情報処理システム論 (12)

文字の扱い

- 地域化(Localization)
  - ある地域の言語が扱えるようにする
  - L10N
- 国際化(Internationalization)
  - 通貨単位、日付、時間も考慮する
  - 言語環境(locale)により地域を切り替え
  - I18N
- 多言語化(Multilingualization)
  - M17N

多言語化の程度(1)

- 表示
  - フォントが存在する
  - フォントを切り替えることができる
- 入力
  - キーボードが対応している
  - (漢字変換サーバがある)
- 印刷
  - (プリンタにフォントがある)

多言語化の程度(2)

- システム起動時に切り替え
- アプリケーション起動時に指定
- アプリケーション動作中に切り替え
- 複数言語を同時使用可能

図形文字と文字コード

- 図形文字
  - A
  - 京
- 文字コード
  - A : 4/1 (0100 0001, 0x41)
    - ASCII文字コード体系の場合
  - 京 : 3/5 7/14 (0011 0101 0111 1110, 0x357e)
    - JIS X 0208 情報交換用漢字符合

文字集合とコード系(1バイト)

- ASCII
  - ANSI(米国規格協会)による規格
  - ISO646 BCT(Basic Code Table) 国際標準
  - 94文字
- ISO 8859
  - ASCIIにヨーロッパ系の文字を追加したもの
  - 8ビットを使い、最上位ビットが1の部分に割当て
  - ISO 8859-1～10
  - 94+96文字

ASCII

- 7ビットで表現し、最上位ビットは0。

00「　　32個の制御文字領域
01　　」
02　！”＃＄％＆’（）＊＋，－．／
03 ０１２３４５６７８９：；＜＝＞？
04 ＠ＡＢＣＤＥＦＧＨＩＪＫＬＭＮＯ
05 ＰＱＲＳＴＵＶＷＸＹＺ［＼］＾＿
06‘ａｂｃｄｅｆｇｈｉｊｋｌｍｎｏ
07 ｐｑｒｓｔｕｖｗｘｙｚ｛｜｝

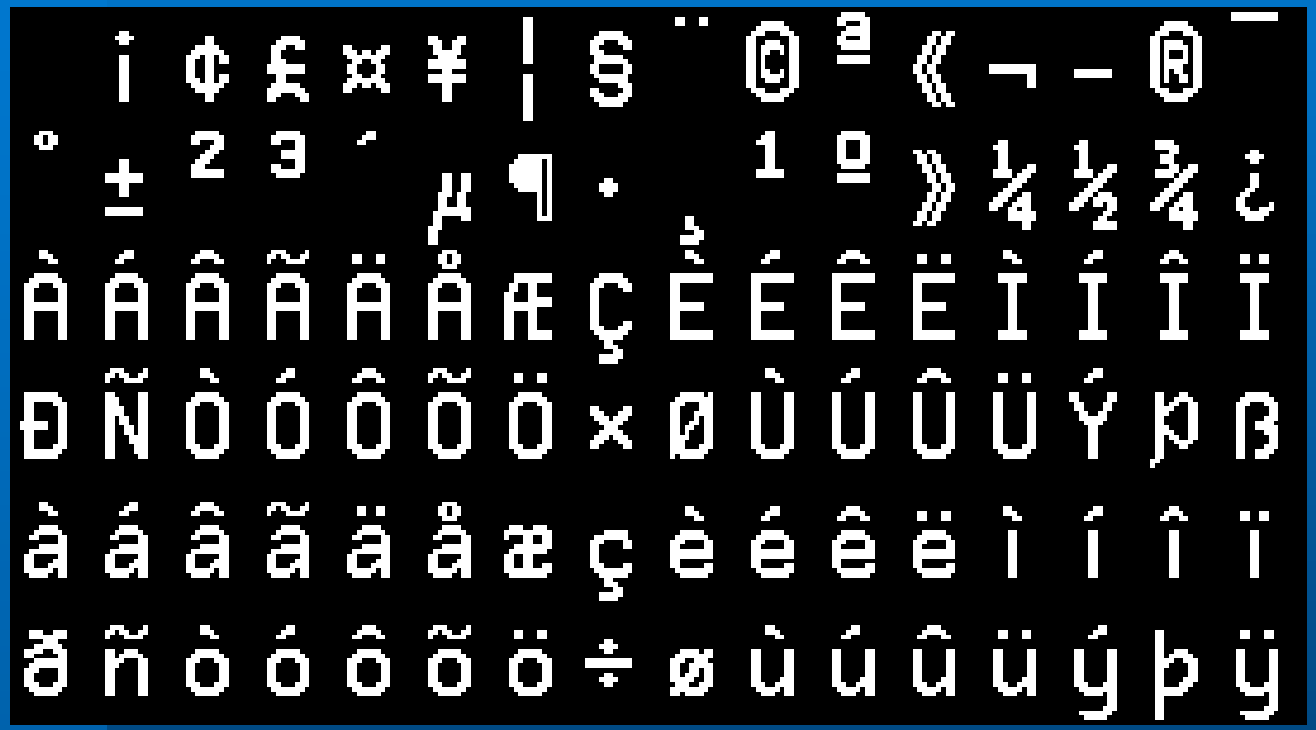
ISO646 BCT(Basic Code Table)

