情報基礎

情報の符号化 (2)
文字コードとその周辺

Modified by Harumi Murakami
Originally written by Kota Abe

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
  - 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?

文字のディジタル化の例

JIS X 0201コード表に基づいて、1文字ずつ文字コードに置き換える。

This is a pen.

文字コード
(10進数表記)

84	104	105	83	32	105	83	32	97	32	112	101	110	46

文字列データ

84,104,105,83,32,105,83,32,97,32,112,101,110,46

2進数の文字列データ

1010100,1101000,1101001,1010011,0100000,1101001,1010011,
0100000,1100001,0100000,1110000,1100101,1101110,0101110

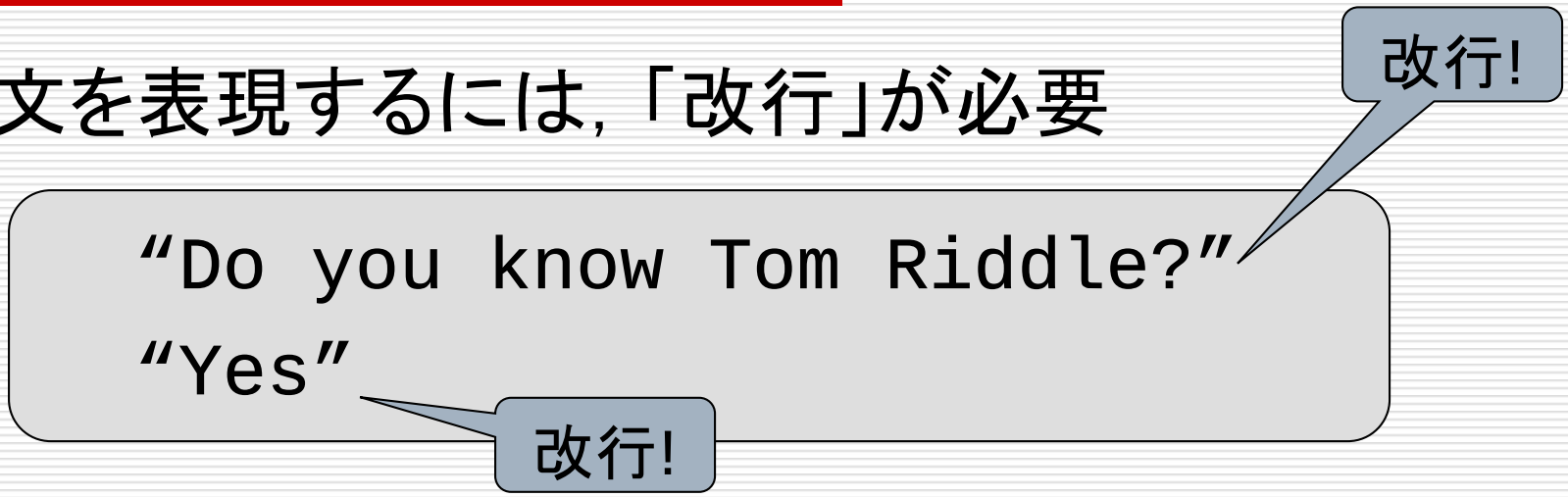
パソコンに記憶

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



- 「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

主な文字コード

- 最初はASCII：英数字（米国）
- 英語以外はどうする？→多言語を扱うために様々な文字コードができた
  - 日本の場合
  - ISO-2022-JP：電子メール
  - Shift-JIS：Windows
  - EUC-JP：UNIX
- 大変不便！→多言語対応のUnicodeへ

主な文字コード

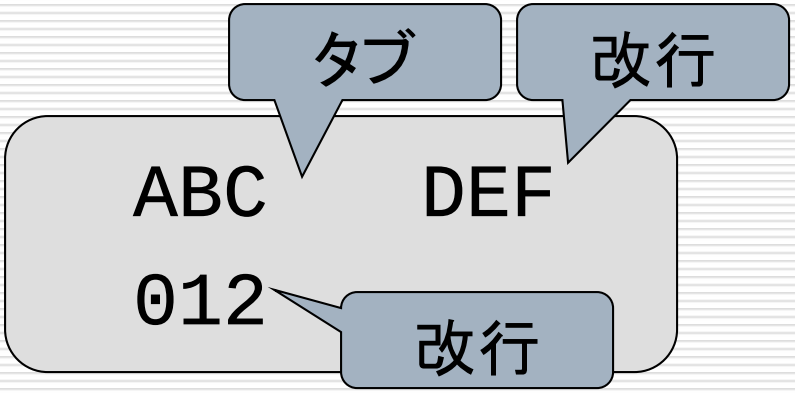
- ASCII
  - 英数字
- ISO-2022-JP
  - 電子メール
- Shift-JIS
  - Windows
- EUC-JP
  - Unix
- Unicode (UTF-8, UTF-16)
  - 近年、多言語対応

