ЕВ-СТРАНИЦ

[HTML, HTML-документ, тэги, HTML 5, XML]

4.1. HTML – язык гипертекстовой разметки документа

Гипертекстовый WWW-документ может содержать стилизованный и форматированный текст, графику и гиперсвязи с различными ресурсами Internet. Чтобы реализовать все эти возможности, был разработан специальный компьютерный язык, названный *HyperText Markup Language (HTML)*, – язык разметки гипертекста. Документ, написанный на HTML, представляет собой текстовый файл, содержащий собственно текст, несущий информацию, и флаги разметки (*markup tags*). Флаги представляют собой определенные стандартом HTML последовательности символов, заключенные между знаками **<** и **>**. Согласно флагам разметки программа располагает текст на экране, включает в него изображения, хранящиеся в отдельных графических файлах, и формирует гиперсвязи с другими документами или ресурсами Internet. Файл на языке HTML принимает «гипертекстовый облик» только тогда, когда он интерпретируется программой просмотра (браузером). HTML предназначен для написания гипертекстовых документов, публикуемых в World Wide Web. Документ на языке HTML может включать следующие компоненты:

- стилизованный и форматированный текст;
- команды включения графических и звуковых файлов;
- гиперсвязи с различными ресурсами Internet.

4.1.1. Составляющие HTML-документа

Документ, написанный на HTML, представляет из себя текстовый файл, который можно писать и редактировать при помощи любых текстовых редакторов. Он включает в себя:

- собственно текст,
- специальные последовательности символов,
- флаги разметки.

Графическая и звуковая информация, включаемая в HTML-документ при помощи специальных команд, хранится в отдельных файлах. Программы просмотра HTML-документов интерпретируют флаги разметки и располагают текст и графику на экране соответствующим образом. Для файлов, содержащие HTML-документы, принято расширение ***.html*** (на серверах с операционными системами *UNIX*, *WindowsXXX* и др.). В свое время на серверах с операционной системой *DOS* использовался суффикс ***.htm***.

Текст. Последовательность символов, составляющая текст, может состоять из пробелов, табуляций, символов перехода на новую строку, символов возврата каретки, букв, знаков препинания, цифр, и специальных символов (например, +, #, **$**, **@**), за исключением следующих четырех символов, имеющих в HTML специальный смысл:

< (*меньше – Less Than*),

> (*больше – Greater Than*),

& (*и, амперсанд – Ampersand*),

" (*двойные кавычки – Double Quote*).

Символ табуляции, символ возврата каретки и символ перехода на новую строку считаются эквивалентными пробелу, а несколько следующих друг за другом пробелов и/или табуляций

и/или символов возврата каретки и/или символов перехода на новую строку эквивалентны ровно одному пробелу, за исключением случая предварительно отформатированного текста.

Тэги (Tags, Флаги). Флаги предназначены для форматирования и разметки документа. Любой флаг начинается символом < и заканчивается символом >. В названиях флагов строчные и прописные буквы считаются эквивалентными. Например, флаг
 может быть записан как
. Существует два вида флагов: *парные* и *непарные*. Действие любого парного флага начинается с того места, где встретился открывающий флаг и заканчивается при встрече соответствующего закрывающего флага (признаком которого является символ /, следующий сразу после <) или конца файла. Например, текст, следующий за флагом курсивного начертания <I> и продолжающийся до его закрывающего парного флага </I>, выводится на экран курсивом.

Непарный флаг вызывает «единичное» действие в том месте, где он встречается. Например, флаг
 служит для перехода на новую строку при выводе текста.

Многие флаги могут включать дополнительные *параметры*, или *атрибуты*, модифицирующие эффект данного флага, например: <P> – флаг начала параграфа; <P ALIGN=CENTER> – флаг начала параграфа, выровненного по центру окна.

4.1.2. Структура HTML-документа

Типичный HTML-документ имеет *головную часть* и *тело*. Начало документа отмечается флагом <HTML>, а конец – флагом </HTML>.

