XML (eXtensible Markup Language) — «расширяемый язык разметки».

Спецификация XML описывает XML-документы и частично описывает поведение XML-процессоров (программ, читающих XML-документы и обеспечивающих доступ к их содержимому).

XML разрабатывался как язык с простым
формальным синтаксисом, удобный для создания и обработки документов
как программами, так и человеком, с акцентом на использование в Интернете. Язык называется расширяемым, поскольку он не фиксирует разметку, используемую в документах: разработчик волен создать разметку в соответствии с потребностями к конкретной области, будучи ограниченным лишь синтаксическими правилами языка.

Расширение XML — это конкретная грамматика, созданная на базе XML и представленная словарём тегов и их атрибутов, а также набором правил, определяющих, какие атрибуты и элементы могут входить в состав других элементов. Сочетание простого формального синтаксиса, удобства для человека, расширяемости, а также базирование на кодировках Юникод для представления содержания документов привело к широкому использованию как, собственно, XML, так и множества производных специализированных языков на базе XML в самых разнообразных программных средствах.

Физическая и логическая структуры документа

С физической точки зрения документ состоит из сущностей, из которых каждая может ссылаться на другую сущность. Единственный корневой элемент — документная сущность. Содержание сущностей — символы.

С логической точки зрения документ состоит из комментариев,
объявлений, элементов, ссылок на сущности и инструкций обработки. Всё это в документе структуризуется разметкой.

Физическая структура

Сущность — мельчайшая часть в документе. Все сущности что-нибудь содержат, и у всех них есть имя (существуют исключения, напр. документная сущность). Проще говоря, термин «сущность» описывает «сущую вещь», «что-то».

Документ состоит из сущностей, содержание которых — символы. Все символы разделены на два типа: символы данных и символы разметки.

К разметке относятся:

1. **теги (англ. tags)** <- обозначают границы элементов
2. объявления и инструкции обработки, включая их атрибуты (англ. attributes)
3. ссылки на сущности
4. комментарии
5. а также последовательности символов, обрамляющие секции «CDATA»

Часть документа, не принадлежащая разметке, составляет символьные данные документа.

Логическая структура

Все составляющие части документа обобщаются в пролог и корневой элемент. Корневой элемент — обязательная часть документа, составляющая всю его суть (пролог, вообще говоря, может отсутствовать). Корневой элемент может включать (а может не включать) вложенные в него элементы, символьные данные и комментарии. Вложенные в корневой элемент элементы, в свою очередь, могут включать вложенные в них элементы, символьные данные и комментарии, и так далее. Пролог может включать объявления, инструкции обработки, комментарии. Его следует начинать с объявления XML, хотя в определённой ситуации допускается отсутствие этого объявления.

Элементы документа должны быть правильно вложены: любой элемент, начинающийся внутри другого элемента (то есть любой элемент документа, кроме корневого), должен заканчиваться внутри элемента, в котором он начался. Символьные данные могут встречаться внутри элементов как непосредственно так и в специальных секциях «CDATA». Объявления, инструкции обработки и элементы могут иметь связанные с ними атрибуты. Атрибуты используются для связывания с логической единицей текста пар имя-значение.

Символы разметки

Разметка всегда начинается символом `<` и заканчивается символом `>`.

Наряду с символами `<` и `>`, специальную роль для разметки играет также символ `&`. Угловые скобки обозначают границы элементов, инструкций обработки и некоторых других последовательностей. Амперсанд позволяет выполнить замену текста при помощи сущностей.

Решение проблемы неоднозначности разметки

Употребление разметочных символов в символьных данных затрудняет распознавание конструкций разметки и может создать проблему неоднозначности структуры. В XML эта проблема решается следующим образом: <, > и & не могут присутствовать в символьных данных и в значениях атрибутов в их непосредственном виде, для их представления в этих случаях зарезервированы специальные *сущности*:

Символ	Замена
<	<
>	>
&	&

Кроме того, для употребления апострофов и кавычек внутри значений атрибутов используются следующие *сущности*:

'	'
"	"

Правило замены разметочных символов на их обозначающие *сущности* не распространяется на символьные данные в секциях «CDATA», зато выполняется во **всех** остальных местах документа.

Числовые ссылки на символы

Числовые ссылки на символы указывают кодовую позицию символа в наборе символов документа. Числовые ссылки на символы могут принимать две формы[7]:

1. синтаксис «**&#D;**», где D — десятичное число;
2. синтаксис «**&#xH;**» или «**&#XH;**», где H — шестнадцатеричное число (шестнадцатеричные числа в числовых символьных ссылках не чувствительны к регистру).

Примеры числовых ссылок на символы:

- **å** — (в десятичной форме) представляет букву «а» с маленьким кружком над ней (используется, например, в норвежском языке);
- **å** — (в шестнадцатеричном) представляет собой тот же символ;
- **å** — (в шестнадцатеричном) также представляет тот же символ;
- **И** — (в десятичной форме) представляет заглавную букву кириллицы «I»;
- **水** — (в шестнадцатеричном) представляет китайский иероглиф «вода»;

Имена

В языке XML все имена должны начинаться с буквы, символа подчёркивания (_) и продолжаться только допустимыми для имён символами, а именно: они могут содержать только буквы, входящие в секцию букв кодировки Unicode, арабские цифры, дефисы, знаки подчёркивания, точки. Так как буквы не ограничены исключительно символами ASCII, то в именах можно использовать буквы из любого языка.

Объявление XML

Объявление XML указывает версию языка, на которой написан документ. Поскольку интерпретация содержимого документа зависит от версии языка, то Спецификация предписывает начинать документ с объявления XML. В первой (1.0) версии языка использование объявления не было обязательным, в последующих версиях оно обязательно. Таким образом, версия языка определяется из объявления, и если объявление отсутствует, то принимается версия 1.0.