文字コード略歴

よこやままさふみ

社内勉強会
2012/05/18

自己紹介

- 横山昌史
- 入社4年目
- プログラマ etc...
- 所属プロジェクト
  - Java、UNIX、雑用 etc...
  - 文字コードの "るつぼ"

Rabbitについて

- プレゼンテーションツール
- 実装: Ruby/GTK
- 動作: UNIX/Win/Mac
- 文章とデザインの分離
  - バージョン管理しやすい

文字コードとは

- 文字をコンピュータで扱うための符号化方式
- エンコード、キャラクターセットとも呼ばれる

符号化

- 文字や音声などを0と1のデジタルデータに変換すること
- 16進数で記述されることが多い
  - Javaなどでは、頭に"0x"を付けると16進数として扱われる

よく使われる文字コード

- ✓ ASCII
- ✓ Shift_JIS
- ✓ UTF-8（Unicode）
- ✓ EBCDIC

ASCII

ASCII

American
Standard
Code for
Information
Interchange

ASCII

- 英語を表現するための文字コード
  - 英字アルファベット、記号
  - いわゆる半角文字

ASCII

- 1文字につき7ビットの1バイトコード
  - 7ビット = 2^{7} = 128
  - 16進数で言うと0x00～0x7Fまで
  - 1バイト = 8ビットのため、1ビット余り

ASCIIの例

- Heisei 24
  - 48 65 69 73 65 69 20 32 34
    - 16進数（0xは省略）

JIS X 0201

- 日本工業規格
- 一部の符号位置がASCIIと違う
- 半角カナが使える
  - 8ビット目を拡張

ASCIIと異なる文字

- 符号位置0x5C
  - ASCII:半角バックスラッシュ
  - JISX0201:半角円記号

ASCIIと異なる文字

- 符号位置0x7E
  - ASCII:半角チルダ
  - JISX0201:半角オーバーライン

ASCIIと異なる文字

- ASCIIかJISX0201かは曖昧
  - 環境によって表示が変わる
- 日本のフォントは円記号
- 外国のフォントはバックスラッシュ
- 7Eは日本でも大抵チルダ

Shift_JIS

Q & A

- Q.あなたの母語は何語ですか？
  - A.日本語
- Q.日本語は英字アルファベットだけで表現できますか？
  - A.いいえ

ひらがなや漢字が必要

- 常用漢字 2,136文字（2010年改定）
- 1バイト = 8ビット = 2^8 = 256
- 1バイトでは表現できない

2バイトコード

- 1文字を2バイトで符号化
- 2バイト＝16ビット＝2^{16}＝65,536

JIS X 0208

- 日本工業規格
- JIS第1・第2水準漢字を定義
- 最新版では6,879文字を収録
- 1983年に大幅な変更
  - 異字体の符号位置入れ替え
  - 字形の変更

JIS X 0213

- ✓ 日本工業規格
- ✓ JIS X 0208の拡張（後方互換）
- ✓ JIS第3・第4水準漢字を定義
- ✓ 環境によってはJISX0213に対応していない（JIS第3・第4水準漢字が使えない）

Shift_JISの成り立ち

JIS X 0201
＋
JIS X 0208（JIS X 0213）

Shift_JISの特徴

- ✓ 日本語が表現できる
- ✓ 半角カナが使える
  - ✓ JISX0201との互換性

Shift_JISの例

- 平成 24
  - 95 BD 90 AC 20 32 34

Shift_JISの欠点

- 全角半角問題
  - 「A」と「A」、「ｱ」と「ア」など
  - 全角文字を扱える文字コード共通の問題
- JISX0201の副作用
  - だめ文字

Shift_JISの派生

- WindowsではShift_JISを拡張した文字コードが使われている
  - Windows31-JやMS932やCP932などと呼ばれる
  - 重複符号化(㈱問題)
    - 同じ文字に複数の符号位置

UTF-8

Q & A

- Q.今はどんな時代ですか？
  - A.国際化時代
- Q.Shift_JISの欠点はどこですか？
  - A.日本語しか扱えない

Unicode

- 世界中の言語を表現できる文字コードの仕様がUnicode
- 110,181文字（2012年1月）

Unicode

- Unicodeの実装の一つがUTF-8
  - 他にもUTF-16など

UTF-8の特徴

- ASCIIを拡張
  - Shift_JISとは違い、JISX0201の拡張ではない
  - 半角カナなどの符号位置がShift_JISと違う
  - だめ文字がない

1文字のバイト数

- ASCII
  - 全て半角文字 = 全て1バイト
- Shift_JIS
  - 半角 = 1バイト
  - 全角 = 2バイト

1文字のバイト数

- UTF-8
  - 半角 = 主に1バイト
  - 全角 = 日本語は3バイト
    - 記号は3バイトか2バイト

UTF-8の例

- 平成 24
  - E5 89 B3 E6 88 90 20 32 34

1バイトでない半角文字

- ￥(半角)が2種類
  - 5C (ASCII): 規格上はバックスラッシュ
  - C2 A5 (UTF-8): 規格上は円記号
    - 2バイトの半角文字

1バイトでない半角文字

- ~（半角）も2種類
  - 7E (ASCII): 規格上はチルダ
  - E2 80 BE (UTF-8): 規格上はオーバーライン
    - 3バイトの半角文字

半角カナ

- ｱ（半角）
  - B1 (Shift_JIS)
  - EF BD B1 (UTF-8)
    - UTF-8の半角カナは全て3バイト
- 単純なバイト数チェックでは、半角か全角か判別できない

