

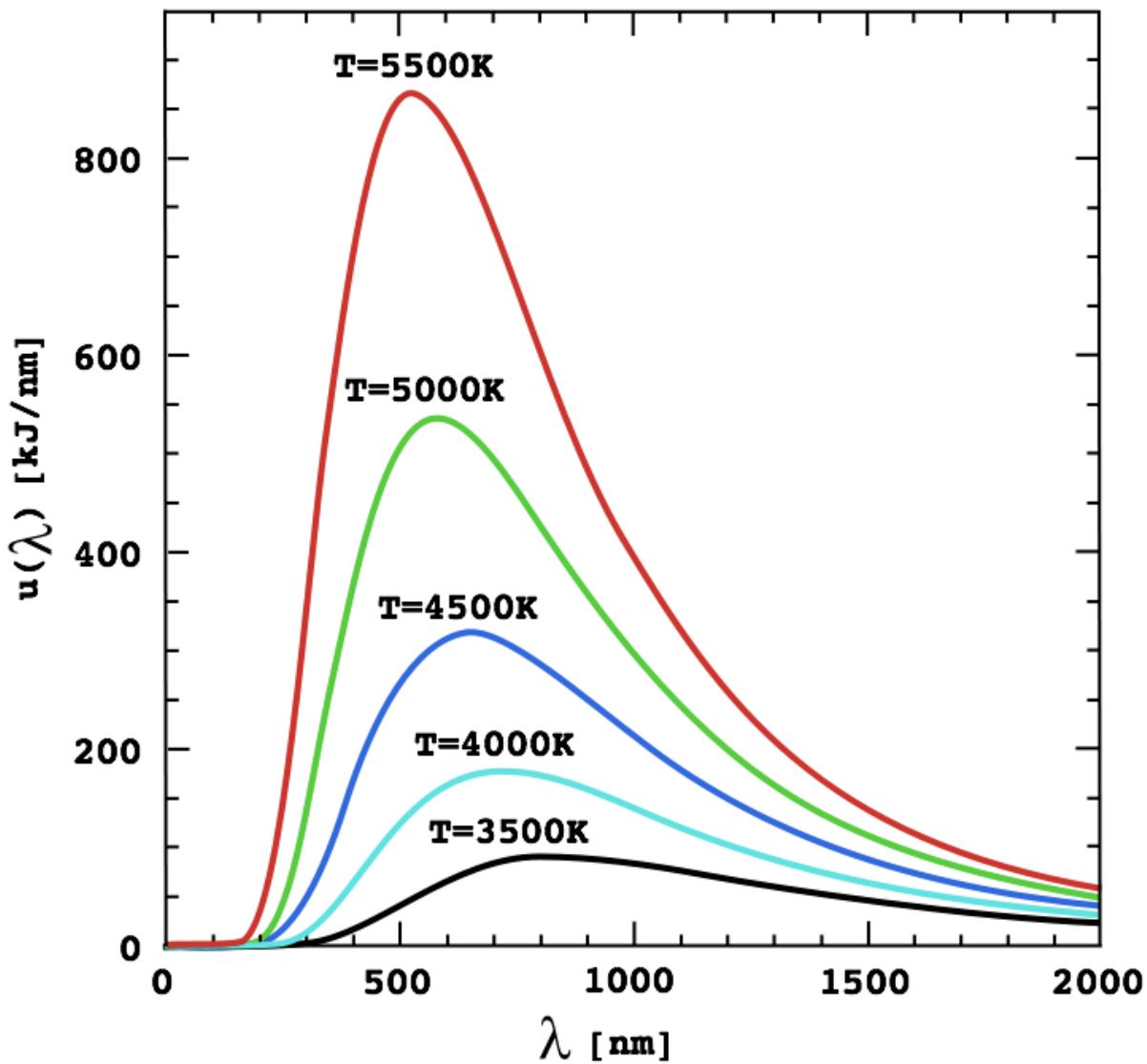


Andrei Zmievski
Chief Architect
Outspark, Inc

PHP 6

or

*“Im in ur endginn,
playin wif ur stringz!”*



$$I(\lambda, T) = \frac{2hc^3}{\lambda^5} \frac{1}{e^{\frac{hc}{\lambda kT}} - 1}$$





PHP 6 = PHP 5 + Unicode



PHP 5 = PHP 6 - Unicode



Unicode = PHP 6 - PHP 5



What is Unicode?

and why do I need?



