

Konwersja ISO->UTF i na odwrót

Autor: Jokris
17.01.2008.

O konwersji systemów kodowania znaków pisałem wielokrotnie na Forum Jokris.info, np. TUTAJ. Jeden z użytkowników mojego forum napisał: "Próbowałem skonwertować plik polish.php do utf-8, ale nie przyniosło to efektu. Krzaczki były zastępowane przez trochę inne krzaczki. Program do konwersji (Edit pad pro) twierdzi, że plik polish.php jest kodowany (original encoding) w Windows 1252: Western European." Chociaż każdy ma swój sposób na brak polskich znaków, podam wam opis metody, którą preferuję ja. I przynosi zawsze pozytywny efekt końcowy. Temat ten jest jak najbardziej aktualny, bo ostatnio, słusznie czy nie słusznie, część użytkowników przestawiła się na kodowanie UTF-8 dla Joomla. Na początek ważna informacja: Żeby przekonwertować (przekodować) tekst z ISO-8859-2 na UTF-8 należy najpierw dokonać konwersji tekstu z ISO-8859-2 na Windows-1250. Teraz trochę teorii na temat systemów kodowań znaków. Windows-1250 jest to strona kodowa której używacie na co dzień np. w Notatniku systemowym, Wordzie, czy chociażby wprowadzając tekst z klawiatury do postu na Forum. Czyli z polskimi znakami diakrytycznymi, jak np "Ś, ś, ą, ź" i.t.d. Bo musicie wiedzieć, że ISO-8859-2 jest kojarzone ze standardem polskich znaków, ale tak w rzeczywistości to wszystkie litery z ogonkami zastąpione są innymi znakami, które mają z językiem polskim tyle wspólnego, co nic. Dopiero przeglądarka internetowa rozkodowuje tekst (po to mamy w menu przeglądarki listę z kodowaniem znaków), i przedstawia go już z polskimi ogonkami. Dlatego nazywa się to "SYSTEM KODOWANIA ZNAKÓW". ISO-8859-2, czy też UTF-8 są to systemy kodowania, które zamieniają polskie ogonki (w przypadku ISO-8859-2) i również polskie nietypowe znaki, jak np. "ł, ż" (w przypadku UTF-8) na ich odpowiedniki zakodowane (po prostu zamienione) w określonym w w/w systemach kodowania. I zamiast literki ą będziesz miał znak ą. Trochę teorii a teraz do konkretów. Jeśli te "konkrety" interesują Was, kliknijcie w poniższy odnośniczek...

- Do przekodowywania pomiędzy systemami kodowania znaków używaj jedyne, sprawdzonego programu o nazwie Gzegzółka XP. Gzegzółka to konwerter standardów kodowania polskich znaków diakrytycznych (zwanymi potocznie ogonkami). Oprócz ogonków obsługuje również większość standardów kodowania stosowanych na całym świecie.

- Jak robić konwersję?. Po prostu wystarczy otworzyć np. taki plik "polish.php" za pomocą drugiego, jedyne i sprawdzonego programu (a co, jak za komuny) jakim jest Notatnik SP PL. Ja w tym edytorze przetłumaczyłem wiele plików jak i też edytuję kod PHP lub HTML. Nie znam lepszego. (oczywiście to tylko moja opinia, nie zawsze jedyna i słuszna).

- Po otwarciu pliku z kodowaniem znaków ISO-8859-2, wystarczy zaznaczyć w menu "Konwersja" => kodowanie Windows-1250". I już zamiast dziwnych znaków będziesz miał polskie literki. A musicie to zrobić dlatego, aby Gzegzółka XP poprawnie rozpoznała źródło, i system znaków jakie ma przekonwertować.

```
{mosimage cw=300 popup=1 popupTyp=dhtml}
```

- Po tej operacji (zawsze róbcie kopie oryginalnych plików) uruchamiasz program Gzegzółka XP i pojawi Ci się jego okienko. Wrzucasz do niego metodą "przeciągnij - upuść" plik "polish.php", ale oczywiście ten w zestawie znaków Windows-1250. Na dole masz 2 pola typu "lista", rozwijalne. W lewym polu wybierasz zestaw znaków "Windows 1250(Europa Środkowa)", natomiast w prawym polu, który jest docelowym, wybierasz interesujący Cię zestaw znaków, czyli "Unicode UTF-8". Naciskasz tylko "Start". Potwierdzasz, że chcesz dokonać konwersji, i w miejscu pliku "polish.php" zakodowanym w standardzie Windows 1250 pojawi się "polish.php" w kodowaniu UTF-8. Program automatycznie wykonuje kopię oryginalnego pliku "polish.php". I po wszystkim.

Proste?. No pewnie teraz tak. Musicie pamiętać, że po przekodowaniu pliku przykładowego "polish.php" z ISO-8859-2 na UTF-8, w zawsze słusznym Notatniku SP PL nie zobaczycie polskich znaków, tylko bardzo dziwne "krzaczki", które to w/w edytor nie potrafi wyświetlić. Zresztą jak i wiele innych edytorów. Sprawdzicie poprawność kodowania w Notatniku systemowym. Tam powinniście mieć polskie znaki. Jeśli są, oznacza to, że konwersja udała się. Notatnik systemowy obsługuje kodowanie UNICODE. Szkoda że nie mają tej cechy edytory, chociażby dostępne na mojej stronie.

Jokris (Krzysiek Stachyra). 18 stycznia 2008r.