Синтаксис

```
<HTML>
<HEAD><TITLE>...</TITLE>
</HEAD>
<BODY>...</BODY>
</HTML>
```

Пример:

```
<HTML>
<HEAD><TITLE>Vasya's Homepage</TITLE>
</HEAD>
<BODY>Добро пожаловать ко мне в гости!<BR>Рад вас видеть у себя дома.
</BODY>
</HTML>
```

Головная часть документа (Head). Головная часть документа является служебной. Она обычно включает в себя *название документа* (см. далее). Кроме того, в нее часто помещается *<META>-информация* – ключевые слова и описание документа, которые читаются некоторыми программами-роботами.

Синтаксис

```
<HEAD>...</HEAD>
```

Пример:

```
<HEAD><TITLE>Caucasian Ovcharka Homepage</TITLE></HEAD>
```

Название документа (Title). Название документа помещается внутри его головной части. Оно выводится не вместе с самим документом, а в полосе заголовка окна программы просмотра. Оно также используется и для других целей. Например, когда программа просмотра создает так называемую *закладку,* то есть запоминает местонахождение

документа, к которому предполагается в дальнейшем вернуться, этой закладке присваивается имя, которое берется из названия документа. Поле *title* в документе является обязательным. Его не следует делать длиннее 64 символов. Название документа должно быть осмысленным, поскольку это поле читается программами-роботами и заносится в базы данных поисковых систем.

Синтаксис

```
<TITLE>...</TITLE>
```

Пример:

```
<HEAD><TITLE>Caucasian Ovcharka Homepage</TITLE></HEAD>
```

Тело (Body). Определяет "видимую" часть HTML-документа. В документе должно быть только одно тело.

Синтаксис

```
<BODY>...</BODY>
```

Пример:

```
<BODY>Это крошечный HTML-документ.</BODY>
```

Комментарий (Comment). Комментарий – это текст, который игнорируется программой просмотра. Комментарий предназначен в первую очередь для самого автора документа и может содержать дату создания, версию, замечания и т.п. Комментарии могут находиться в любой части документа, но не внутри флагов.

Синтаксис

```
<!-- текст_комментария -->
```

Пример:

```
<!-- This document was created from RTF source by rtftohtml version 2.7.5 -->
```

В виде комментария в HTML-файл могут быть помещены различные инструкции для WWW-сервера и других служебных программ. Например, при сответствующей настройке WWW-сервера комментарий `<!--#exec cgi="/cgi-bin/counter"-->`, включенный в HTML-файл, будет вызывать запуск программы `counter` каждый раз, когда кто-нибудь "берет" этот файл с сервера.

4.1.3. Таблицы (Tables) в HTML-документах

Таблицы в HTML-документе являются удобным средством форматирования как собственно таблиц, так и «нетабличной» информации. В последнем случае таблицы используют для того, чтобы добиться жестко заданного взаимного расположения частей Web-страницы в окне программы просмотра. **Таблица** – совокупность *ячеек (cells)*, каждая из которых занимает заданное число *строк (rows)* и *столбцов (columns)*. Таблица может включать ячейки двух видов:

1) ячейки, содержащие *подзаголовки* частей таблицы *(headers)*;

2) «обычные» *ячейки с данными (data cells)*.

Ячейки могут содержать как форматированный в соответствии с правилами HTML текст, так и графику. Таблица может иметь *подпись (caption)*.

4.2. Основные особенности HTML 5

После многих лет успешного использования *HTML 4* в Internet постепенно приходит пора *HTML 5*. Смена стандарта обусловлена, прежде всего, развитием в сети мультимедийности и интерактивности. Разработка пятой версии стандарта происходит в

сотрудничестве компаний *Mozilla*, *Opera*, *Apple* и *Google*, а также консорциума *W3C* (именно эта организация внедряет новые технологические стандарты в мировой Паутине).

Основная задача HTML 5 – правильно интегрировать мультимедийный контент. Пока для этого требуются дополнительные *плагины*[1], самый популярный из которых – *Adobe Flash Player*. В *HTML 5* присутствуют специальные теги и в этом случае использование дополнительных плагинов не потребуются.