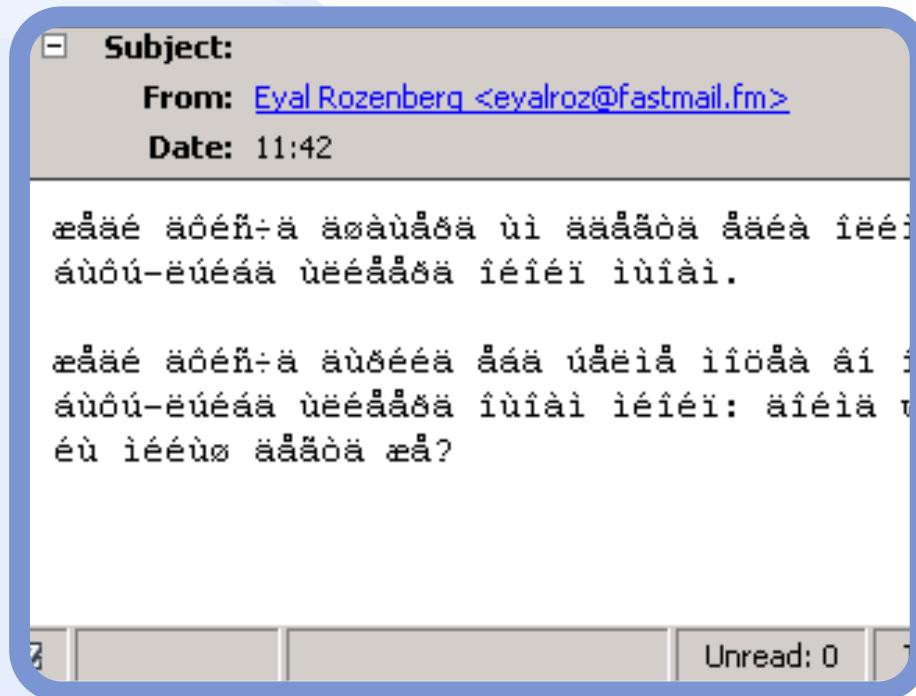
mojibake

もじばけ

mojibake

phenomenon of incorrect, unreadable characters shown when computer software fails to render a text correctly according to its associated character encoding

