КОДИРОВКА ТЕКСТОВ

Носиков Андрей Алексеевич

студент, филиал федерального государственного бюджетного образовательного учреждения высшего образования Национальный исследовательский университет «МЭИ» в г. Смоленске, РФ, г. Смоленск

TEXT ENCODING

Andrey Nosikov

Student, Branch of the Federal State Budgetary Educational Institution of Higher Education National Research University MPEI in Smolensk, Russia, Smolensk

Аннотация. Кодировка текста является основным рабочим инструментом веб-мастеров. Это связано с тем, что верстка html-документов и web-программирование практически всегда подразумевает работу с кодировкой файлов, ведь при неверно выбранной кодировке исходного текста существует вероятность некорректного воспроизведения информации различными браузерами. Это происходит из-за того, что программы не всегда способны автоматически определить кодировку исходного текст а в автоматическом режиме. В случае неверно выбранной кодировки, пользователь, работающий с информацией, увидит хаотично-напечатанный текст вместо ожидаемого текста. Именно кодировке текстов и анализу различных ее вариаций посвящена данная статья.

Abstract. Text encoding is the main working tool of webmasters. This is due to the fact that the layout of html documents and web programming almost always involves working with the encoding of files, because if the source code is incorrectly selected, there is a possibility of incorrect reproduction of information by different browsers. This is due to the fact that programs are not always able to automatically detect the encoding of the source text in automatic mode. In the case of an incorrectly selected encoding, the user working with the information will see randomly printed text instead of the expected text. This article is devoted to text encoding and analysis of its various variations.

Ключевые слова: Кодировка текстов, браузеры, текст, информация, воспроизведение.

Keyword: Text encoding, browsers, text, information, playback.

В современных электронно-вычислительных машинах символы способны храниться исключительно в виде последовательности бит, также, как и числа. Каждому символу текста должна соответствовать оригинальная последовательность нулей и единиц с целью корректной передачи на компьютерном языке. Для достижения этой цели были разработаны и созданы уникальные таблицы кодировок.

Количество символов, задаваемых длиной n можно вычислить по формуле C(n)=2n. Исходя из этого, относительно требуемого количества символов – напрямую зависит емкость используемой памяти.

Изначально, при первых попытках разработать кодировку текста, на каждый символ отводилось по пять бит. Данный факт был связан с малой оперативной памятью компьютеров тех лет. В эти 6465 символа могли умещаться только управляющие символы, а также строчные буквы английского алфавита.

Параллельно с ростом мощностей компьютеров стали появляться таблицы кодировок, имеющие большее количество символов. Первой семибитной кодировкой является ASCII7. В нее уже вошли прописные буквы английского алфавита, арабские цифры, знаки препинания.

Далее, на ее базе была создана ASCII8, в которой уже стало возможным хранение 256256 символов: 128128 основных и ровно столько же расширенных. Первая часть таблицы осталась без каких-либо изменений, а вторая способна иметь различные варианты (каждый имеет свой номер). Эта часть таблицы стала заполняться символами национальных алфавитов.

Для большинства языков, к примеру, японского, арабского, китайского, такое количество символов не является достаточным, именно поэтому развитие кодировок не переставало останавливаться, что впоследствии привело к разработке UNICODE.

Модернизация кодировок текстов происходило параллельно с формированием сферы информационных технологий. За это время они смогли перетерпеть немалое количество изменений. Относительно истории, все началось с EBCDIC, которая предоставляла возможность кодировать буквы латинского алфавита, арабские цифры и знаки пунктуации с управляющими символами.

Несмотря на это, отправной точкой для развития кодировок текстов наших дней стоит считать знаменитую ASCII. Именно с ее помощью описываются первые 128 символов из наиболее часто используемых англоязычными пользователями — латинские буквы, арабские цифры и знаки препинания.

Еще в эти 128 знаков, описанных в ASCII, попадали некоторые служебные символы навроде скобок, решеток, звездочек и т.п.

Дальнейшее развитие кодировок текста было связано с тем, что набирали популярность графические операционные системы и необходимость использования псевдографики в них со временем пропала. В результате возникла целая группа, которая по своей сути по-прежнему являлись расширенными версиями Аски (один символ текста кодируется всего одним байтом информации), но уже без использования символов псевдографики.

Они относились к так называемым ANSI кодировкам, которые были разработаны американским институтом стандартизации. В просторечии еще использовалось название кириллица для варианта с поддержкой русского языка. Примером такой может служить Windows 1251.

Она выгодно отличалась от используемых ранее CP866 и KOI8-R тем, что место символов псевдографики в ней заняли недостающие символы русской типографики (окромя знака ударения), а также символы, используемые в близких к русскому славянских языках (украинскому, белорусскому и т.д.)

Эти тысячи знаков языковой группы юго-восточной Азии никак невозможно было описать в одном байте информации, который выделялся для кодирования символов в расширенных версиях ASCII. В результате был создан консорциум под названием Юникод (Unicode — Unicode Consortium) при сотрудничестве многих лидеров IT индустрии (те, кто производит софт, кто кодирует железо, кто создает шрифты), которые были заинтересованы в появлении универсальной кодировки текста.

Различные наборы символов сложились исторически и вследствие естественного развития компьютерной техники за последний полувек. Кодировка текста ASCII — один из первых наборов, разработанный в 1963 году и используемый до сих пор.

Первоначально таблица содержала всего 128 символов, среди которых были буквы латинского алфавита, цифры и специальные символы. В дальнейшем это число было расширено до 256 — это позволило использовать буквы национальных алфавитов, в том числе и русского. Однако порядок и способ указания подобных символов не был регламентирован, что породило несколько несовместимых между собой кодировок: Windows-1251, КОИ-8. Помимо указанных кодировок, существовали также несовместимые (не-ASCII) варианты — например, CP866.

Стандарт Unicode (Юникод) был разработан для решения этих проблем. На нём основаны наборы символов UTF-8, UTF-16, UTF-32, самым популярным из которых является UTF-8. Обычно его и применяют для вёрстки современных web-страниц; на нём также основана работа большинства систем, таких как WordPress и Joomla.

Кодировка текста UTF-8 поддерживает множество специальных символов (например, диакритические знаки и псевдографику), иероглифы и т.д. На сегодняшний день Юникод — самая универсальная кодировка текста.

Список литературы:

1. Агеев В.М. Теория информации и кодирования: дискретизация и кодирование измерительной информации. — М.: МАИ, 1977.

2. Кузьмин И.В., Кедрус В.А. Основы теории информации и кодирования. — Киев, Вища школа, 1986.

3. Простейшие методы шифрования текста/ Д.М. Златопольский. – М.: Чистые пруды, 2007 – 32 с.