

#2 bit, Byte, フォーマット

2004, Spring term, Yutaka Yasuda

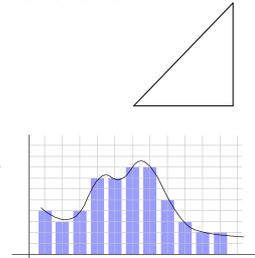
## デジタル処理の原理

### 情報のデジタル化=符号化=数値化

- 三角形なら  
(0,0),(100,0),(100,210)

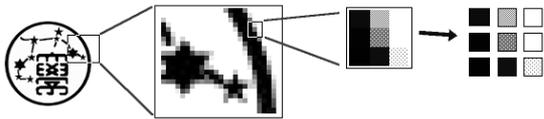
- 音声なら  
4,3,4,7,7,8,8,5,3,2,2...

- 確定的な数値として表現  
- 欠点と利点の双方をもつ



## 画像のデジタル表現

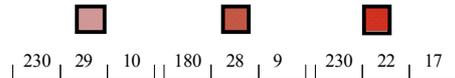
- 絵は画素(Pixel : Picture Element)ごとに分解
- 各画素ごとに数値化
  - 1-3-5,1-2-5,1-1-4 のごとし
  - 空間的なサンプリングと考える



## カラー画像のデジタル表現



一画素ごとに赤・青・黄  
に色分解して各色256段階  
で記録



空間的なサンプリングと考  
えれば良い

動画も簡単にデータ化で  
きますね？

## 文字のデジタル表現

- 数値化された文字、とは？
  - あり得る文字にすべて番号を振る
  - 番号付け=コード化 (符号化)
- 元もと文字はデジタルな存在？
- (例)
  - ABC = 1,2,3 とすれば 26 で足りる
  - abc = 27,28,29.. で 52 まで
  - 0,1,2 = 53,54 で 62 まで
  - 漢字はたいへんだが 6 万もあれば？

## 文字データの例

AB123 → “A” “B” “1” “2” “3”  
65 66 49 50 51 (ASCII)

漢字 → “漢” “字”  
180 194 187 250 (EUC)

- 標準枠 = Byte
- 1 Byte = 8 bit が主流 (歴史的経緯)

## 標準枠としての Byte

- Byte
  - 慣例的に決まった 0-255までの256種類の値を入れられる枠 (8bit)
  - 255を超える値は二桁 (2Bytes) 使う
  - Byte is not 'Bite', bit is not bit
- ASCIIは 1 バイト
- 漢字は (普通は) 2 バイト
  - 「フロッピー1枚は新聞何枚に相当し、」

## データの内部処理について

## 音楽CDは何バイトあるか？

- さまざまなもののバイト数
  - 広辞苑 (第二版)
    - 24字 x 50行 x 4段 x 2400ページ=11,520,000 字
    - 一文字 2 Bytesとして 23 Mega Bytes
  - 音楽CD
    - 44KHz x 65536段階(2Bytes) x 2ch = 176KB/sec
    - 176KB x 3600sec = 633,600 KB = 634MB
- さまざまなものが bit にかわる姿を想像できたろうか？

## 電子回路でのデータ表現

- コンピュータはスイッチの on/off で動作している
  - スイッチ一つ分、電線一本分、二進数一桁分のデータの枠が bit (binary digit) となる
- 理由
  - 二値動作の単純さ (誤動作への耐性)
  - 汎用性の高さ

## データは bit から電気信号へ

- 構造
  - デジタルデータは数値列として表現
  - データは標準枠 Byte 単位に格納される
  - 1 Byte = 8 bits & 1 bit = on/off
- 結論
  - データは二値化されて電子回路に格納される
  - すなわち内部は二進法で数値処理が行われる
- デジタルデータが電気的 on/off によって処理される姿が想像できたろうか？

## 話を論理的な層に戻します

## データ

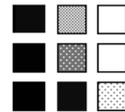
- その実体は数値（記号）の列
  - 音声：111,121,122,89,80,82,75....
  - 静止画：10,240,22,30,34,80...
  - 音声付き動画：12,33,45,1123,488...
  - 文字：33,38,42,60,32,39,55,80...
- これだけでは無意味
  - 符号化ルールとデータは常に一体
- このルールがフォーマット（書式）を生む

## フォーマット（書式）

同じ画像データでも数え方を違えると全く違う数字列に

- 1-3-5,1-2-5,1-1-4
- 1-1-1,3-2-1,5-5,4

符号化ルールと一致する復号化をしないと異なる結果に



## フォーマット（書式）

- デジタルデータを解釈するには
  - 解釈（復号）ルールが必要
- フォーマット（書式）
  - つまりデータにはフォーマットがある
  - フォーマットを間違えて解釈すると間違った結果が導き出される
  - 異なるアプリケーションでデータが扱えない理由
  - データにおける「互換性」という概念の実体

## 文字におけるフォーマットの問題

- 統一されていないルール
  - 文字番号表（この字を何番とするか）はいくつかある
  - バイト単位での並べ方（次の1バイトは漢字の前半か、後半か）にも幾つか
- ASCII コード
- JIS漢字表
  - JIS/EUC/Shift-JIS漢字コード
- いわゆる文字化けの原因

## まとめ

- デジタルデータとフォーマットの関係
  - デジタルデータはメディアからは自由になったがフォーマットが重要になった
- データはどのように内部処理されるか
  - 標準枠としての Byte
  - 二値電子回路のための bit
  - 表層から内部までの連続性を感じることを大切に
- 互換性の問題
  - フォーマットこそ互換性が存在する理由である

## 事例紹介

- Microsoft の HD DVD への進出
  - 動画フォーマットとしての Windows Media Series 9 の提出と(仮)承認が意味するものは何か？
  - NEC / 東芝は MPEG など公開の場で作られたフォーマットを推している
  - 何故か？