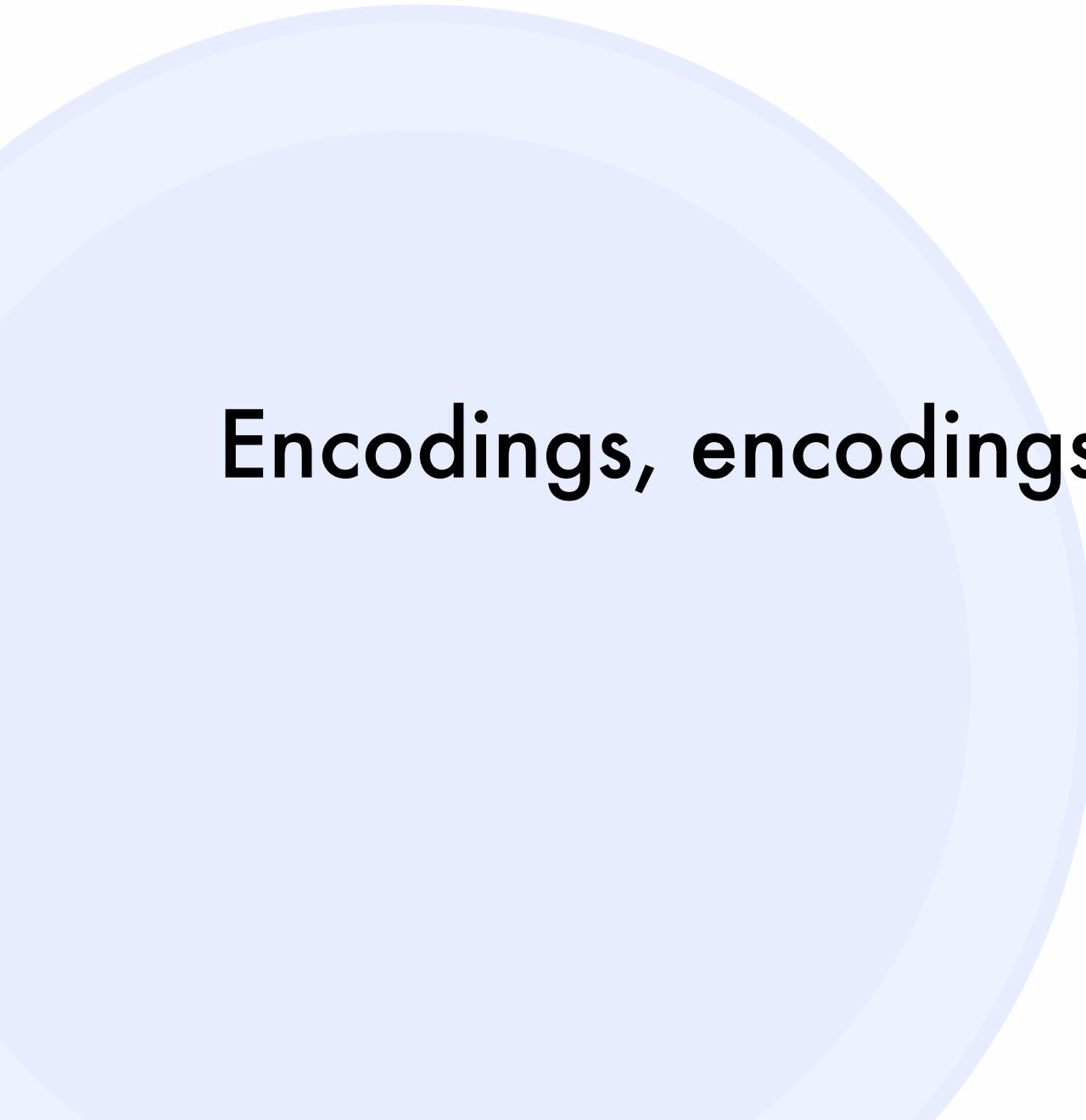




mojibake



Computers deal with numbers



Encodings, encodings, encodings

Encoding soup

1. provinciality
2. computer limitations
3. inertia

What is Unicode?

Latin

ASCII
ISO-8859-1
Windows-1252

Τι είναι το Γιούνικοντ;

Greek

ISO-8859-7
Windows-1253

Что такое Юникод?

Cyrillic

ISO-8859-5
Windows-1251
KOI8-R

Ç'është Unicode?

Albanian

ISO-8859-2
Windows-1250

מה זה יוניקוד?

Hebrew

ISO-8859-8
Windows-1255

ما هي الشفرة الموحدة "يونيكود"؟

Arabic

ISO-8859-6
Windows-1256

유니코드에 대해?

Korean

EUC-KR

ISO-2022-KR

MS Code Page 949

什么是统一码？

Simplified Chinese

GB18030

GB2312

GBK

EUC-CN

ISO-2022-CN

யூனிக்கோடு என்றால் என்ன?

