

PHPのマルチバイト処理

— 分類と落とし穴

埜 与志夫 (hnw)

第59回PHP勉強会 2012/05/25 発表資料

問題意識


- PHPのマルチバイト処理は関数ごとに挙動が異なることがある
 - 実装の差が原因
 - 意外と知られていない
- 分類がトラブル解決の助けになるはず

今日のゴール

- PHPの文字コード処理を分類する
- ありがちなトラブル事例を共有

自己紹介

- hnw
- bugs.php.netでバグ報告18件



php.net | support | documentation | report a bug | advanced search | search howto | statistics | login

go to bug id or search bugs for

Showing 1-18 of 18

ID#	Date	Last Modified	Package	Type	Status	PHP Version	OS	Summary	Assigned
48254 (edit)	2009-05-13 03:47 UTC	2009-05-14 01:42 UTC	Arrays related	Bug	Closed	5.2.9	*	Inconsistent handling of huge numerical keys for array	mattwil
52826 (edit)	2010-09-13 09:32 UTC	2010-09-13 09:42 UTC	Bzip2 Related	Bug	Closed	5.3.3	any	phpinfo is incorrect for bzip2 stream wrapper	aharvey
53249 (edit)	2010-11-07 04:24 UTC	2010-11-07 06:21 UTC	Date/time related	Doc	Closed	5.3.3	-	No documentation about comparison between two Datetime objects	frozenfire
53370 (edit)	2010-11-21 13:38 UTC	Not modified	Date/time related	Bug	Open	5.3.3	Linux and MacOSX	Some relative date/time format returns incorrect result at the end of DST	
46930 (edit)	2008-12-22 23:09 UTC	2009-05-03 14:45 UTC	Documentation problem	Doc	Closed	5.3.0alpha3	any	5.3.0alpha3's strtotime() returns inconsistent result with some relative items	derick
46932 (edit)	2008-12-23 06:18 UTC	2009-09-15 06:52 UTC	Documentation problem	Doc	Closed	5.2.8	any	strtotime() has inconsistency between 'next Monday' and '+1 Monday'	
47370 (edit)	2009-02-12 16:22 UTC	2009-05-15 17:10 UTC	Documentation problem	Doc	Closed	5.2.9	*	array_unique has backward compatibility problem, and SORT_REGULAR is confusing	
46478 (edit)	2008-11-04 12:56 UTC	2009-12-22 05:50 UTC	Feature/Change Request	Req	Closed	5.2.6	*	htmlentities() uses obsolete mapping table for character entity references	moriyoshi
47745 (edit)	2009-03-21 23:34 UTC	2009-03-31 10:06 UTC	Filter related	Bug	Closed	5.2.9	*	FILTER_VALIDATE_INT doesn't allow minimum integer	dmitry
47752 (edit)	2009-03-23 05:40 UTC	2009-11-25 10:41 UTC	Filter related	Bug	Closed	5.2.9	*	FILTER_VALIDATE_INT doesn't allow "+0" and "-0"	pajoye

自己紹介

- hnw
 - bugs.php.netでバグ報告18件
 - 好きな物：バグ, カレー
 - 好きな境界値：2の53乗+1

マルチバイト処理とは

- マルチバイト文字1文字を認識する
 - 例：UTF-8として何文字か調べる
- 文字コードの変換
 - 例：UTF-8文字列をShift_JISに変換

マルチバイト処理の分類

- 大別すると4つ
 - シングルバイト処理
 - OSの持つ多言語対応の仕組みに依存
 - 外部ライブラリに依存
 - 自前実装

①1バイトずつ処理する関数

- 例：strlen, substr
- 要は文字コードを意識しないだけ

②ロケール依存の関数

- ロケール=OSの多言語対応の仕組み
- 例：fgetcsv、ucfirst
- OSの実装およびロケール設定に依存
- ロケール依存の関数一覧：
「PHPのロケールに関するまとめ」
(<http://d.hatena.ne.jp/hnw/20120501>)

③mbstring系関数

- 例：mb_convert_encoding
- libmbflがベース、実質PHP独自実装
(廣川さんやmoriyoshiさんが改造)
- 約60の文字コードに対応
- PHPマニュアル「サポートされる文字エンコーディング」に詳細あり

④mb_ereg系関数

- 例：mb_ereg_match
- 正規表現エンジン鬼車を利用
 - 約30の文字コードに対応
 - PHPマニュアル「サポートされる文字エンコーディング」に詳細あり

