆる全角文字
- 日本語情報処理の基本
  - 常用漢字，人名用漢字を含む
- 約7000文字
- 収録文字種
  - 各種記号，アラビア数字，ローマ字，ひらがな，カタカナ，ギリシャ文字，キリル文字，罫線素片
    第1水準漢字，第2水準漢字

JIS漢字コード(JIS X 0208) (2)

- 2バイトで1文字を表現
  - 第1バイト+第2バイト
  - それぞれ0x21～0x7eを使う
- 歴史
  - 1978年制定 78JIS (旧JIS)
  - 1983年改訂 83JIS (新JIS)
- コード表:
  - http://www.infonet.co.jp/ueyama/ip/binary/x0208txt.html

JIS漢字コード表(一部)



関連規格

- JIS X 0212 (補助漢字)
  - “¡”, “©”, …
  - ほとんど使われていない
- JIS X 0213 (JIS2000)
  - 「7ビット及び8ビットの2バイト情報交換用符号化拡張漢字集合」
  - 第3水準，第4水準 (4344文字)
  - JIS X 0208 と一緒に使う
    - “ゕ”, 半濁音つき「かきくけこ」など
  - まだ普及していない

注意すべきこと

- 包摂
  - どこまで同一の文字と見なすか
  - クチ高とハシゴ高
  - 土吉と士吉
    - 吉野家の吉は？

髙
吉

- JIS X 0201 と JIS X 0208で重複する文字が存在
  - アルファベット，数字，カタカナ，空白，記号
  - コードは全く異なる
    - コンピュータからは異なる文字と見なされる

各種エンコーディング

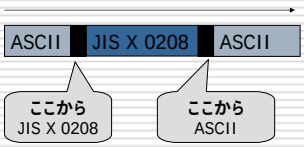
- JIS漢字とASCII文字などを混在して使用したい
- 各種のエンコーディング
  - エンコーディング：文字コードをファイルに格納 (or ネットワークで伝送) する方式のこと
- ISO-2022-JP
  - 電子メールなどで用いられる
- Shift_JIS, EUC-JP
  - エスケープシーケンスが不要
    - エスケープシーケンスの処理は大変!
  - コンピュータ内部でよく用いられる

ISO-2022-JP

International Standardization Organization

- 電子メールなどで広く使用
  - 半角カナは使えない
- JIS漢字コードとISO-2022-JPを同一視する場合もある
- JIS漢字コードとASCIIはどちらも0x21〜0x7eを使用するので…
  - ASCIIとJIS漢字を切り替えるマークを挿入
    - エスケープシーケンス

エスケープシーケンス



ASCII	ESC (B
JIS X 0201の0x20〜0x7e	ESC (J
JIS X 0208(78年版)	ESC $ @
JIS X 0208(83年版)	ESC $ B

0x1b

実験

- エディタ(kwrite)で，次のようなテキストを入力せよ

ABCあいう
えおDEF
改行
改行

- セーブするときにエンコーディングとして jis7 を選択してセーブせよ
- khexedit でエスケープシーケンスがどうなっているか確認せよ

Shift_JIS

- エスケープシーケンス不要
- JIS X 0201の空き領域を使用
  - JIS X 0201と混在可能
  - いわゆる半角カナが使える
  - 漢字は2バイトで表現
    - 第1バイトは0x81〜0x9F, 0xE0〜0xEFの範囲
- 日本語版Windowsの標準
  - MS漢字コードとも呼ばれる
  - 日本のデファクトスタンダード (事実上の標準)
- コード表: http://www.rtpro.yamaha.co.jp/RT/docs/misc/kanji-sjis.html 等

EUC-JP

- EUC = Extended UNIX Code (拡張UNIX符号)
  - EUC-JP(日本語EUC), 韓国語EUC(EUC-KR)などがある
- エスケープシーケンス不要
- UNIX系OSで広く用いられている
- ASCIIと重ならないので扱いやすい
- コード表: http://www.rtpro.yamaha.co.jp/RT/docs/misc/kanji-euc.html 等

実験

- Kwriteを使って, 先ほどのテキストをShift_JIS, EUC-JPでセーブし, khexeditで観察してみよう

演習

- コード表を参照し, 次のShift_JISコードで書かれた文字列を解読せよ
  - 82 b2 96 bc 93 9a 21 21 0a
- この文字列はWindows, UNIXどちらの環境で作られたものか?

