情報処理システム論 (14)

コード表と符号化方式

- コード表
  - 文字に番号を割りたてたもの
  - JIS X 0208
- 符号化方式
  - 番号のビット列への変換方法
  - ISO-2022-JP
  - EUC-Japan
  - Shift JIS

符号化方式の識別

- 世の中には様々な符号化方式がある
- 自動識別が可能なものもあるが一般的な自動識別方法はない
  - EUC-Japan と Shift JIS の重複
  - EUC-xxxxx の識別
- 符号化方式を識別するための情報を提示する方法が必要

MIME (Multipurpose Internet Mail Extensions)

- メールの本文の識別方式
- メールのヘッダの符号化方式および識別方式

- メールに限らずニュースやWWWでも用いられている
- （MIME に従ってさえいればなんでもよいというわけではない）

メールの本文の識別方式

MIME-Version: 1.0
Content-Type: text/plain; charset=ISO-2022-JP

ヘッダ

これは日本語の文章です。

本文

MIME-Version: 1.0
Content-Type: text/plain; charset=US-ASCII

ヘッダ

This is a English text.

本文

Charset

- ISO-2022-JP
- EUC-JP
- ISO-10646-J-1
- Windows-31J など
- IANA (Internet Assigned Numbers Authority) に登録
  - http://www.isi.edu/div7/iana/descript.html
  - x-**** は自由に使ってよい

WWWの場合

- Content-Type ヘッダを同様に利用する
  - HTTP で通知
  - HTML で通知
- HTTP/1.1
  - Content Negotiation
    - クライアントに対して複数の charset を返し選ばせることができる

HTTPで通知

GET /index.html HTTP/1.0

HTTP/1.0 200 OK
Date: Mon, 20 Oct 1997 01:39:35 GMT
Server: Apache/1.1.3 BSDI/3.0
Content-type: text/html; charset=ISO-2022-JP

ようこそ

HTMLで通知

- META の利用

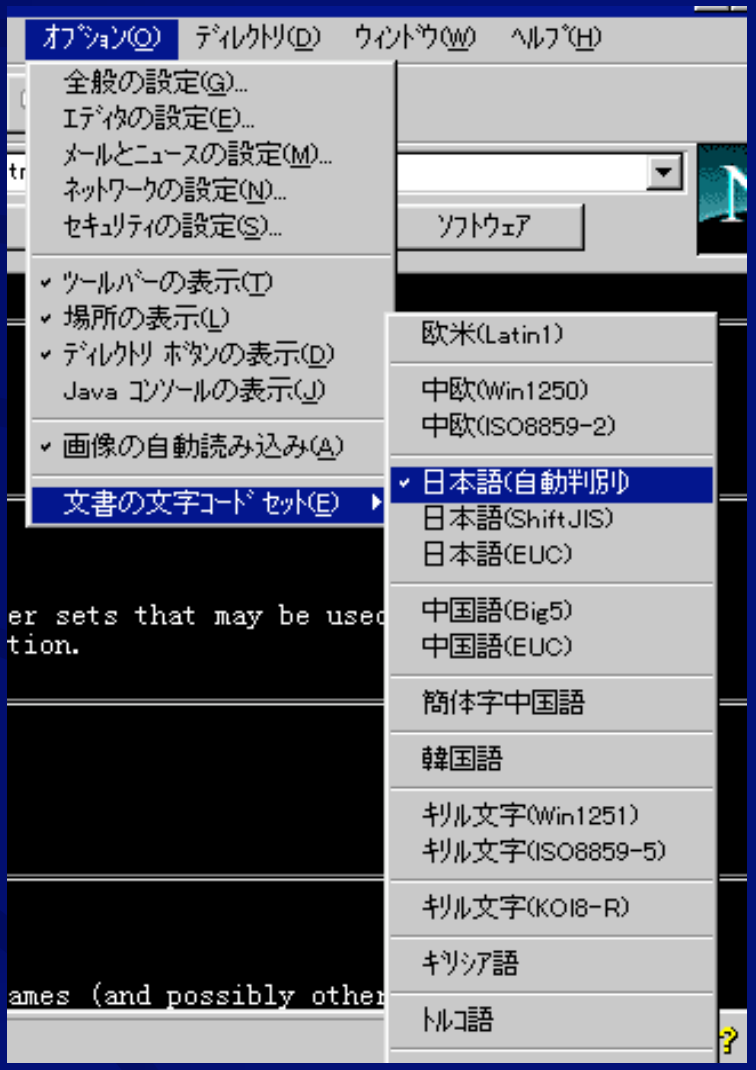
```
<META HTTP-EQUIV="Content-Type"
  CONTENT="text/html; charset=ISO-2022-JP">
```

問題点

- 実際にはあまり利用されていない
  - サーバに自動判別機能がない
  - ページの作成者が指定してくれない
    - 特に困っていない
  - 指定したくてもできない
- 間違って指定されていると読めなくなる
- 自動コード変換プロキシとの相性が悪い

クライアント側で識別

- ユーザが選択する
- 知識が必要



文字コードと内容

- UNICODEにするだけでいいのか？

- 文字コードは文字を図形として交換・保存する手段
- 内容が何であるかを判断するのは、それを参照する側の仕事である

JIS X 0208 で書けるけど...

- 「私の名前は中村です」は、中国語では「我的名前是中村」と書きます。
- 「湯」は、日本語では熱い水を意味し、中国語ではスープを意味します
- My Name is Nakamura.

すべてが日本語じゃないはず。

UNICODE でも

- すべての文章は UNICODE だけで表現できるだろう
- ある部分が、どの言語で書かれているかは知的に判断しないといけない？
- せっかく書く人はどの言語で書いているのかを知っているのだから、その情報を残したい → Language Tag

Language Tag

- Tags for the Identification of Languages
  - RFC1766
  - PrimaryTag*(-SubTag)
- PT: ISO-639 の言語コードを用いる
  - ja, en, fr, de,…
- ST: ISO-3166 の国コードを用いる
  - jp, us, fr, de,...

SGML形式で指定

- RFC1766 的方式
- HTML4 的方式

中国語では<LANG ZH>「我的名前是中村」</LANG>と書きます。

<P LANG="en">English</P>
<P LANG="fr">le francais</P>
<P LANG="ja">日本語</P>
<P LANG="zh">中国語</P>

MIME と Language Tag

Content-Type: multipart/alternative; differences=Content-Language;
boundary="limit"
Content-Language: en, fr

--limit
Content-Language: fr

Le renard brun et agile saute par dessus le chien paresseux
--limit
Content-Language: en

The quick brown fox jumps over the lazy dog
--limit--

余談:LOCALEと言語コード

- setenv LANG=C

  % date

  Mon Oct 20 11:30:58 JST 1997

- setenv LANG=ja_JP.ujis

  % date

  平成9年10月20日 (月), 午前 11時31分10秒

もう一つのUNICODEの弊害

- 異なる文字なのに unify されている
  - 骨
- 実はJIS自体にもある
  - 柿
    - かき
    - こけら

参考文献

- いま日本語が危ない
  ～文字コードの誤った国際化～
  太田昌孝, 丸山学芸図書
  ISBN 4-89542-146-5, \2000