Введение в мультимедийные базы данных

7. XML и XML-базы данных

- 7.1. XML
- 7.2. XML-термины
- 7.3. Хранение XML в СУБД

XML

- XML = e**X**tensible **M**arkup **L**anguage (расширяемый язык разметки)
- Поддерживается W3C (World Wide Web Consortium); первая рекомендация (описание) – 1998 год
- Подмножество SGML (**S**tandard **G**eneralized **M**arkup **L**anguage): упрощенная версия SGML
- XML - метаязык
- Документ – дерево элементов
- Элемент состоит из открывающего и закрывающего тегов
- Описание элемента может быть дополнено атрибутами, помещаемыми в открывающий тег
- Элемент может ссылаться или описывать мультимедийный объект
- Отличия от HTML (также базирующегося на SGML):
  - HTML - для описания внешнего представления документа ; XML – для описания структуры и семантики документа
  - Расширяемый: можно задавать свои собственные теги
  - HTML - язык для публикации в Веб; XML - для более широкого применения

XML

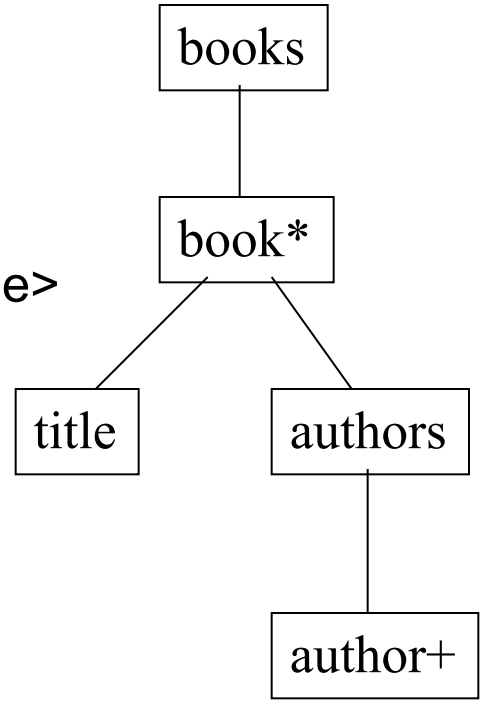
- Термин “XML” иногда используется неправильно:

  Ошибочные представления о XML:
  - XML – не язык программирования (но язык программирования можно описать с помощью XML-разметки)
  - XML – не протокол передачи данных (но типичная задача XML описывать структуру документов/данных, передаваемых по компьютерным сетям)
  - XML – не структура базы данных (но XML может храниться в бд и можно выполнять различные запросы к XML-данным)

XML

Пример:

```
<?xml version=”1.0”?>
<books>
   <book isbn=”1558604669">
    <title>Principles of Multimedia Database Systems</title>
    <authors>
      <author>Subrahmanian</author>
    </authors>
   </book>
   <book isbn=”1558603123”>
    <title>Multimedia and Imaging Databases</title>
    <authors>
      <author>Khoshafian</author>
      <author>Baker</author>
    </authors>
   </book>
</books>
```



XML

Языки разметки базируемые на XML:

- Wireless Markup Language (WML): формат данных для (беспроводных) устройств, работающих с протоколом WAP (мобильные телефоны)
- Synchronized Multimedia Integration Language (SMIL):
  - Задает временную разметку, внешнее вид и т.д. для мультимедийных презентаций
  - Определяет порядок воспроизведения мультимедийных файлов
  - Для просмотра требуется SMIL-совместимый плеер (AMBULANT, MS IE6)
  - Руководство и примеры: http://www.multimedia4everyone.com/
- Scalable Vector Graphics (SVG): для описания двухмерной векторной графики
- Mathematical Markup Language (MathML): для описания математических обозначений (формул)
- Chemical Markup Language (CML): для представления химических формул
- ... множество других

XML-термины

- DTD = Document Type Definition:
  - Определяет структуру документа
  - Коллекция похожих документов обычно имеет одинаковый DTD
  - Задает описание документа: какие элементы, какие атрибуты, в каком порядке, количестве и т.д.