Также весьма интересной и полезной представляется функция «*Canvas*». Она описывает размеченную на Web-сайте область, а движок браузера отображает в реальном времени графическое наполнение, например, чертежи, графики или даже простые игры. В дальнейшем планируется и реализация 3D-графики. Для этого разрабатывается стандарт *WebGL*, который для обработки сложных 3D-сцен в свою очередь будет обращаться к OpenGL.

Для того, чтобы скрипты «Canvas» не тормозили браузер, предусматривается поддержка многопоточности. Эта опция носит название «*Web Workers*», она выполняет скрипты и Web-приложения параллельно. Таким образом, Web-сайт со сложным оформлением грузится быстрее, прокрутка страниц делается более плавной и нет задержек при вводе текстовой информации.

Полностью меняется способ хранения информации. Сейчас она пишется в небольшие файлы – *cookies*. А по новой технологии *Web Storage* на стороне клиента будут храниться до 10 Мбайт данных. В Cookies информация сохраняется в виде текстовых файлов, теперь же будет использоваться специальная база данных. С её помощью можно даже хранить специальные Web-приложения и работать с ними без подключения к Internet.

HTML 5 обеспечивает безопасность компонентов. Самая большая угроза в сети исходит из тегов *iFrame* (в этой области отображается содержимое стороннего сайта). Если в этой области содержится вирус, то он может проникнуть на компьютер. В новом стандарте в теги iFrame добавлен фильтр *Sandbox*, который ограничивает действие скриптов с внешних Web-сайтов.

Ещё одна новинка – технология *Web Forms 2.0*. Она более эффективно выполняет обработку введенных пользователем данных, что также обеспечивает более высокую скорость. Количество тегов, используемых в HTML 5, увеличится по сравнению с HTML 4.

4.3. Язык XML

Технология XML (*Extensible Markup Language*) представляет собой нечто большее, чем просто способ представления Web-страниц; с помощью XML набор документов превращается в базу данных (БД). Содержимое документа XML располагается между стандартными тэгами: столь строгая структура кода позволяет всем приложениям без труда выбирать и использовать в своих целях это содержимое. Каждый документ XML становится хранилищем данных, к которому можно обращаться с запросами подобно тому, как можно было бы обратиться к любой БД. К сожалению, правила, регламентирующие порядок упаковки данных, на Web-страницах и методы обработки этих, данных до сих пор не определены и не систематизированы. В результате, Internet сегодня представляет собой беспорядочную и причудливую смесь технологий HTML, *JavaScript* и *Java* на клиентских системах и весьма широкий набор компилируемых языков и языков сценариев на стороне Web-сервера.

XML позволяет разобраться в этих нагромождениях и, упорядочив хаос, организовать его в единую унифицированную сеть. Данные, которые ранее были бессистемно разбросаны по

1 *Плагин* (*plug-in*, *англ.*, от *plug in* – «подключать») – независимо компилируемый программный модуль, динамически подключаемый к основной программе и предназначенный для расширения её возможностей.

страницам HTML, теперь размещаются на строго структурированных документах XML. Все популярные Web-браузеры поддерживают спецификации XML и способны обрабатывать информацию гораздо эффективнее по сравнению со своими предшественниками, которые манипулировали лишь конструкциями HTML.

Эти данные правильно интерпретируются не только Web-браузерами, но и другими XML-совместимыми приложениями. Новое поколение служб электронного обмена данными (EDI) способно связать при помощи средств XML различные бизнес-процедуры, определить соответствующие API-интерфейсы и форматы сообщений.

В некоторых источниках XML представляется упрощенной версией стандартного языка обобщенной разметки *SGML*[2], который, собственно, и положил начало HTML. Тем не менее, XML нельзя считать обычным представителем семейства языков гипертекстовой разметки. И эта технология быстро превращается в основную движущую силу развития объектно-ориентированной мировой Паутины.

Строгая система управления наборами документов лежит в основе большинства операций, требующих особой точности.

Хотя язык HTML создан на основе спецификаций DTD (спецификации определений типов документов – *Document Type Definitions*) для SGML, браузеры никогда не отличались их однозначной интерпретацией. Web-страницы не ограничивали полета творческой фантазии, и любой пользователь мог погрузиться в игру с Web. Но сейчас, когда HTML давно завоевал статус официального языка Internet, подобные вольности уже недопустимы.

