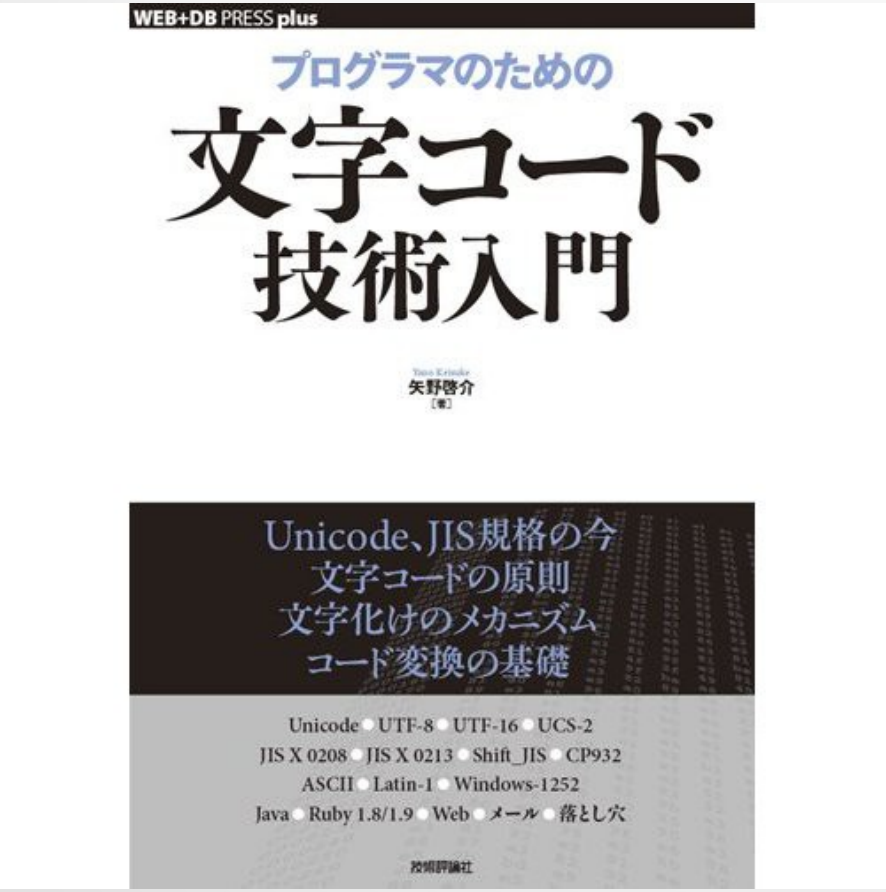
文字コード(1)

林部祐太 (NAIST)

@国立国会図書館 関西館 電子図書館課
2013/7/26

参考書

『プログラマのための文字コード技術入門』（技術評論社、2010年）



※特に注記がない場合，本スライドの図表は本書からの引用

2

予定

- 第1回　文字コードとは (2013-7)
  - 文字とコンピュータ
  - 文字コードの編成 (文字集合・符号化文字集合・文字符号化方式)
- 第2回　代表的な符号化文字集合・文字符号化方式 (2013-9)
  - 代表的な文字符号化方式(EUC-JP, ISO-2022-JP, Shift_JIS, UTF-16, UTF-32, UTF-8)
- 第3回　インターネットと文字コード (2013-11)
  - 文字コードの自動判別の仕組み
  - インターネットと文字コードの関係

