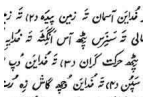
Определить кодировку стороннего сайта PHP

Автор: Administrator
23.11.2012 11:06



Иногда при написании парсера необходимо знать кодировку сайта донора. В большинстве случаев парсер пишется на какой то определенный сайт и его кодировку мы смотрим сами в исходном коде. Но бывают и другие случаи, к примеру перед вами стоит задача **получать титл** с совершенно случайного сайта и тут без знания его кодировки нам необойтись (в противном случае мы возможно будем наблюдать каракули крокозяблы).

В интернете я видел различные функции для определения кодировки, они у меня не работали. Порой определяли правильно, а порой нет. Возможно в этом моя вина я что то делал не так.

И так перед тем как писать код который будет определять кодировку нам необходимо знать каким образом это делать. У каждого грамотно созданного сайта есть в исходном коде meta тег который и сообщает браузеру кодировку сайта.

```
<meta http-equiv="content-type" content="text/html; charset=utf-8" />
```

Раз уж этот мета тег говорит браузеру о кодировке то почему бы и не сделать чтоб он сообщал и нашему скрипту ее. На сколько я знаю сайты делаются в основном в двух кодировка это utf-8 и win-1251. Сайты в других кодировках я не видел, если видели вы то покажите мне я с удовольствием взгляну и пополню свои знания.

И так исходя из наших рассуждений нам необходимо получить исходный код сайта и вытащить из его meta тега кодировку. Вытаскивать будем с помощью регулярного выражения. Так же необходимо учитывать что старые сайты или просто сайты сделанныйе новичками могут и не содержать meta тега с кодировкой, а это нужно учитывать.

И так функция:

```
function detect_charset($str){
```

```
 if(preg_match("/windows-1251/i", $str)){
 return "windows-1251";
 }
 elseif(preg_match("/utf-8/i", $str)){
 return "utf-8";
 }
 else{
 return "windows-1251";
 }
 }
```

Функция из переданного ей исходного кода парсит кодировку, если она кодировку не находит то возвращает windows-1251, обычно сайты без кодировки и сделаны в кодировке виндовса. По сути в лучшем случае регулярное выражение нужно доработать, чтоб парсить именно из meta тега (сделаем это чуть позже).

Функция готова необходимо ее испытать. Испытания будем проводить с помощью функции парсинга title сайта. Вот она:

```
function title($urls){
 if($urls!=''){
 $content = file_get_contents($urls);
 if(detect_charset($content)=="windows-1251"){
 $content = iconv("cp1251", "UTF-8", $content);
 }
 preg_match('|<title>(.*)</title>|Uis', $contents, $title);
 if ($title[1]==''){
 $contents=$content;
 preg_match('|<title>(.*)</title>|Uis', $contents, $title);
 }
 return "<h3 class='name_site'>".$title[1]."</h3>";
 }
 }
```

Суем обе функции в один файл и проверяем работу передавая ей сайты в различноей кодировке:

```
echo title("http://.yandex.ru");
```

Представленный в этой теме код подходит лишь для ознакомления. Для его использования в реальном проекте необходимо его доработать.