半角と全角

- 俗に,
  - 半角文字: ASCII や JIS X 0201 (ｺﾝﾅﾓｼﾞ)の文字
  - 全角文字: JIS X 0208 (JIS漢字)の文字

  と呼ぶことがある.
- 昔はASCII文字などの幅をJIS漢字の幅の半分にするのが一般的だったため
- 実際は使用するフォントによって文字の幅は異なるので, 半角・全角と呼ぶのは避けたほうが良いという意見もある

外字

- コード表の空き領域を用いて文字を定義
- 機種依存文字
  - メーカーが独自に定義
  - 丸付き数字 ①②③④ など
    - Windowsの拡張(元々はNECの拡張)
    - Macintoshでは同じコードで丸付き曜日 (月)(火)(水)が表示される
  - iモードの絵文字
- ユーザ定義文字
  - ユーザが独自に定義
- 情報交換の障害となる
  - 電子メールやWebページで使わないこと

Unicode(1)

- 様々なエンコーディング
  - 韓国　EUC-KR
  - 中国　GB18030
  - 台湾　BIG5
  - タイ　TSCII
  - ヨーロッパ　ISO-8859-1 等
- 多言語対応のソフトウェアを作るのが大変!

Unicode(2)

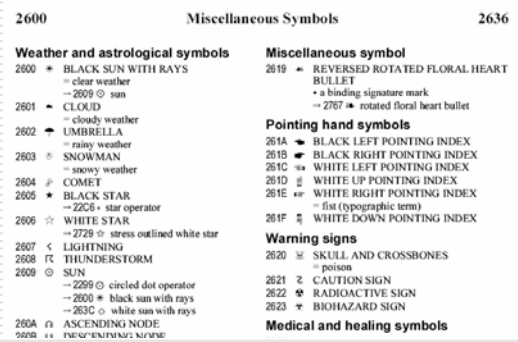
- 世界中の全ての文字を網羅した文字コードを作ってしまおう
  - Microsoft, Apple, Sun Microsystems, etc.
  - Unicode Consortium http://www.unicode.org/
  - Windows2000やXP, MacOS X は内部Unicode
  - Java言語もUnicodeが標準
  - これからの標準になる見込み
  - 最近のエディタはUnicodeをサポートしているものも多い

Unicode(3)

http://www.unicode.org/charts/

- Basic Latin
- Latin-1 Supplement
- Latin Extended-A
- Latin Extended-B
- IPA Extensions
- Spacing Modifier Letters
- Combining Diacritical Marks
- Greek
- Cyrillic
- Cyrillic Supplement
- Armenian
- Hebrew
- Arabic
- Geometric Shapes
- Miscellaneous Symbols
- Dingbats
- Miscellaneous Mathematical Symbols-A
- Supplemental Arrows-A
- Braille Patterns
- Supplemental Arrows-B
- Miscellaneous Mathematical Symbols-B
- Supplemental Mathematical Operators
- Miscellaneous Symbols and Arrows
- CJK Radicals Supplement
- Kangxi Radicals
- Ideographic Description Characters

Unicode(4)



2600 Miscellaneous Symbols 2636

Weather and astrological symbols

2600 ☀ BLACK SUN WITH RAYS
= clear weather
→ 2609 ☉ sun
2601 ☁ CLOUD
= cloudy weather
2602 ☂ UMBRELLA
= rainy weather
2603 ☃ SNOWMAN
= snowy weather
2604 ☄ COMET
2605 ★ BLACK STAR
→ 22C6 ⋆ star operator
2606 ☆ WHITE STAR
→ 2729 ☆ stress outlined white star
2607 ☇ LIGHTNING
2608 ☈ THUNDERSTORM
2609 ☉ SUN
→ 2299 ⊙ circled dot operator
→ 2600 ☀ black sun with rays
→ 2639 ☼ white sun with rays
260A ☊ ASCENDING NODE