- XML Schema:
  - DTD с более широкими возможностями
  - Замена DTD

- Сравнение XML Schema и DTD:
  - DTD описывается не на XML (т.е. с помощью своего не XML формата)
  - XML Schema позволяет создавать более расширяемые и гибкие описания
  - XML Schema позволяет определять ограничения на тип данных

XML-термины

Пространства имен (namespaces):

- Позволяют избегать многозначных толкований элементов и атрибутов (с одинаковыми именами)
- Группируют понятия, относящиеся к одному и тому же приложению (объекту, понятию и т.д.)
- Используются уникальные идентификаторы – определяющие пространства имен
- Элемент или атрибут однозначно идентифицируется по своему имени плюс по пространству имен к которому элемент/атрибут относится

XML-термины

XSL = E**X**tensible **S**tylesheet **L**anguage (расширяемый язык таблиц стилей):

- Преобразование XML-документа
- Изменение структуры XML-документов; например, для представления одних и тех же XML-данных на различных носителях (экране, бумаге, мобильном телефоне)
- XML-документ после преобразования может оказаться не XML-документом (например, документом в формате PDF)
- XSL-спецификация состоит из двух частей:
  - XSL Transformations (XSLT): реорганизация структуры и содержимого
  - XSL Formatting Objects (XSL-FO): визуальное представление

XML-термины

XPath:

- Язык для адресации определенных частей в XML-документах
- Используется в XSLT, XPointer, XQuery

CSS - каскадные таблицы стилей (Cascading Style Sheets):

- Применяется в HTML
- Более ограниченные возможности по сравнению с XSL-FO

XLink:

- Ссылки между документами
- Определяет действия, связанные с ссылками (например, как отображать документ по ссылке)

XPointer:

- Расширение XPath, используемое XLink, для указания на любой фрагмент в другом документе

XHTML:

- HTML, переформулированный на XML
- Поддерживается W3C с 2000 года

Программные средства для работы с XML

XML-Парсер:
- Проверки структуры документов и типов данных, задаваемых DTD/XML Schema
- Пример: Xerces

XSLT-процессор:
- Преобразование XML-документа в другой тип документа (XML, HTML, текстовый и т.д.)
- Пример: Xalan

Процессор форматирования (formatting objects processor):
- Основан на XSL-FO
- Результат форматирования: PDF, PCL, PS, SVG и ряд других
- Пример: Apache FOP

XML-редактор:
- Множество разных
- Создание, редактирование XML-документов, DTD, XML Schema и т.д.
- Пример: XML Spy

XML-браузер:
- Преобразование XML в HTML
- Реализовано во всех основных - MSIE, Firefox, Opera

Программные интерфейсы (API) для XML

- XML-документы – обычные текстовые файлы; в принципе можно обойтись без API
- Тем не менее, в почти каждом приложении, работающим с XML-данными, используются стандартные операции для доступа к XML-данным
- Document Object Model (DOM):
  - Рассматривает документы как объекты
  - Строит древовидную структуру документа в памяти
  - Предоставляет методы для движения по дереву и манипуляцией с узлами дерева
  - Также применима к грамматически правильным HTML-документам
  - Спецификация от W3C
- Simple API for XML (SAX):
  - Две версии: SAX1 и SAX2
  - Основана на модели событий (под событием понимается начало тега, конец тега и т.д.)
  - Элементы обрабатываются в том порядке в каком встречаются
- Streaming API for XML (StAX):
  - Лишено ряда недостатков DOM и SAX

Хранение XML в СУБД

Два способа хранить XML-данные в базе данных:

- Преобразование схемы (схем) XML-документов в схему базы данных:
  - Использование своей собственной модели данных – реляционной, иерархической, объектно-ориентированной
  - Например, хранение документа с описанием заказа (в XML) в реляционной база данных – данные распределяются в реляционной бд по таблицам «Заказы», «Наименования», «Клиенты» и т.д.
  - База данных, поддерживающая этот способ, называется базой данных с XML-поддержкой (XML-enabled); XML-функциональность может быть добавлена к бд с помощью стороннего программного обеспечения
  - Единица хранения (в случае реляционной модели данных): запись (строка в таблице)

Хранение XML в СУБД

Два способа хранить XML-данные в базе данных:

- Использование определенного набора структур, позволяющего хранить любой XML-документ:
  - XML-модель данных
  - Используется набор таблиц, позволяющий хранить произвольные XML-документы (Элементы, Атрибуты, Текст, и т.д.)
  - Хранение документов с произвольной схемой или даже с неизвестной (отсутствующей) схемой
  - База данных, поддерживающая этот способ, - нативная (или прирожденная) XML-база данных (native XML database)
  - Единица хранения: XML-документ

Хранение XML в СУБД

База данных для хранения данных или документов?