Tamil

ISCII
TSCII

Unicode

provides a unique number
for every character:

no matter what the platform,

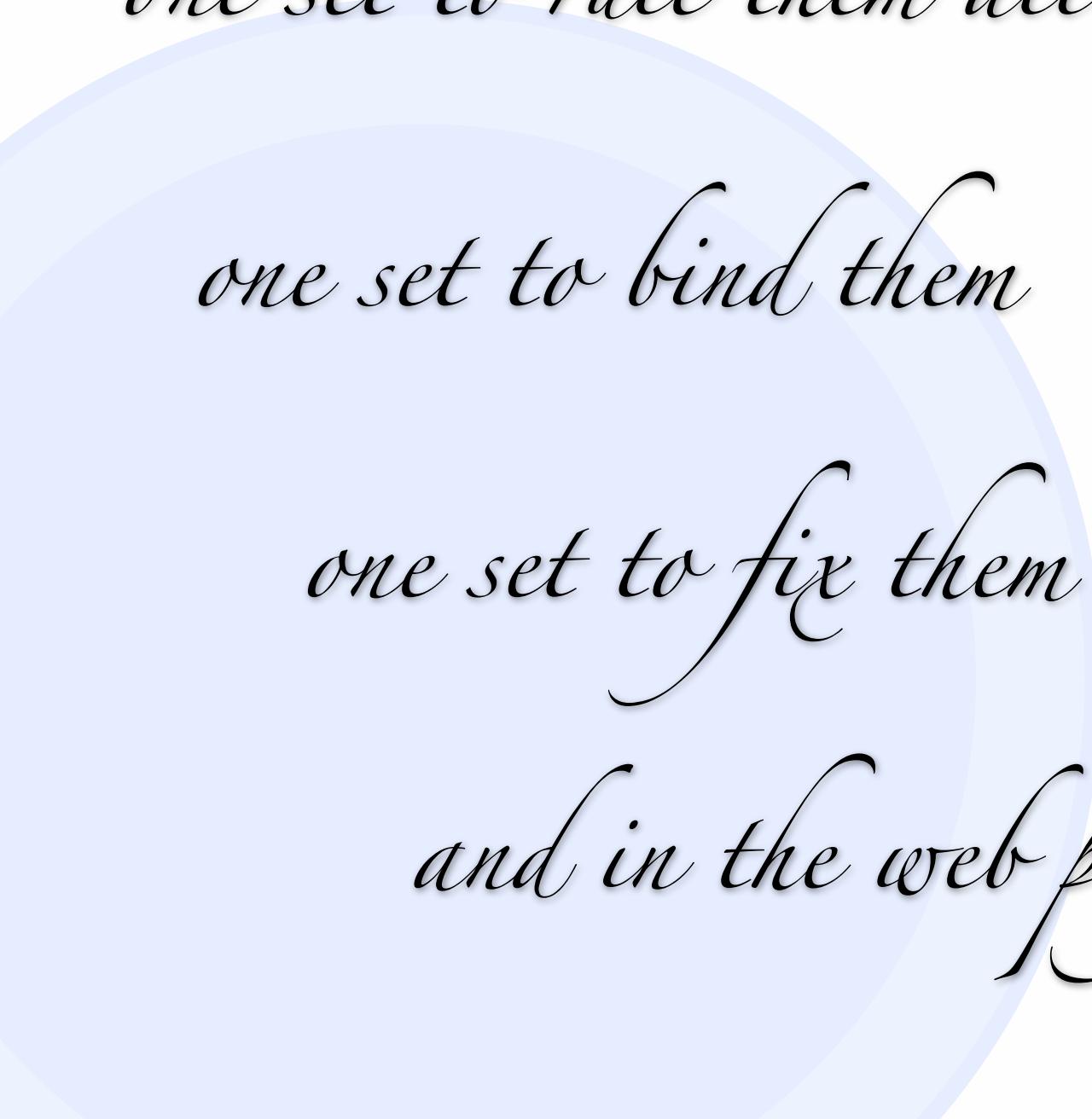
no matter what the program,

no matter what the language.