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

演習

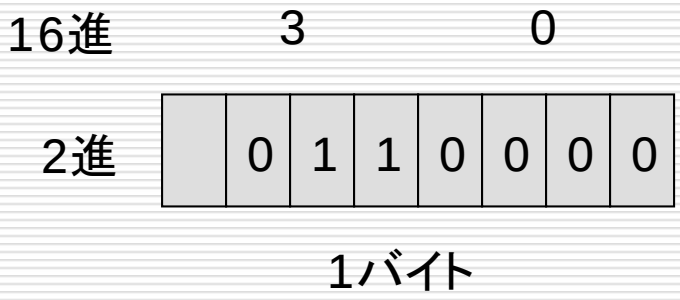
- エディタで次のようなファイルを作成し適当なファイル名で保存



- バイナリエディタ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
  - 内容を確認（2進数と16進数でどうなっているのか確認）

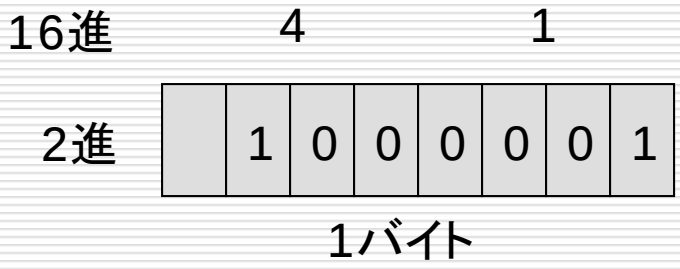
ASCIIの場合

- 「0」



0

- 「A」



A

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)

改行コードの違い

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

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

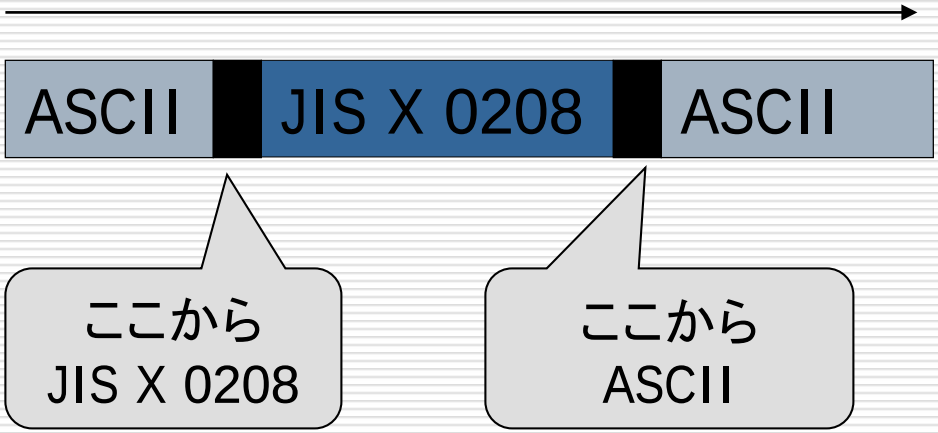
JIS漢字コード(JIS X 0208) (2)

- 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)

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
  - エスケープシーケンスが必要
- Shift_JIS, EUC-JP
  - エスケープシーケンスが不要
  - Shift_JIS：Windows, EUC-JP：UNIXで使われる

エスケープシーケンス

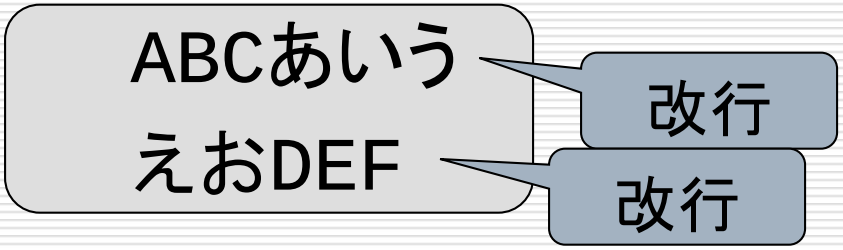


ASCII	ESC (B
JIS X 0201の0x20～0x7e	ESC (J
JIS X 0208(78年版)	ESC $ @
JIS X 0208(83年版)	ESC $ B