- **ASCII と異なる 12 文字**

2/3	2/4	4/0	5/11	5/12	5/13	5/14	6/0	7/11	7/12	7/13	7/14
#	$	@	[	\	]	^	`	{	|	}	~
#	$	@	[	¥	]	^	`	{	|	}	‾
£	$	@	[	\	]	^	`	{	|	}	~

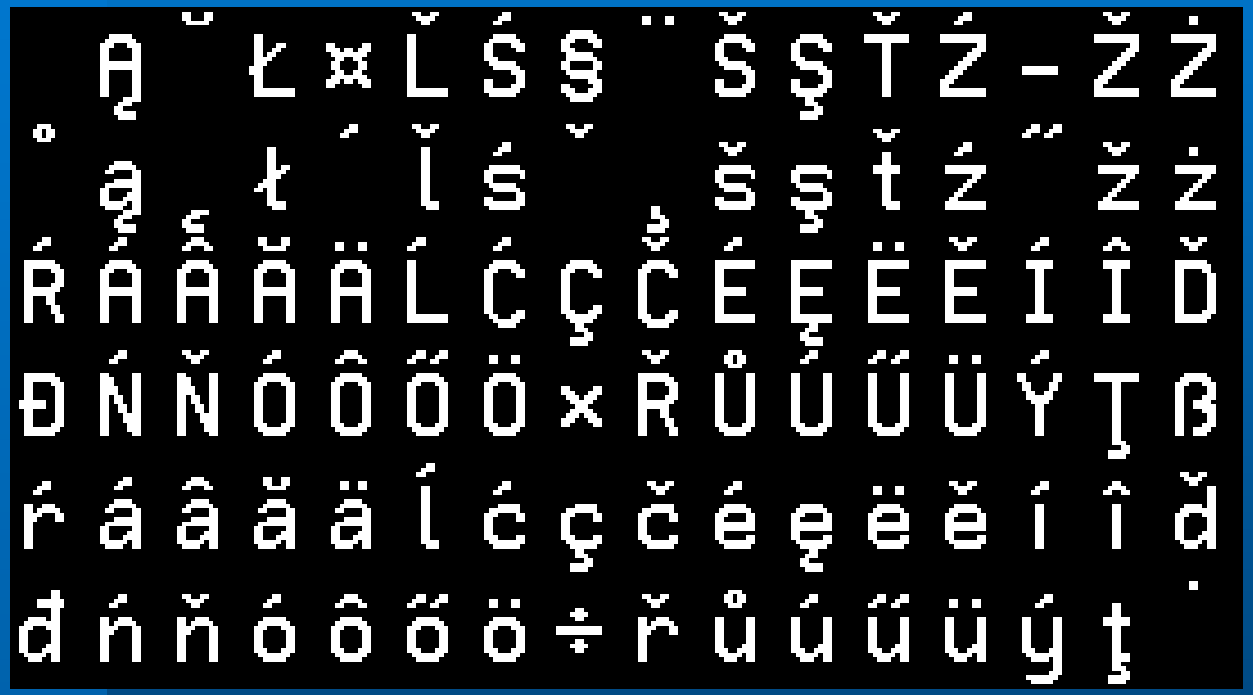
上が ASCII
中央が JIS X0201ローマ字
下が BSI 4730

ISO 8859-1



アイスランド、アイルランド、イタリア、オランダ、スウェーデン、スペイン、
デンマーク、ドイツ、ノルウェー、ポルトガル、フィンランド、フェロー、フランス

ISO 8859-2



アルバニア、スロバキア、スロベニア、チェコ、ドイツ、ハンガリア、ポーランド、ルーマニア

JIS X0201片仮名

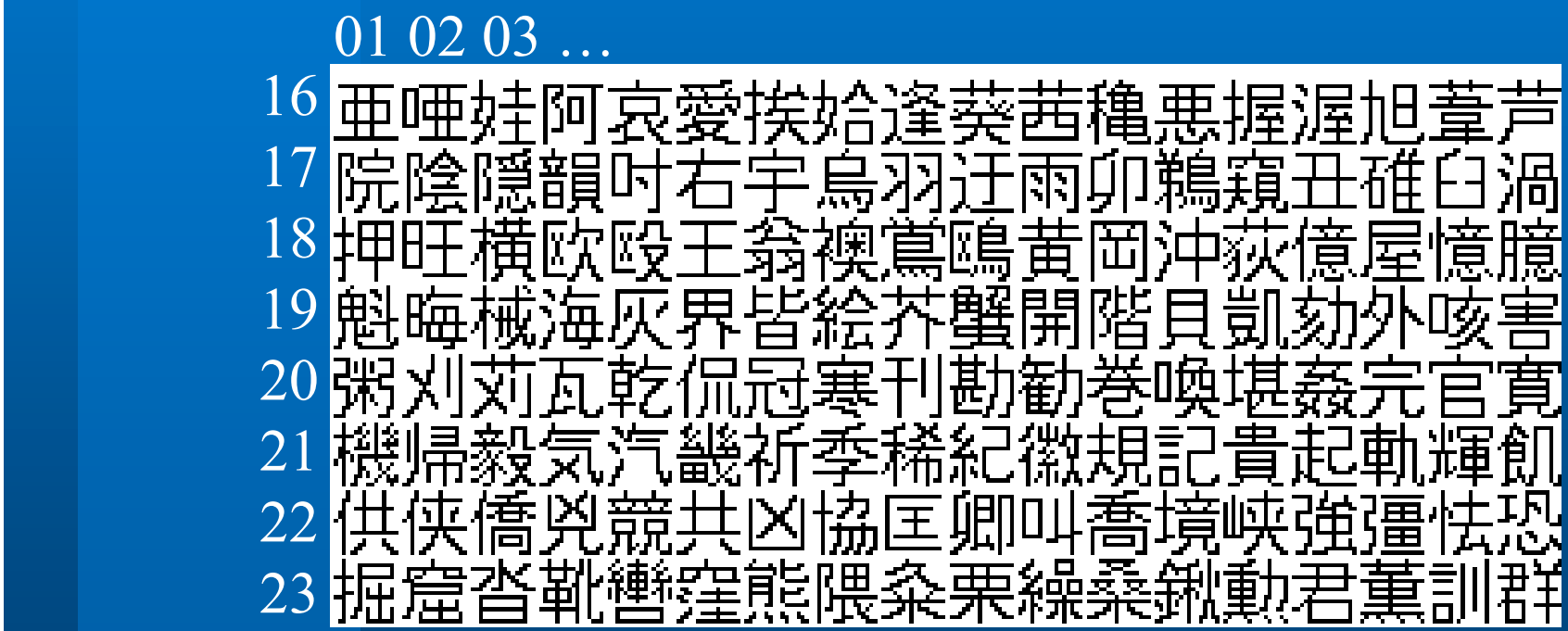
- **Shift JISでは1バイトコードの後半に割り当てて利用**

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
10		。	「	」	、	・	ヲ	ァ	ィ	ゥ	ェ	ォ	ャ	ュ	ョ	ッ
11	ー	ア	イ	ウ	エ	オ	カ	キ	ク	ケ	コ	サ	シ	ス	セ	ソ
12	タ	チ	ツ	テ	ト	ナ	ニ	ヌ	ネ	ノ	ハ	ヒ	フ	ヘ	ホ	マ
13	ミ	ム	メ	モ	ヤ	ユ	ヨ	ラ	リ	ル	レ	ロ	ワ	ン	゛	゜

多バイトコード

- 94+96文字で表現できない文字集合は多バイトで表現する
  - 日本語
  - 韓国語
  - 中国語（簡体字、繁体字）
- 区点コード（日本語）
  - 区（2桁、1－94）×点（2桁、1－94）＝ 8836
  - コードの開始は 0x2121（漢字の空白）

日本語 (JIS X 0208-1983)



数字は区点コード

日本語といっても…

- 文字集合にはいくつも種類がある
  - JIS X 0208-1978
  - JIS X 0208-1983
    - 字体の変更、記号の追加