Unicode

'jʊnɪkəd	ન્યૂડે	યુનિકોડ	યુનિકોડ	يونىكود
Ūnicōdē	GhAઉ	Ioúníkouṁ	યૂનિકોડે	Юникод
*统一码	统一码	યૂનિકોડ	유니콘	યુનિકોડ
يونيكود	統一碼	યૂનિકેડ	യુનિકોડ	युनाइकोड
统一码	统一碼	યૂનિકેડ	যুনিকোড	યુનિકોડ
统一码	统一碼	યૂનિકોડ	统一碼	统一碼

• • •



one set to rule them all

one set to bind them

one set to fix them all

and in the web page hide them



business challenges

- Supporting languages needed for business
 - here or abroad
- Adding new languages (customers), easily



business challenges

- Increasingly, users are not satisfied with incorrect spellings, or restrictions to write their names, addresses, and other information in ASCII or incompatible encodings



technical challenges

- Differences in character encodings
- Require different algorithms
- Imply different code in each market
- High error rate and poor quality



four-letter words



data



easy



find

A large, semi-transparent light blue circle is centered in the background, partially overlapping the word "give".

give



keep



unicode benefits

- allows for multilingual text using any or all the languages you desire
- invoice or ticketing applications can print customer information in their native languages from a single database



unicode benefits

- one way to process text
- one version of the product can be used worldwide



unicode benefits

- text in any language can be exchanged worldwide
- eliminates data corruption and other problems due to incompatible code pages or missing conversion tables



unicode benefits

- support of Unicode by modern technologies extends code life and broadens integration possibilities
- easier to take advantage of new technologies and integrate with other applications



unicode benefits

- 
- **internet-ready**
 - XML, JavaScript, Firefox, Java, and now PHP, all Unicode-based



unicode standard

- Developed by the Unicode Consortium
- Covers all major living scripts
- Version 5.0 has 99,000+ characters
- Capacity for 1 million+ characters



unicode standard

- One character set for worldwide use
- Standard encodings: UTF-8, UTF-16, UTF-32
- International Standard – ISO 10646
- Precisely defined
- Widely supported by standards & industry



unicode standard

- Required by Web & modern applications
- International Domain Names
- DOM



unicode character set

- **Code Points 0 to 10FFFF, (Maximum 21 Bits)**
 - Unicode notation for code point is U+hhhh
 - 17 Planes of 64K (FFFF) code points
- **Basic Multilingual Plane, BMP (U+0000 – U+FFFF)**
 - Commonly used characters in living scripts
- **1st Supplementary Plane (U+10000 – U+1FFFF)**
 - Archaic, fictional characters
- **2nd Supplementary Plane (U+20000 – U+2FFFF)**
 - Ideographs



Organized by scripts into blocks

Example Unicode Characters

ASCII	ABCDEFGHIJKLMNP
Latin-1	ÀÁÃÄÅÆÇÈÉÊËÌÍÏЇ
Latin-2	āĂăĀăĆćĈĉĊċƉďƉ
Greek	ΪΑΒΓΔΕΖΗΘΙΚΛΜΝΞ
Cyrilllic	рстуфхцҹш҃ъыъэյя
Thai	ກມຢຣຖລງວສໜສຫ້ວ່າ
CJK	北丂丢丄丂丂丂丂丂丂丂丂丂丂
Korean	감갑값갓갓갓강갓갓

The primary scripts currently supported by Unicode 4.0 are:

- Arabic
- Armenian
- Bengali
- Bopomofo
- Buhid
- Canadian Syllabics
- Cherokee
- Cypriot
- Cyrillic
- Deseret
- Devanagari
- Ethiopic
- Georgian
- Gothic
- Greek
- Gujarati
- Gurmukhi
- Han
- Hangul
- Hanunóo
- Hebrew
- Hiragana
- Kannada
- Katakana
- Khmer
- Lao
- Latin
- Limbu
- Linear B
- Malayalam
- Mongolian
- Myanmar
- Ogham
- Old Italic (Etruscan)
- Osmanyia
- Oriya
- Runic
- Shavian
- Sinhala
- Syriac
- Tagalog
- Tagbanwa
- Tai Le
- Tamil
- Telugu
- Thaana
- Thai
- Tibetan
- Ugaritic
- Yi

••• generative

- Composition can create “new” characters
- Base + non-spacing (combining) character(s)

A + ° = Å

U+0041 + U+030A = U+00C5

a + ^ + . = â

U+0061 + U+0302 + U+0323 = U+1EAD

a + ˇ + ˇ = ˇ

U+0061 + U+0322 + U+030C = ?



unicode != i18n

- Unicode simplifies development
- Unicode does not fix all internationalization problems

• • • definitions

Internationalization

I18n

To design and develop an application:

- ✓ without built-in cultural assumptions
- ✓ that is **efficient** to localize

Localization

L10n

To tailor an application to meet the needs of a particular region, market, or culture

date formats



- USA: 2/16/05
- France: 16.2.05 or 16-2-05
- Japan, China: 2005年2月16日

calendars



- Gregorian 2007
- Thailand: 2550 (Buddhist Year)
- Taiwan: 96 (1911-based)
- Hebrew: 5767
- Also Hijri (Islamic), Lunar (Asia)
and many others

time formats



- USA: 4:00 P.M.
- France: 16.00
- Japan: 1600
- Don't forget to identify the time zone

number formats



- England: 12,345.67
- Germany: 12.345,67
- Switzerland: 12'345,67
- Swiss money: 12'345.67
- France: 12 345,67
- India: 12,34,567.89

currency

- Symbol placement
- Symbol length (1-15)
- Number width
- Number precision:
 - Spain, Japan 0
 - Mexico, Brazil 2
 - Egypt, Iraq 3

US \$12.34
12.345,67 €
12\$34€
¥123

sorting

English:

ABC...RSTUVWXYZ

German:

AÄB...NOÖ...SßTUÜV...YZ

Swedish/Finnish:

ABC...RSTUVWXYZÅÄÖ

- Languages may sort more than one way
 - traditional vs. modern Spanish
- Japanese stroke-radical vs. radical-stroke
- German dictionary vs. phone book

- Swedish: $z < ö$
- German: $ö < z$
- Dictionary: $öf < of$
- Phonebook: $of < öf$
- Upper-first: $A < a$
- Lower-First: $a < A$
- Contractions: $H < Z$, but $CH > CZ$
- Expansions: $OE < œ < OF$

locale data ● ● ●

- I18N and L10N rely on consistent and correct locale data
- Problem with POSIX locales: not always consistent or correct

- Hosted by Unicode Consortium
- Goals:
 - Common, necessary software locale data for all world languages
 - Collect and maintain locale data
 - XML format for effective interchange
 - Freely available
- Latest release: July 2007 (CLDR 1.5)
- 394 locales, with 135 languages and 149 territories



PHP 6



APC

bundled by default



register_globals





magic_quotes_*





safe mode



A large, light blue semi-transparent circle is centered on the word "dl()", partially overlapping the text below.

dl()

disabled, except in CGI, CLI, and EMBED



unicode support

- Everywhere:
 - in the engine
 - in the extensions
 - in the API



unicode support

- Native and complete
 - no hacks
 - no mishmash of external libraries
 - no missing locales
 - no language bias

string types



Unicode

- text

- default for literals, etc

Binary

- bytes

- everything \notin Unicode type

string types



- internal processing: Unicode
- interface to outside world: binary

● All string literals are Unicode

```
$str = "Hello, world!"; // Unicode string
echo strlen($str);      // result is 13
```

```
$jp = "検索オプション"; // Unicode string
echo strlen($jp);       // result is 7
```