Miscellaneous symbol

2619 ❧ REVERSED ROTATED FLORAL HEART BULLET
• a binding signature mark
→ 2767 ❧ rotated floral heart bullet

Pointing hand symbols

261A ☚ BLACK LEFT POINTING INDEX
261B ☛ BLACK RIGHT POINTING INDEX
261C ☜ WHITE LEFT POINTING INDEX
261D ☝ WHITE UP POINTING INDEX
261E ☞ WHITE RIGHT POINTING INDEX
= fist (typographic term)
261F ☟ WHITE DOWN POINTING INDEX

Warning signs

2620 ☠ SKULL AND CROSSBONES
= poison
2621 ☡ CAUTION SIGN
2622 ☢ RADIOACTIVE SIGN
2623 ☣ BIOHAZARD SIGN

Medical and healing symbols

Unicode(5)

- (最初は)16ビット固定長(最大65536文字)
  - 足りなくなってきたので今は違う
  - 文字をU+261A のように表記する
- JIS漢字コードに含まれる文字は全て収録されている
- コードを節約するために日本，中国，韓国の漢字を一旦バラバラにして統合
  - CJK統合漢字 (Chinese-Japanese-Korean)
  - JIS漢字との変換には変換表が必要
  - 日本語と中国語を混ぜられない?
    - 「高低」と「高低」が同じコードになってしまう

Unicode(6)

- Windows2000やXPでは「文字コード表」で閲覧可能
  - プログラム⇒アクセサリ⇒システムツール
  - 適当なフォントを選ぶこと
    - 日本語: MSゴシック
      中国語: SimSun
      韓国語: Gulim

Unicode(7)

- Unicodeで使われるエンコーディング
  - UTF-8
    - 1〜6バイトの可変長でエンコード
    - ASCII文字は1バイトで済む
  - UTF-16
    - (基本的に)2バイト固定長でエンコード

文字コードの周辺

プレーンテキストファイルとバイナリファイル

- プレーンテキストファイル
  - 以下の条件を満たすテキストファイル
    - 標準的なエンコーディングのみを使用
    - アプリケーション独自の制御コードを含まない
  - フォントを変更したり，センタリングしたりはできない
  - いろいろなアプリケーションで使える
- バイナリファイル
  - プレーンテキストファイルでないもの
  - ワープロのファイル，画像，音楽 etc.
- テキストエディタ
  - kwrite，メモ帳，etc.
  - プレーンテキストファイルを扱う
- バイナリエディタ
  - khexedit，etc.

Webページと文字コード

- <meta>タグを使ってエンコーディングを指定
  - ブラウザはこれを頼りに表示する
- 文字化けの原因
  - Metaタグを使用していない
  - Metaタグのエンコーディング指定と本文のエンコーディングが異なっている

```
<meta HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=ISO-2022-JP">
<meta HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=EUC-JP">
<meta HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=Shift_JIS">
```