IBM版とMS版

- IBM-Unicode（一般的なUnicode）とMS-Unicode（マイクロソフト版Unicode）で符号位置が異なる文字がある
  - いわゆる波ダッシュ問題の要因

波ダッシュ問題

- Windowsとそれ以外のOS間での通信時などに文字化け
- 対象文字は10文字程度（環境による）
  - ～ ― − ‖ ∥ ¦ ¬ ¢ £

EBCDIC

EBCDIC

- IBMによって定義された文字コード
  - IBM製のメインフレーム（汎用機）などで現在も使用されている

EBCDIC

- ✓ 半角文字の符号位置がASCIIと異なる
- ✓ 全角文字の表現方法がShift_JISやUTF-8と異なる
- ✓ 基本的にJIS第3・第4水準は含まれない

半角文字

- すべて1バイト
  - 8ビット目まで使用
- 制御文字エリアが大きい
  - 0x00～0x3Fと0xFF
  - 汎用機で使用される特殊な制御文字が含まれている

全角文字

- 半角文字との区別は制御文字で行う
  - 全角の開始位置がシフトアウト（0x0E）
  - 全角の終了位置がシフトイン（0x0F）
- 略してSO/SIなどと呼ばれる

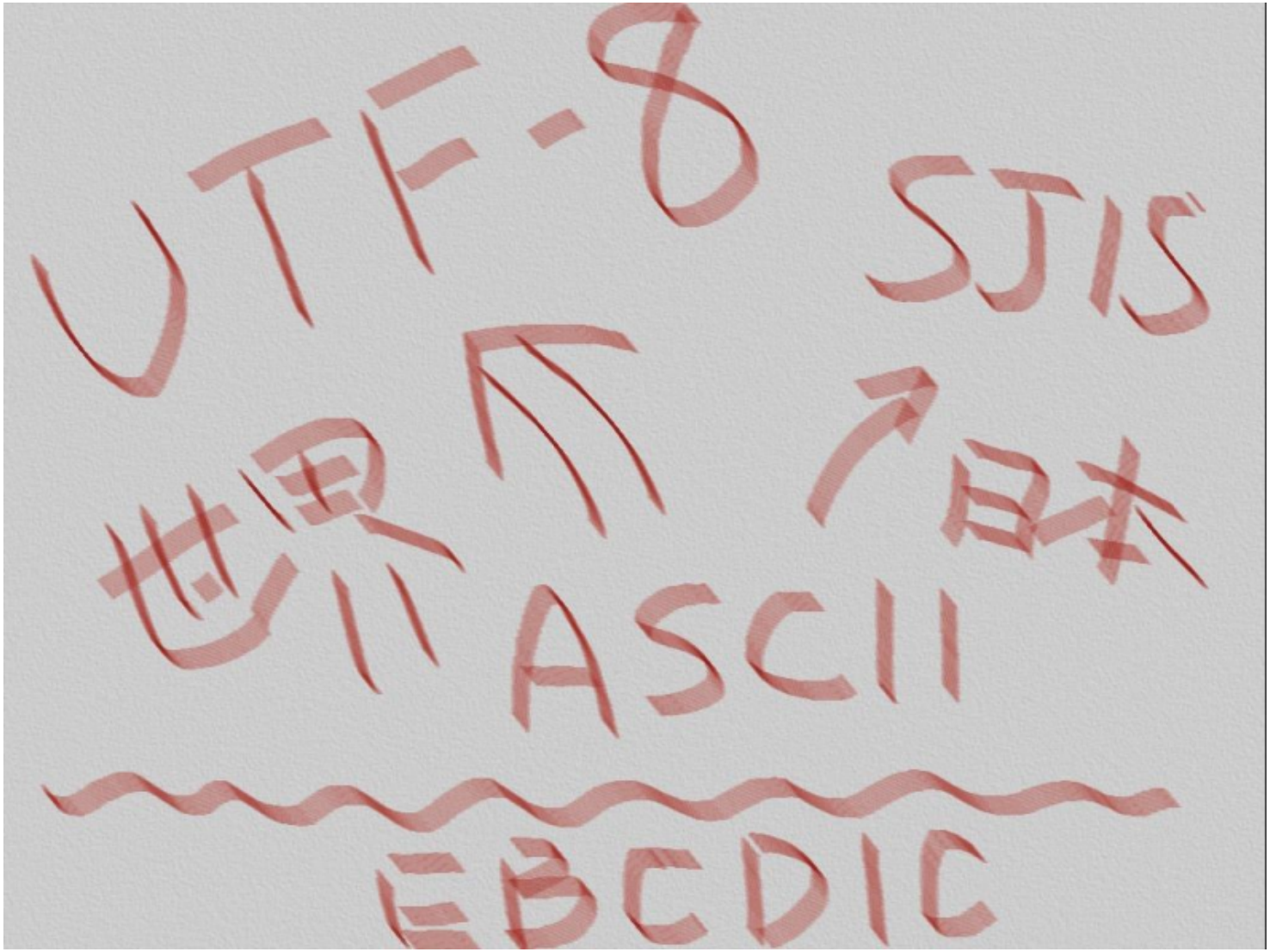
SO/SI

- 平成 24
  - 0E 45 8D 45 BA 0F 40 F2 F4

うわっ…

- SIの欠如
  - 0E 45 8D 45 BA
- SO/SIのネスト
  - 0E 45 8D 0E 45 BA 0F 0F
- SO/SIを対として扱うのではなく、モード切替文字として扱うことで対応

まとめ



参考

- プログラマのための文字コード技術入門
- 正規表現クックブック（66ページ）
- AIX 5L 日本語コード一覧表
  - jp_codebookで検索

ご静聴ありがとうございました。