雑に文字コードについて考えてみる

注意

雑に考えるので正確でない表現、情報があります。
まぁ、こんなもんかと流してもらえれば〜 😅
気になる方は調べてみてください。

文字コードとは？

コンピュータ上で文字（キャラクタ）を利用する目的で各文字に割り当てられるバイト表現のこと。

文字コード - Wikipedia

文字コードとは？

コンピュータ上で文字をどう表すか文字毎に決めたもの。

文字	バイト表現（2進数）
A	0100 0001
B	0100 0010
C	0100 0011
…	…

文字コードの類語

- コードセット (CodeSet)
- キャラクタセット (charset, charseter set)
- 文字マップ (Character Map)
- コードページ (CodePage)
- エンコーディング (encoding)

文字コードの有名どころ

- UTF-8
- UTF-16
- UTF-32
- ISO-2022-JP == JIS
- Shift_JIS !== CP932
- EUC-JP !== CP51932

PHP との関係

文字コード	PHP でサポートされているエンコーディング (CI)
UTF-8	UTF-8
ISO-2022-JP	JIS (ISO-2022-JP は半角カナが化ける)
Shift_JIS	SJIS
EUC-JP	EUC-JP
CP932	SJIS-win
CP51932	eucJP-win

- 丸数字等化けるので変換時は `*-win` を使ってください。

どの文字をコンピュータ上で扱うか
決めないといけない。

文字集合

コンピュータ上で扱う文字を集めたもの。

符号化文字集合

文字集合を定義し、その集合内の各文字に一意の符号化表現を関連付けたもの。

文字集合の類語

- 文字セット (charset, charseter set)
- 符号点 (code point) 符号化文字集合内の、文字を割り当てうる個々の点

charset に限らず

文字コードの話か、文字集合の話か確認する必要があることもあります。

文字集合の有名どころ

- Unicode
- JIS X 0208
- JIS X 0212 （補助漢字など）
- JIS X 0213 （JIS X 0208 に第三・第四水準漢字などを追加）

文字集合と文字コードの関係

文字集合	文字コード
Unicode	UTF-8, UTF-16, UTF-32
JIS X 0208	ISO-2022-JP, Shift_JIS, EUC-JP, CP932, CP51932
JIS X 0212	EUC-JP, ISO-2022-JP-2, ISO-2022-JP-1
JIS X 0213	ISO-2022-JP-2004, Shift_JIS-2004, EUC-JIS-2004

まとめ

文字集合（符号化文字集合）

コンピュータ上で扱う文字を集めたもの。
（かつ、各文字に一意の符号化表現を関連付けたもの）

文字コード

上記で決めた文字集合内の文字をコンピュータ上でどう表すか文字毎に決めたもの。

おまけ

改行コード

改行コードとは、コンピュータなどで、改行を表す制御文字のこと。
[改行コード - Wikipedia]

コード名	コード	16進数	使われてる環境
LF	\n	0x0A	Unix 系
CR + LF	\r\n	0x0D0x0A	Windows 系
CR	\r	0x0D	古い Mac (OSX 未満)

PSR-12

PSR-12: Extended Coding Style - PHP-FIG

“ All PHP files MUST use the Unix LF (linefeed) line ending only. ”

全ての PHP ファイルは、Unix LF（ラインフィード）の行末のみを使用する必要があります。

まとめ

UTF-8, LF で書く。

文字コード、文字集合は魔境

先人に感謝