⑤preg系関数

- 例：preg_match
- 正規表現エンジンPCREを利用
 - ASCIIとUTF-8に対応
 - u修飾子をつけると"."がUTF-8の1文字にマッチするようになる

⑥iconv系関数

- 例：iconv、iconv_strlen
- 文字コード変換ライブラリiconvを利用
 - 多数の文字コードに対応
 - 実装が複数あり互換性に疑問

⑦GNU recode関数

- 例：recode_string
- 文字コード変換ライブラリらしい
 - 約150の文字コードに対応
- 利用実績は疑問

⑧XML系関数

- 例：SimpleXML
- XML処理ライブラリlibxml2を利用
 - 約40の文字コードに自前対応
 - Shift_JISなど日本語未対応
 - iconvが有効ならiconvに丸投げ

⑨htmlspecialchars系

- htmlspecialchars/htmlentitiesの第3引数に対応する処理
 - マルチバイト1文字を調べる処理
 - 14の文字コードに対応

分類まとめ

- PHPの文字コード処理は9種類！
 - ①無視 ②ロケール依存 ③mbstring
 - ④鬼車(mb_ereg) ⑤PCRE(preg)
 - ⑥iconv ⑦GNU recode ⑧libxml2
 - ⑨htmlspecialchars
- この分類で過去のバグの原因がわかった人はいませんか？僕にも教えて下さい！

実装の多さは問題か？(1)

- 文字コードの処理が複数あると潜在的なバグのリスク、良くはない
 - 例：htmlspecialcharsがShift_JISのXSSを最近まで防げていなかった件
 - PHPプログラマの混乱の元

実装の多さは問題か？(2)

- libxml2などは他言語でも依存している
 - 実装の混在自体は仕方がない
- 透過的に扱えるなら問題にならない
 - 関係者の努力で少しずつ前進している

トラブル事例の紹介(1)

- iconvとmb_convert_encodingとを同時に使ったら文字化け
 - 原因：両者の変換表が違う
 - 対策：どちらかに統一する

iconvのイヤなところ

- 波ダッシュ「~」の問題

`iconv("CP932", "UTF-8", $c)`



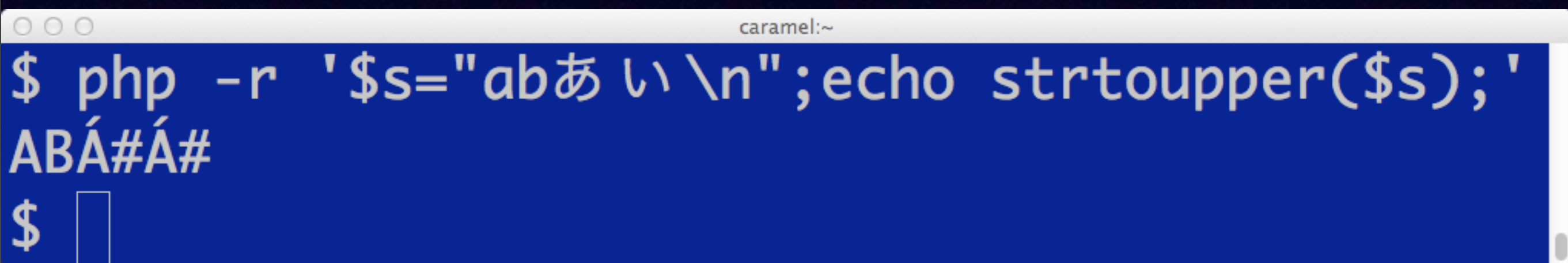
`mb_convert_encoding($c, "UTF-8", "SJIS-win")`



混ぜるな危険！

トラブル事例の紹介(2)

- strtoupperで日本語が化ける



```
caramel:~  
$ php -r '$s="abあい\n";echo strtoupper($s);'  
ABÁ#Á#  
$
```

- 原因：UTF-8ロケールが壊れてる
- 対策：setlocale(LC_ALL, "C")

ロケールは鬼門

- ロケールはOSの多言語対応の仕組み
 - アジア圏向けのテストがぬるい傾向
 - 壊れてることもあります！
 - Debian系だと入ってないことも

トラブル事例の紹介(3)

- fgetcsv関数がSJISのCSVを読めない
 - setlocale忘れか、
SJISロケールが壊れてるのでは
 - 対策：UTF-8に文字コード変換
(<http://d.hatena.ne.jp/hnw/20090317>)

まとめ

- PHPの文字コード処理を9種類に分類
- よく見るトラブルは次の2個？
 - mbstringとiconvの違い
 - ロケール周り
 - 他にあったら教えてください！

ご清聴
ありがとうございます
ございました