0x1b

演習（オプション）

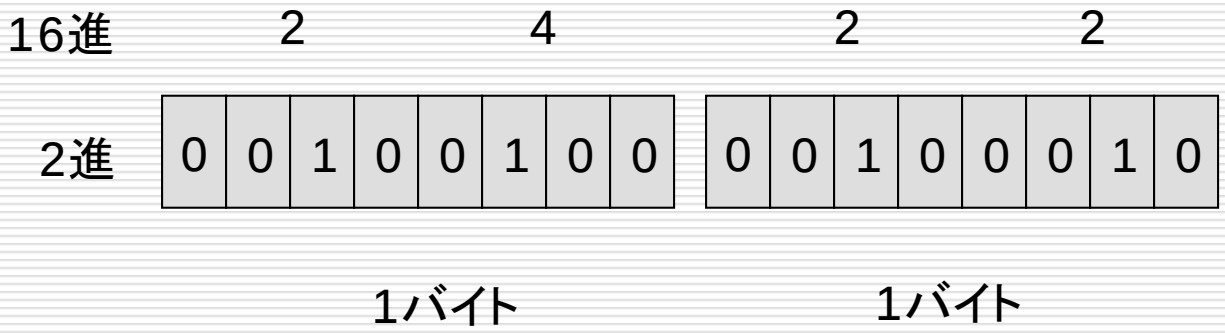
- JISで保存できるエディタがある人向け
- エディタで，次のようなテキストを入力せよ

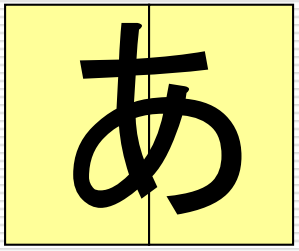


- セーブするときにエンコーディングとして JIS (ISO-2022-JP) を選択してセーブせよ
- バイナリエディタでエスケープシーケンスがどうなっているか確認せよ
- JISで保存できるエディタがない人はとばして次へ進んでください

ISO-2022-JPの場合

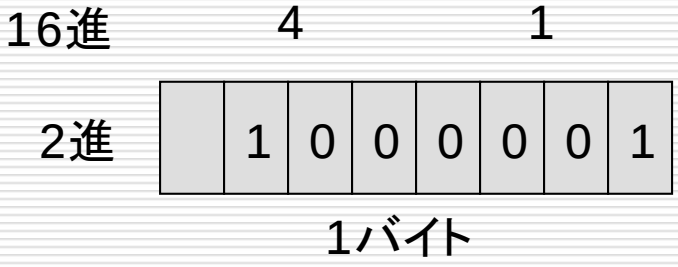
- 「あ」





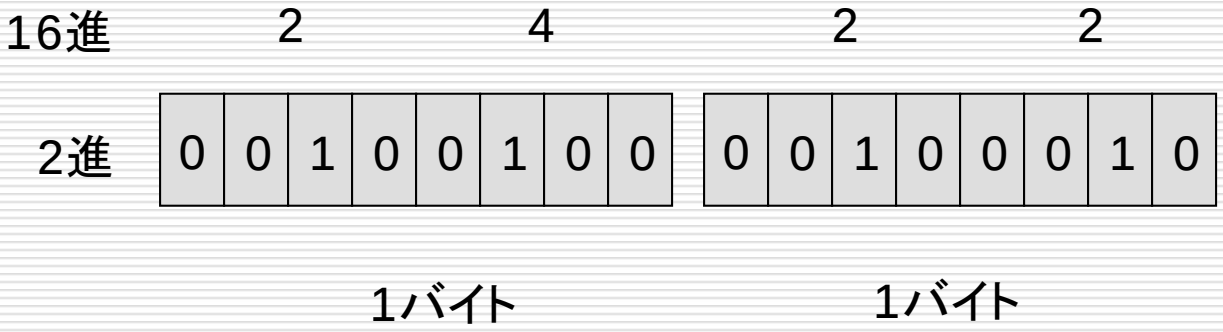
ASCIIとISO-2022-JPの比較

ASCIIの「A」



A

ISO-2022-JPの「あ」



あ

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

機種依存文字とユーザ定義文字

- 機種依存文字
  - コード表の空き領域にメーカーが独自に文字を定義
  - Windowsの丸付き数字 ①②③④ など（JIS2000で標準に）
  - 携帯電話の絵文字（Unicodeで規格に）
- ユーザ定義文字
  - コード表の外字領域にユーザが定義
  - Windowsの外字エディタ
- 情報交換の障害となる
  - 電子メールやWebページで使うとトラブルに

