

Večjezičnost

давайте говорить по русский

Gašper Kozak
<?php konferenca 2010

Kdo

- Gašper Kozak, 31 let
- fatg na php-si.com in pehape.si
- v Tobonetu razvijam spletne oglaševalske sisteme
- v prostem času: WideImage, LayerCache, varnost spletnih aplikacij

Oris delavnice

- uvod
- zaznava jezika in predstavitev večjezičnosti
- kodni nabor in operacije z nizi
- jezik besedil vmesnika in podatkov
- izpis in vnos večjezičnih podatkov
- problemi
- izdelane rešitve
- zaključek

Uvod

- $i18n$ = internationalization
 - priprava aplikacije na večjezičnost
- $L10n$ = localization
 - prilagoditev za posamezen jezik
- $i18n + L10n = g11n = globalization$

Zaznava jezika: IP naslov

- država != jezik
 - Belgija: NIZ 60%, FRA 40%
 - Švica: NEM 63%, FRA 20%, ITA 6%
- roaming
- odsvetovan, težji in največkrat uporabljen pristop
 - inštalacija paketa: geoip
 - zastonjska verzija ni 100% natančna





Zaznava jezika: Accept-language

- brskalnik pošlje strežniku
 - sl;q=0.8,en;q=0.6,en-us;q=0.6
 - IE: sistemska nastavitvev
 - Firefox: glede na inštalacijo
- priporočena in enostavna rešitev
- ... ampak redkeje uporabljena rešitev. Zakaj?
- večina uporabnikov ima to nastavljeno

Zaznava jezika: cookie/URL

- shranimo v piškotek
- nastavitve prijavljenega uporabnika
- zahtevnejši uporabniki cenijo to možnost
 - geeki radi uporabljamo aplikacije v angleščini
- jezik v URL naslovu (vsekakor)
 - `novica.php?lang=sl&id=234`
 - `/sl/news/234`

Predstavitev večjezičnosti

- naj bo očitno, da znamo русский
- zastavice? Ne.
 - Brazilci  govorijo portugalsko. 
 - angleška zastava:  UK: 
- Oznake: Slovenščina, SL, SLV
- v URL naslovu:
 - informacija uporabniku
 - pajki nimajo piškotkov (načeloma)

Izbira kodnega nabora

- UTF (presenečeni?)
- UTF povsod:
 - PHP datoteke (BOM in nevidni znaki)
 - PHP izvajanje (mb/iconv internal encoding)
 - struktura baze in collation na tabelah in poljih
 - podatki v bazi
 - MySQL povezava: set names utf8
 - HTTP in HTML headerji
- paziti na Accept-encoding ... ampak ne zares :)

MySQL - kodni nabor in razvrščanje

- character set: kodni nabor znakov
 - kodni nabor, v katerem poteka povezava
 - SET NAMES utf8
 - podatki v bazi so vedno v UTF
- collation: pravilo za razvrščanje nizov
 - na nivoju baze, tabele, stolpca, povezave in znotraj poizvedbe za posamezno polje

Operacije z nizi

- strlen, substr ... adijo!
- mb_strlen, iconv_strlen ...
- razvrščanje se zakomplicira
 - MySQL: COLLATE
 - PHP:
 - setlocale + sort + SORT_LOCALE_STRING
 - Collator (intl extension)
 - lastna rešitev (usort)

intermezzo: locale

- problem s threaded Apache ali Windows
 - locale se spreminja celemu procesu
- Okna ne podpirajo UTF preko setlocale
 - le cp1250
 - vpliva na primerjavo UTF nizov (sort)
 - deluje za števila in datume (cp1250)
- sistemski poseg: inštalacija ustreznih paketov na OS
- ne upoštevajo ga vse funkcije (date, number_format, ...)

Jezik besedil na vmesniku

- enostavni katalog: const, PHP array
 - enostaven, ampak nestandarden format
- gettext
- kak drug datotečni katalog (XML; XLIFF)
- baza
- MessageFormatter (intl)
- prevajalci niso geeki:
 - poenostavitev in avtomatizacija: export + prevod + import

Jezik besedil v podatkovni bazi

- sprememba podatkovne baze
 - polje *jezik* v tabeli
 - polja *naslov_en*, *naslov_sl*, ... (haha!)
 - *tabela* + *tabela_lang*
 - oteži poizvedbe
- vključitev v URL naslove
 - */en/news/2010/july/9/new-battery*
 - */sl/novica/2010/julij/9/nova-baterija*
 - */sl/news/123* (enostavno, združeno)

Izpis numeričnih podatkov

- števila, datumi, valute
 - number_format, sprintf
 - NumberFormatter (intl)
 - date, strftime
 - IntlDateFormatter (intl)
 - money_format (locale aware)
 - ampak privzame valuto ...
 - framework pomaga

Izpis večjezičnih besedil

- `select ... where lang = 'sl' ...`
- ... to je to :)

Vnos večjezičnih podatkov

- števila, datumi, valuta
 - sscanf (locale, sprintf format)
 - strptime (locale, strftime format)
 - ročno razčlenjanje (nooooooooooooooooooooo)
 - NumberFormatter
 - IntlDateFormatter
- spelling
 - Enchant
 - pspell

Problemi

- število
 - You have received %d new message(s).
 - Prejeli ste 1 sporočil.
 - Število sporočil, ki ste jih prejeli, je enako 1.
 - gettext zna! Ampak samo po koščkih.
 - MessageFormatter raztura
- spol
 - Oseba Gašper Kozak je napisala. WTF?

Izdelane rešitve

- gettext, strftime, number_format, ...
- intl razširitev
- PHP-FLP in podobni projekti
- frameworki
 - Zend_Locale + Zend_Translate
 - Symfony
- kitajski pristop: link na Google Translate

Zaključek

- vgraditi čim prej (vendar ne prekmalu)
- orodja:
 - pogledj, kaj zna framework
 - intl
 - setlocale + php funkcije (fuj)
 - ročno (hec)