- **Дата-ориентированные** документы (data-centric):
  - Документы, использующие XML для транспортировки данных
  - Преимущественно для машинной обработки
  - Примеры: торговые заказы, расписания рейсов, биржевые индексы, научные данные (не всегда), многие контентно-насыщенные документы (вроде страницы с описанием книги на Ozon.ru)
  - Регулярная структура
  - Порядок элементов обычно не имеет значения
  - Многие контентно-насыщенные документы
- **Документо-ориентированные** документы (document-centric):
  - Документы с которыми работают люди
  - Примеры: книги, сообщения электронной почты, реклама, почти все создаваемые вручную XHTML-документы
  - Менее регулярная или иррегулярная структура
  - Порядок элементов в большинстве случаев важен

На практике, разделение между дата-ориентированными и документо-ориентированными документами не всегда однозначно

Тем не менее, разделение важно:

- Данные: хранить в базе данных с XML-поддержкой
- Документы: в нативной XML-базе данных

Хранение XML в СУБД

Нативная XML база данных:

- Формальное техническое определение отсутствует
- Задает (логическую) модель XML-документа; как минимум, модель должна включать в себя элементы, атрибуты, секции PCDATA, и порядок документа; примеры моделей – модель данных XPath, модели, основанные на DOM и событиях в SAX 1.0
- Фундаментальная единица (логического) хранения - XML-документ (в то время как в реляционной базе данных – запись)
- Использование определенной физической структуры хранения не обязательно; например, нативная XML-бд может строиться на основе реляционной, иерархической или объектно-ориентированной базе данных или использовать собственный формат хранения, например, в виде индексированных и заархивированных файлов

Хранение XML в СУБД

Нативная XML-база данных или традиционная (в первую очередь, подразумеваем, реляционную) база данных?

- Открытый вопрос
- Реляционная модель данных (начало 1970-х годов) – устоявшаяся технология: множество решений, методик, продуктов, множество специалистов, ...; солидный математический и научный базис
- XML-модель данных (стандарт с 1998г) – первые шаги ...
- Причины не использовать нативную XML-бд: слабые гарантии производительности при больших и очень больших объемах документов; зачаточные возможности индексирования (над совершенствованием сейчас ведется активная работа)
- Рекомендация на данный момент: при выборе реляционная бд с (или даже без) поддержкой XML должна иметь приоритет, но**(!)** активно использовать XML как формат описания данных
- Тем не менее, для ряда приложений нативная XML-бд хороший вариант: в частности, когда требуется интенсивное выполнение запросов к XML-данным

Хранение XML в СУБД

Пример нативной XML-базы данных:

Sedna[1]**:**

- Открытый исходный код
- Разработана «с нуля» (не на основе какой-то бд)
- Поддержка XQuery
- Поддержка ACID транзакций
- Безопасность (пользователи, роли, привилегии)
- Индексы по структуре и по значениям
- API к нескольким языкам программирования

---

[1] http://modis.ispras.ru/sedna/

XQuery

- Язык запросов XML
- Разработан в W3C; первая версия - XQuery 1.0 в 2003г.
- Надмножество XPath
- Совместим с другими XML-стандартами
- Изначально предназначен для извлечения информации и не включал средств для модификации существующих документов XML
- XQuery аналог SQL для баз данных
- XQuery поддерживается тремя главными производителями бд (IBM, Oracle, Microsoft), а также многими другими бд

Ссылки на литературу

[1] Aiken and Allen. XML in Data Management: Understanding and Applying Them Together (The Morgan Kaufmann Series in Data Management Systems). Morgan Kaufmann, 2004

[2] R. Bourret. XML and Databases. http://www.rpbourret.com/xml/XMLAndDatabases.htm, Сентябрь 2005, *есть перевод на русский*