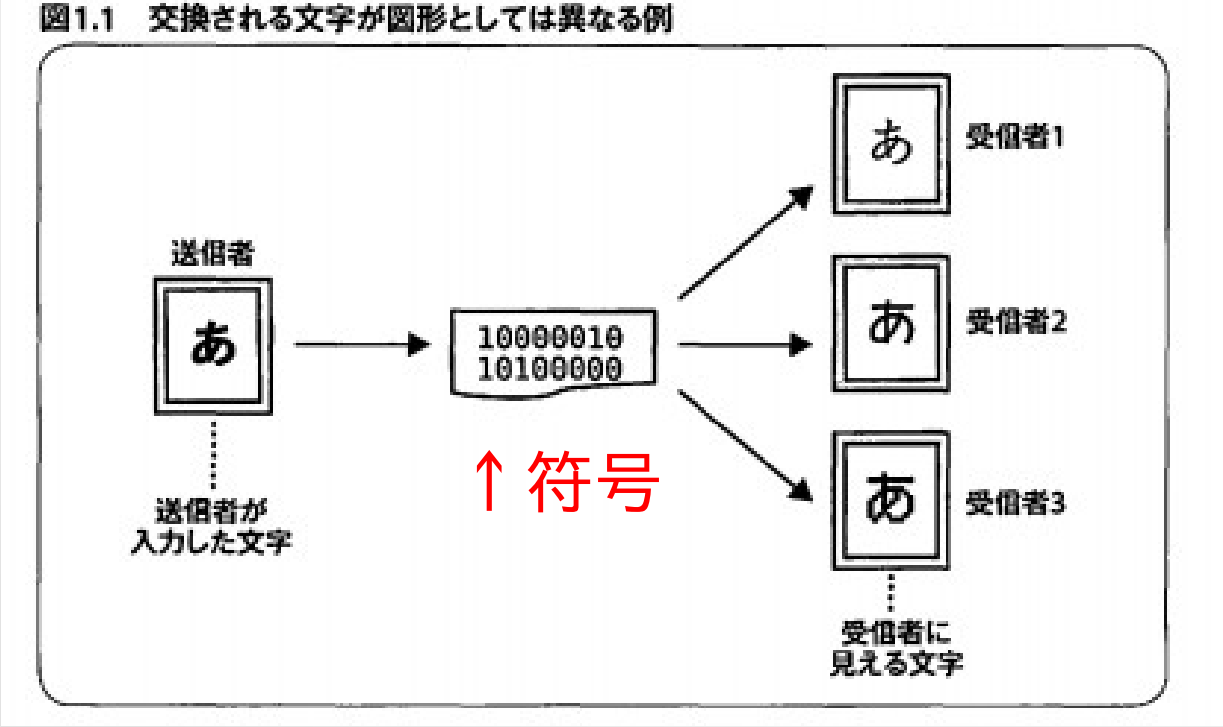
導入

文字とは

- そもそも文字とは？
  - 詳細な議論はしない
  - 「視覚的に伝達されるもの」
- 文字コード
  - 文字種を，図形としての細部を抽象化
  - それをコンピュータで扱うための符号

「文字」をコンピューターで表現/やりとりする

- 図形を交換するのでなく、符号を交換する
- 文字の図形の細部は伝わらない
  - 送信者と受信者で同じ形になるとは限らない



「コンピュータでは
0と1で表す」
少し説明します

コンピュータ内部での数の表現

- 「コンピュータは0と1の2進数」
  - よく言われるけど，具体的には??
- 内部的に，電圧が高いか・低いかで，2状態を表現
  - もちろん，3分割して，「高・中・低」も可能だが…
- 2進数のメリット
  - ノイズに強い
  - 構成部品が簡単
  - 演算装置の作成が簡単

ビットとバイト

0 0 0 1 1 0 0 1
=>8bit = 1byte

- 2進数の
  - 1けたをbit
  - 8けたをbyte

0 0 0 1 1 0 0 1 (2進)
<=> 25=2^4+2^3+2^0 (10進)

- コンピュターは, byte単位で処理
- 1bitは2種類の情報を表現できる
  - Nbit は 2^n種類の情報を表現できる

文字	対応するビット列の並び
A	001
B	010
C	011
D	100
E	101

16進数表記

- 沢山0と1を並べるのは大変…
- そこで, 4bitを16進数の1ケタと見なして
- 16進数
  - 0, 1, …, 8, 9, A, B, C, D, E, F
- 例
  - 0000 0001 0000 0011 1111 0000
  - 01 03 F0
  - **0x**0103F0 (16進であることを明示するとき)

ちょこっと演習

- それぞれ何種類の状態を表現できますか?
  - 3bit
  - 2byte
- 以下の数字を，2進数・16進数で表現してください
  - 3
  - 12
  - 140
- 次の数値を10進数で表現してください
  - 1011 (2進数)
  - 0x2F
- 次のbit列を16進数で表現してください
  - 0011 1100 1111 1001