XML придает технологии SGML дополнительную строгость и точность, не ограничивая возможность манипулирования огромным количеством HTML-страниц, накопленных в Internet к сегодняшнему дню. Этого удалось добиться за счет упрощения правил определения DTD. Таким образом, чтобы добиться совместимости миллиардов уже размещенных в Internet страниц HTML со спецификациями XML, достаточно приложить минимум усилий.

Вот, например, типичный фрагмент HTML:

```
<img src=/img/fig1.jpg>
```

Эквивалентная конструкция на языке XML будет выглядеть так:

```
<img src="/img/fig1.jpg"/>
```

Внесение небольших изменений превращает конструкции HTML в код XML. Заключение атрибута /img/fig1.jpg, представляющего собой ссылку на графический файл fig1.jpg, в кавычки и добавление в конце косой черты позволяет избежать двусмысленности при автоматическом синтаксическом анализе операторов XML. Страницу XML (так же, как и весь документ) можно считать базой данных, поскольку каждое идентифицируемое в процессе синтаксического анализа поле содержит специфическую информацию, которая распознается, обрабатывается и преобразуется в нужный вид другими приложениями.

Web-браузер *Internet Explorer* способен преобразовать страницу XML в объект, который можно непосредственно обрабатывать средствами таблиц стилей *Extensible Stylesheet Language* (*XSL*). Допускается также косвенная обработка за счет извлечения нужных фрагментов страницы при помощи сценариев *Microsoft VBScript* или *ECMAscript* и последующего встраивания их в модель *Document Object Model* браузера.

Однако Web-браузер решает лишь часть задач. Web-узлы обслуживают массу других бизнес-процедур: с их помощью отслеживается доставка экспресс-отправлений, закупаются

2 *SGML* (*Standard Generalized Markup Language, англ.* – стандартный обобщённый язык разметки) – метаязык, на котором можно определять язык разметки для документов. SGML – наследник разработанного в 1969 году в IBM языка *GML* (*Generalized Markup Language*),

товары, проводятся операции с ценными бумагами. Появляются все новые и новые задачи, многие из которых выполняются безо всякого вмешательства человека.

Например, на Web-узле английской компании *Harvey Bowring Online*, специализирующейся на страховании кредитов, используют инструментарий *GlobalAccess*, разработанный компанией D&B. На каждом этапе работы узла, как данные, так и протоколы запросов и ответов представляются в терминах XML. Это означает следующее. Система может работать где угодно. В технологии электронного обмена данными (EDI) для D&B нет ничего нового. Компания применяет ее уже в течение многих лет. Однако как D&B, так и ее клиенты не могут и, вероятно, не смогут получить глобальный доступ к сетям EDI.

Для обращения к данным применяется протокол HTTP. Поэтому никаких осложнений с межсетевыми экранами в этом случае не возникает.

Для доступа к данным Web-браузер и приложения используют одни и те же унифицированные технологии. Гарантией совместимости является управление протоколами средствами XML DTD.

Несмотря на то, что система объединяет сервер *WebMethods* на узле D&B и набор инструментов D&B на узле *Bowring*, все ее составные части могут взаимодействовать и с другими средствами, поддерживающими XML.

В предшествующие годы для развития объектно-ориентированной гиперсистемы WWW предлагалось использовать технологии *DCOM*, *CORBA* и *Internet Inter-ORB Protocol*. Однако им не удалось справиться с тем, что, оказалось, по силам XML – не только проектировать Web-страницы, но и решать значительно более сложные задачи.

ИСТОЧНИКИ ИНФОРМАЦИИ

1. [Электронный ресурс] Справочник по HTML [http://htmlbook.ru/] на 31.12.2013
2. Лоусон Б., Шарп Р. Изучаем HTML5. Библиотека специалиста, 2-е изд. – СПб.: Питер, 2012, 304 с.: ил.
3. Эрик Рэй. Изучаем XML. Издат-во O'Reilly, 2001, 403 с.: ил.