Unicode

- 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

- 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, Oracle, etc.
  - Unicode Consortium http://www.unicode.org/
- かなり普及している
  - Windows, MacOS X は内部Unicode
  - 多くのエディタやWebブラウザ，プログラミング言語でもUnicodeをサポート

Unicode

- 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」(日本語)と「高低」(中国語)は同じコード

演習

- エディタで，次のようなテキストを入力せよ

ABCあいう

- UTF-8を選択して保存
- バイナリエディタで内容を確認

Unicode

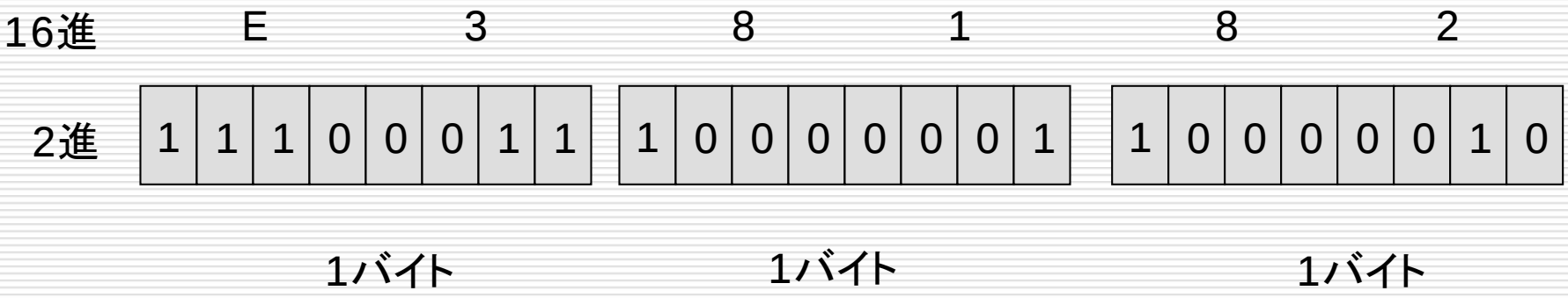
- Windowsでは「文字コード表」
  - 適当なフォントを選ぶこと
    - 日本語：MSゴシック
      中国語：SimSun
      韓国語：Gulim
- Macでは日本語入力の「文字パレット」等

Unicode

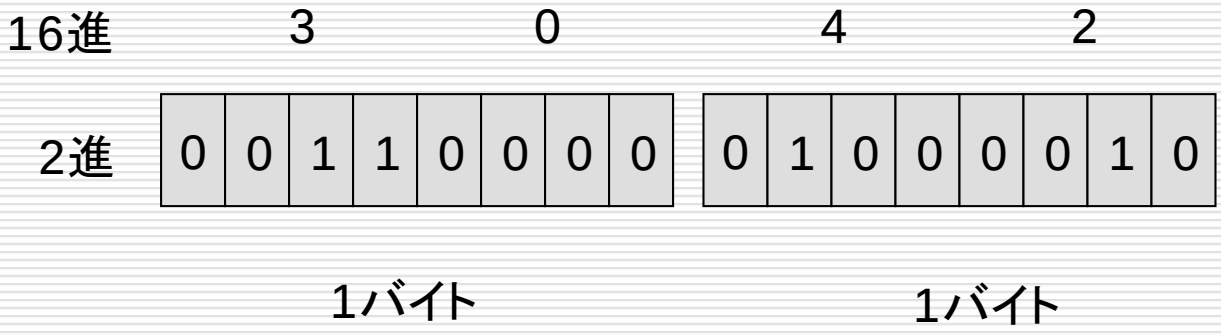
- Unicodeで使われるエンコーディング
  - UTF-8
    - 1～6バイトの可変長でエンコード
    - ASCII文字は1バイトで済む
  - UTF-16
    - (基本的に)2バイト固定長でエンコード

UTF-8とUTF-16

UTF-8の「あ」



UTF-16の「あ」



演習

- 自分の名前をUTF-8でエンコーディングしてみよう
- いろいろなファイルをバイナリエディタで見てみよう