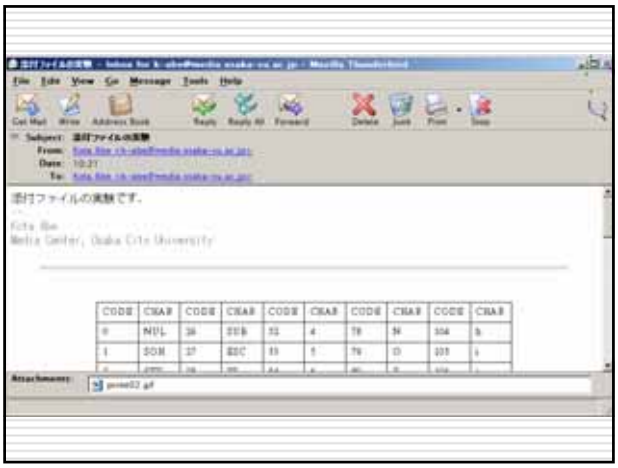
電子メールと文字コード

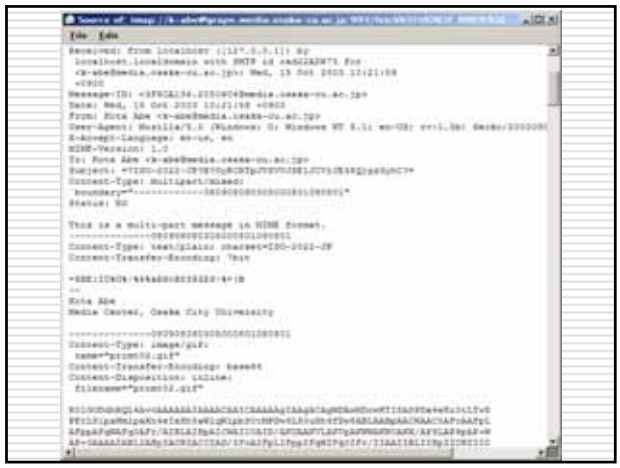
- 原則として7ビットコードを使う
- 世界では様々なエンコーディングが使用されている
  - 本文のエンコーディングをヘッダ部で指定
- 日本語では ISO-2022-JP が標準
  - いわゆる半角カナは使えない
- 電子メールソフトで「ヘッダの表示」をしてみよう

```
Date: Thu, 09 Oct 2003 12:01:22 +0900
From: Kota Abe <k-abe@media.osaka-cu.ac.jp>
Content-Type: text/plain; charset=ISO-2022-JP
Content-Transfer-Encoding: 7bit
```

電子メールの添付ファイル

- 電子メールでバイナリファイルを送りたい
  - いわゆる「添付ファイル」機能
  - 静止画，動画，ワープロのファイル，etc.
- 電子メールを配送するソフトウェアは「文字」にしか対応していない
- バイナリファイルも文字に変換





Base64エンコーディング

- バイト列を文字に変換する
  - バイト列を6ビット毎に区切る
    - 3バイトを4つに区切る
- 6ビットで0〜63まで表現可能
- 0〜63に，異なる文字を割り当てる(64進数)
- サイズはほぼ4/3倍

Value	Encoding	Value	Encoding	Value	Encoding	Value	Encoding
0	A	17	R	34	i	51	z
1	B	18	S	35	j	52	0
2	C	19	T	36	k	53	1
3	D	20	U	37	l	54	2
4	E	21	V	38	m	55	3
5	F	22	W	39	n	56	4
6	G	23	X	40	o	57	5
7	H	24	Y	41	p	58	6
8	I	25	Z	42	q	59	7
9	J	26	a	43	r	60	8
10	K	27	b	44	s	61	9
11	L	28	c	45	t	62	+
12	M	29	d	46	u	63	/
13	N	30	e	47	v		
14	O	31	f	48	w	(pad)	=
15	P	32	g	49	x		
16	Q	33	h	50	y		

参考文献

- 小形克宏の「文字の海、ビットの舟」
  ―― 文字コードが私たちに問いかけるもの
  - http://internet.watch.impress.co.jp/www/column/ogata/index.htm
- 文字コードの話
  - http://euc.jp/i18n/charcode.ja.html
- 深沢千尋，文字コード「超」研究，株式会社ラトルズ，2003年

削除したスライド

混乱(2)

- パソコンのカタログから (富士通FMV-DESKPOWER)

色彩が異なる場合があります。●本カタログに記載の製品のディスク容量は、1MB=1000^{2}byte、1GB=1000^{3}byte換算値です。1MB=1024^{2}byte、1GB=1024^{3}byte換算のものとは表記上同容量でも、実容量は少なくなりますのでご注意ください。●フロッピーディスクは、フォーマットした環境（メーカ、機種、ソフトウェア）が違うと、データを読み

- 3.5インチフロッピーディスクの容量は“1.44MB”．このMBは1024*1024でも1000*1000でもない．
  - 1024*1000バイト (!)

