ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
ГОСУДАРСТВЕННОЕ ОБРАЗОВАТЕЛЬНОЕ
УЧРЕЖДЕНИЕ
ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ
«ВОРОНЕЖСКИЙ ГОСУДАРСТВЕННЫЙ
УНИВЕРСИТЕТ»

А.В. Сычев

WEB-ТЕХНОЛОГИИ

Часть 2

Учебное пособие

Издательско-полиграфический центр
Воронежского государственного университета
2009

ŋqdqɐeⱯ bɆ dedɆbⱯзə d ɕɔɑbɔⱯ Ɉqddɑ̆Ɐqɐè

Глава 6. ВВЕДЕНИЕ В XML

В 1986 году, задолго до того, как идея создания сети Веб была воплощена в жизнь, *универсальный стандартизированный язык разметки* SGML (Standardized Generalized Markup Language) был утвержден в качестве международного стандарта (ISO 8879) определения языков разметки, хотя SGML существовал еще с конца шестидесятых. Он использовался для того, чтобы описывать языки разметки, предоставляя при этом автору возможность давать формальные определения каждому элементу и атрибуту языка.

Язык HTML первоначально был всего лишь одним из SGML-приложений. Он описывал правила, по которым должна быть подготовлена информация для World Wide Web. Таким образом, язык HTML – это набор предписаний SGML, сформулированных в виде определения типа документа (DTD), объясняющих, что именно обозначают тэги и элементы. Схема DTD для языка HTML хранится в веб-браузере.

К недостаткам языка HTML можно отнести следующие:

- HTML имеет *фиксированный набор тэгов*. Нельзя создавать свои тэги, понятные другим пользователям.
- HTML – это исключительно *технология представления данных*. HTML не несет информации о значении содержания, заключенного в тэгах.
- HTML – *«плоский» язык*. Значимость тэгов в нем не определена, поэтому с его помощью нельзя описать иерархию данных.
- *В качестве платформы для приложений используются браузеры*. HTML не обладает достаточной мощью для создания веб-приложений на том уровне, к которому в настоящее время стремятся веб-разработчики. Например, на языке HTML невозможно разработать приложение для профессиональной обработки и поиска документов.
- *Большие объемы трафика сети*. Существующие HTML-документы, используемые как приложения, перегружают Интернет большими объемами трафика в системах клиент-сервер. Примером может служить пересылка по сети большого по объему документа, в то время как необходима только небольшая часть этого документа.

Таким образом, с одной стороны, язык HTML является очень удобным средством разметки документов для использования в веб, а с другой – документ, размеченный в HTML, имеет мало информации о своем содержании. Если тот или иной документ несет достаточно полную информацию о своем содержании, появляется возможность сравнительно легко провести автоматическую обобщенную обработку и поиск в файле, хранящем документ. Язык SGML позволяет сохранять информацию о содержании документа, однако вследствие особой сложности он никогда не использовался так широко, как HTML.

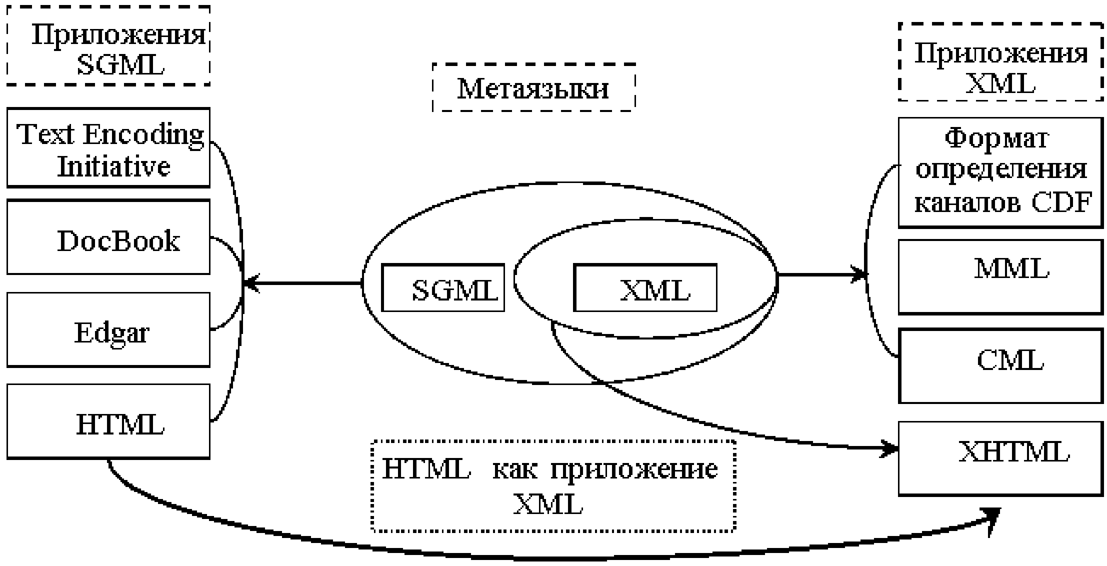

Рис. 6.1. *Диаграмма взаимосвязи между языками разметки*

Важным отличием XML от HTML является то большое внимание, которое уделяется контролю за тем, насколько точно соблюдаются правила языка при разметке документов. В зависимости от этого принято выделять *правильно построенные* и *действительные* XML документы.