  - JIS X 0208-1990（現在の最新）
  - JIS X 0212-1990（補助漢字）

ASCIIと日本語の混在方式

- そのままだと、同じコード空間を共有するので複数の文字セットを区別して扱うことができない
- 区別して扱うための拡張
  - ISO-2022-JP
  - EUC-Japan
  - Shift JIS
  - UNICODE...

ISO-2022-JP

- インターネットでのメッセージ交換で用いる
- RFC1468
- 「10月10日」は、

  1 0 ESC $ B 7 n ESC (B 1 0 ESC $ B F | ESC (B
- ESC $ B (0x1B, 0x24, 0x42)
  - JIS X 0208 への切り替え
- ESC (B (0x1B, 0x28, 0x42)
  - ASCII への切り替え

EUC-Japan

- **EUC：Extended UNIX Code**
  - 多くの UNIX システムで用いられている（いた?）
- 多バイトコードの文字セットについては、最上位ビットを1にして扱う
- 「10月10日」は、
  - 31 30 B7 EE 31 30 c6 FE
  - 1　0　7　n　1　0　F　|

    アンダーラインは実際には最上位ビットが1

Shift JIS

- パソコンで広く用いられている特殊な拡張方式
- 1バイトカタカナを最上位ビットが1の部分に定義し、隙間に漢字を配置
- 漢字の2バイト目だけを見て漢字の一部かどうかを判断することができない
- 「10月10日」は、

  30 31 8E 8C 30 31 F1 93

自動判別...

- 自動判別で区別できない文字
  - 鮃 (E9B7 in Shift JIS)
  - 薹 (E9B7 in EUC Japan)

参考文献

- マルチリンガル環境の実現
  プレンティスホール/トッパン
  4300円
  ISBN4-88735-020-1
- http://m-media.kudpc.kyoto-u.ac.jp/~yasuoka/CJK.html
- http://www.unicode.org