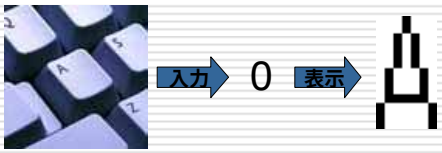
新しい単位

- 1998年 IEC (the International Electrotechnical Commission) が新しい単位を制定
  - 1Ki = 1Kibi = 2^{10} = 1,024
    - Kibi = Kilo Binary
  - 1Mi = 1Mebi = 2^{20} = 1,048,576
    - Mebi = Mega Binary
  - 1Gi = 1Gibi = 2^{30} = 1,073,741,824
    - Gibi = Giga Binary
- まだ一般的ではない

ミニレポート

- 演習(1)〜(4)を次の授業の前までに提出
  - 提出方法は別途指示します
- 質問･感想などを含めてもよい

文字コード(2)



0から“A”が表示される理由

(文字コードとフォントの関係)

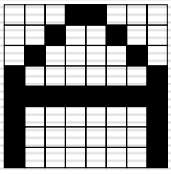
- 文字をディスプレイやプリンタ等に出力するときに使われる書体をフォント(font)と呼ぶ
- フォントデータは，文字コードの順番に並べられている

0	1	2	3	…
“A”の形	“B”の形	“C”の形	“D”の形	…

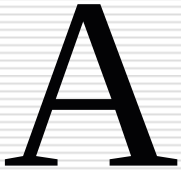
- フォントデータを使って文字の形を画面上に描画することで文字を表現する
- コンピュータは “A” の意味は知らない

ちょっと寄り道(フォントの話)

- ビットマップフォント
  - 高速に描画できる
  - 昔は一般的だった
  - 拡大するとギザギザ
- ベクトルフォント
  - 描画に時間がかかる
  - 最近は一般的
  - 拡大しても綺麗



0x18
0x24
0x42
0x81
0xFF
0x81
0x81
0x81

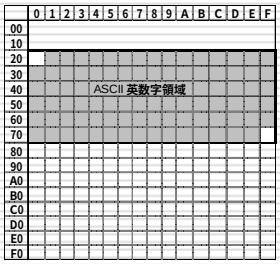


ミニレポート

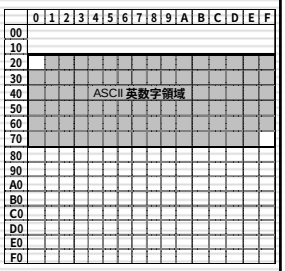
- 宿題(1)〜(2)を次の授業の前までに提出
- 質問･感想などを含めてもよい

JIS

JISコード 1バイト目

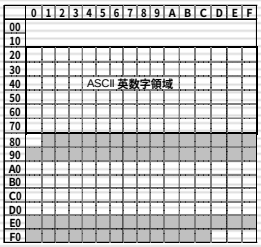


JISコード 2バイト目

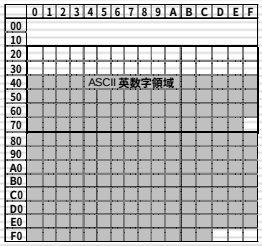


Shift_JIS

Shift-JISコード 1バイト目

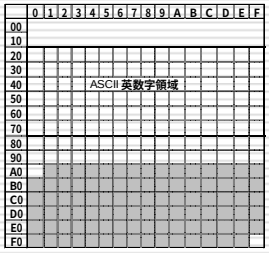


Shift-JISコード 2バイト目



EUC-JP

EUCコード 1バイト目



EUCコード 2バイト目