Документ XML считается *правильно построенным*, если он соответствует всем синтаксическим правилам XML.

Проверка *действительности* документа предполагает выполнение следующих действий:

- Проверка использования только заданного набора дескрипторов.
- Проверка полного соответствия порядка следования элементов и атрибутов содержанию документа или определенным правилам.
- Контроль типов данных (достигается при использовании соответствующей схемы).
- Контроль целостности данных для обеспечения оптимального обмена информацией через Веб с помощью транзакций.

Рассмотрим теперь основные синтаксические правила построения XML документов.

- XML-документ содержит один и только один *корневой элемент*, содержащий все остальные элементы.
- *Дочерние элементы*, содержащиеся в *корневом элементе*, должны быть правильно вложены.
- *Имена* элементов подчиняются правилам:
  - Имя начинается с буквы, знака подчеркивания или двоеточия.
  - После первого символа в имени могут быть буквы, цифры, знаки переноса, подчеркивания, точка или двоеточие.
  - Имена не могут начинаться с буквосочетания XML.

XML-документ имеет следующую структуру:

- Первая строка XML-документа называется *объявлением* XML. Это необязательная строка, указывающая версию стандарта XML (обычно это 1.0). Также здесь может быть указана кодировка символов и внешние зависимости.
- Комментарий может быть размещен в любом месте дерева. XML комментарии размещаются внутри пары тегов <!-- и заканчиваются -->. Два знака дефис (--) не могут быть применены ни в какой части внутри комментария.
- Остальная часть этого XML-документа состоит из вложенных элементов, некоторые из которых имеют атрибуты и содержимое.
- Элемент обычно состоит из *открывающего* и *закрывающего* тегов, обрамляющих текст и другие элементы.
- *Открывающий* тег состоит из имени элемента в угловых скобках;
- *Закрывающий* тег состоит из того же имени в угловых скобках, но перед именем ещё добавляется косая черта.
- *Содержимым* элемента называется всё, что расположено между открывающим и закрывающим тегами, включая текст и другие (вложенные) элементы.
- Кроме *содержания* у элемента могут быть атрибуты – пары *имя=значение*, добавляемые внутрь открывающего тега после названия элемента.
- Значения атрибутов всегда заключаются в кавычки (одинарные или двойные), одно и то же имя атрибута не может встречаться дважды в одном элементе.
- Не рекомендуется использовать разные типы кавычек для значений атрибутов одного тега.
- Для обозначения *элемента без содержания*, называемого *пустым* элементом, необходимо применять особую форму записи, состоящую из одного тега, в котором *после имени элемента* ставится косая черта «/».

К сожалению, описанные выше правила позволяют контролировать только формальную правильность XML-документа, но не содержательную. Для решения второй задачи используются так называемые *схемы*.

Схема четко определяет *имя* и *структуру* корневого элемента, включая *спецификацию* всех его *дочерних элементов*. Программист может задать, какие элементы и в каком количестве *обязательны*, а какие – *необязательны*. Схема также определяет, какие элементы содержат *атрибуты*, *допустимые значения* этих атрибутов, в т. ч. *значения по умолчанию*.

Чаще всего для описания схемы используются следующие спецификации:

- DTD (*Document Type Definition*) – язык определения типа документов.
- XDR (*XML Data Reduced*) – диалект XML, разработанный Майкрософт.
- XSD (*язык определения схем XML*) – рекомендована консорциумом W3C.

XML документ отличается от HTML документа также и тем, как он отображается в веб-браузере. Без использования *CSS* или *XSL* XML-документ отображается как простой текст в большинстве веб-браузеров. Некоторые веб-браузеры такие, как *Internet Explorer*, *Mozilla* и *Firefox* отображают структуру документа в виде дерева, позволяя сворачивать и разворачивать узлы с помощью нажатия клавиши мыши.

Наиболее распространены три способа преобразования XML-документа в отображаемый пользователю вид:

- Применение стилей *CSS*.
- Применение преобразования *XSLT*.
- Написание на каком-либо языке программирования обработчика XML-документа.

6.2. Языки описания схем XML

Идея создания собственных тэгов, имеющих специальное значение и помогающих описать содержание документа, сама по себе просто замечательна. Но если каждый пользователь может создавать свои собственные описания, каким образом их распознавать? С этой целью в спецификации XML для описания подобных «самодеятельных» тэгов используются *схемы*. Они необходимы для того:

- чтобы описать, что именно является разметкой;
- описать точно, что означает разметка.

Наиболее известными языками описания схем являются следующие:

- DTD (*Document Type Definition*) – язык определения типа документов, который первоначально использовался в качестве языка описания структуры SGML-документа.
- XDR (*XML Data Reduced*) – диалект схемы XML, разработанный Microsoft, который поддерживался в Internet Explorer 4 и 5 версий.
- XML Schema или просто XSD (*язык определения схем XML*) – рекомендация консорциума W3C с 2001 года.

6.3. DTD-схема

Схема DTD предоставляет *шаблон* разметки документа, в котором указываются *наличие*, *порядок следования* и *расположение элементов* и их *атрибутов* в документе XML (рис. 6.2).