英語文字を符号化してみよう

- アルファベット文字集合
  - 大文字・小文字 52文字
  - 数字 10 文字
  - 記号類
- 100 文字くらい
  - 7bit (2^7=128) あれば表現できる

ISO/IEC 646 (ASCIIを国際標準化したもの)

図1.4　ISO/IEC 646国際基準版の文字コード表（文字表）

上位3ビット➡ / 下位4ビット⬇				b7	0	0	0	0	1	1	1	1
				b6	0	0	1	1	0	0	1	1
				b5	0	1	0	1	0	1	0	1
b4	b3	b2	b1		0	1	2	3	4	5	6	7
0	0	0	0	0			SP	0	@	P	`	p
0	0	0	1	1			!	1	A	Q	a	q
0	0	1	0	2			"	2	B	R	b	r
0	0	1	1	3			#	3	C	S	c	s
0	1	0	0	4			$	4	D	T	d	t
0	1	0	1	5			%	5	E	U	e	u
0	1	1	0	6			&	6	F	V	f	v
0	1	1	1	7			'	7	G	W	g	w
1	0	0	0	8			(	8	H	X	h	x
1	0	0	1	9			)	9	I	Y	i	y
1	0	1	0	10			*	:	J	Z	j	z
1	0	1	1	11			+	;	K	[	k	{
1	1	0	0	12			,	<	L	\	l	|
1	1	0	1	13			-	=	M	]	m	}
1	1	1	0	14			.	>	N	^	n	~
1	1	1	1	15			/	?	O	_	o	DEL

※ISO/IEC 646:1991に基づいて筆者作成。

Z
1011010
0x710

+0x20

z
1111010

コード値の大小比較によって文字の自然な順番でのソートが可能

重要な用語の定義

- 文字集合
  - 文字を重複なく集めたもの
- 符号化文字集合
  - 概念上の文字の集合から非負整数の集合への写像
  - ASCII, JIS X 0208, UNICODE...
- 文字符号化方式
  - 符号化文字集合で文字に対応付けた非負整数値を、実際にコンピュータが利用できるデータ列（通常、バイト列）に変換する符号化方式
  - ShiftJIS, EUC-JP, UTF-8, UTF-16...

文字の符号化の原則

- 一意な符号化
  - 文字と符号は1体1対応させる
- ただし，似た形であって区別の難しい文字が同一表に存在することはあり得る
  - I (iの大文字)
  - l (Lの小文字)
  - | (縦棒)
- 包摂規準
  - 字体や字形デザインが異なっていても同じ文字コード番号で表す（=同じ文字であると見なす）規準
- 文字コードによって一意に符号化されるものは，
  - 使われる文脈や社会的な常識も加味した総合的な判断によって識別された文字
  - 単なるインクの染みではない
  - 大規模な文字集合を考えるとき，重要になってくる

制御文字

- 目には見えない特殊な文字
- タブ
  - 0x09
- 改行
  - 行を送り、文字位置を行の先頭に戻す
  - 復帰(CR, 0x0D)と改行(LF , 0X0D)
  - 表現方法は3通り. 各システムでよく使われるものは異なる
    - LF: UNIX系のシステム
    - CR+LF: Windows
    - CR : Mac

文字コードが複雑になる理由

- 過去の経緯の積み重ね
  - 文字データ
  - 過去の資産(プログラムやデータ)を無視することができない
  - 過去の資産との互換性や連続性が重要
  - 欠点があっても簡単に治せない
- 文字そのものの難しさ
  - インド系の文字
    - 構成要素を合成して表現する必要がある
  - アラビア文字
    - 1つの文字が語中の文脈に応じて形を変える
    - 右から左に向かって文字を並べるといった書字方向の制御を行う必要がある

まとめ

- 文字をコンピュターで扱うために, 符号化する
- データは, 内部では2進数で表現
  - Nbitで2^N種類の情報が表現できる
  - 簡便のため16進表記する
  - 8bit = 1byte
- 文字集合, 符号化文字集合, 文字符号化方式