● String offsets work on code points

```
$str = "大学";    // 2 code points
echo $str[1];    // result is 学
$str[0] = 'サ'; // full string is now サ学
```

identifiers

Unicode identifiers are allowed

```
class コンポーネント {  
    function ફુંક્શન્સ { ... }  
    function சிவாஜி கணேசன் { ... }  
    function ପ୍ରାଣ୍ୟମା { ... }  
}  
  
$provider = array();  
$provider['רַעֲיוֹלָה שְׁנָה'] = new コンポーネント();
```

functions

● Functions understand Unicode text

- `strtoupper()` and friends do proper case mapping

```
$str = strtoupper("fußball"); // result is FUSSBALL
```

```
$str = strtolower("ΣΕΛΛΑΣ"); // result is σελλάς
```

- `strip_tags()` works on complex text

```
$str = strip_tags("雅<span>είναι</span>通");
```

- `strrev()` preserves combining sequences

```
$u = "Viء\u0302\u0323t Nam"; // Việt Nam
$str = strrev($u);           // result is maN t̄eiV,
                            // not maN ūeiV
```

streams

- Built-in support for converting between Unicode strings and other encodings on the fly
- Reading from a UTF-8 text file:

```
$fp = fopen('somefile.txt', 'rt');  
$str = fread($fp, 100); // returns 100 Unicode characters
```

- Writing to a UTF-8 text file:

```
$fp = fopen('somefile.txt', 'wt');  
fwrite($fp, $uni); // writes out data in UTF-8 encoding
```

- Grab first 5 titles from Reuters China feed, clean up, and send out as JSON

```
$xml = simplexml_load_file(  
    'http://feeds.feedburner.com/reuters/CNTbusinessNews/');  
  
$titles = array();  
$i = 0;  
foreach ($xml->channel->item as $item) {  
    // each title looks like this: (台灣匯市) 台幣兌美元  
    $title = preg_replace('!\p{Ps}.*\p{Pe}\s*!', '', $item->title);  
    $titles[] = $title;  
    if (++$i == 5) break;  
}  
  
echo json_encode($titles);
```



and more...



pecl/intl

- Builds on the ICU library
- Uses CLDR data
- Works in PHP 5 and 6

features



- Locales
- Collation
- Number and Currency Formatters
- Date and Time Formatters
- Time Zones
- Calendars
- Message Formatter
- Choice Formatter
- Resource Handler



PHP upgrade path



Remember:
PHP 6 = PHP 5 + Unicode



Ask yourself:



Can I live with PHP 4?

Until 8/8/8.



If yes, you're done.
Thanks for playing.



If no,
Do I need Unicode support?



No? Think about it.
Really no?
Upgrade to PHP 5.



If yes to Unicode support:
Is PHP 6 out?



Yes: you win.

Collect £200 and a copy of PHP 6.



No: you lose. Improvise.



PHP dev dynamics



what php 6 will *not*
help you do

i can haz sprinkl?



localize for LOL cats

OMGZ0rz WTF n00b
u d0n7 5|*34k l33t!
u r n0t 4 h4><0rz!!!

leverage l33t speak

VENDREDI, MAI 25, 2007

I CAN HATH CHEEZBURGER?

Myn gentil rederes, the joly tyme of Averille and
May hath not been of much jolitee to me – in
feyth, ich haue had but litel tyme to look upon the
newe floures and heere the smale foules doyng
their thinge, for cursid busynesse hath fallen
a-newe vpon me. I was prikked to take thys
biswinkful newe labor by grete nede, for whan ich
madde myn accountes ich discovered gret dettes
and but litel revenue. Thomas, who ys wyth my
Lord John of Gaunt in Spayne, had gret need for

write in Middle English



kthxbye