СТАНДАРТЫ ПЛАТФОРМЫ XML И БАЗЫ ДАННЫХ [*]

М.Р. Когаловский
Институт проблем рынка РАН
Москва, 117418, Нахимовский проспект, 47
e-mail: kogalov@cemi.rssi.ru

Абстракт

В докладе обсуждаются предпосылки создания новой формирующейся технологической платформы Web, базирующейся на языке XML, которая стала основой второго поколения этой глобальной информационной системы. Рассматриваются существо происходящих в Web радикальных перемен, организация XML-платформы, принципы обеспечения расширяемости языка XML и функциональных возможностей платформы, синтаксического единства составляющих ее стандартов. Обсуждаются назначение, классификация, взаимосвязи и состояние разработки основных стандартов XML-платформы. Анализируются подходы к представлению метаданных и описанию семантики XML-документов, предусмотренные для этого средства. Показаны основные сферы применения стандартов платформы XML, в частности, в базовых стандартах других широко распространенных информационных технологий, а также в разработках электронных библиотек. Особое внимание уделяется проблемам интеграции технологий XML и баз данных. Оцениваются перспективы XML-платформы.

Предпосылки создания новой технологической платформы Web

Создание World Wide Web стало одним из крупнейших научно-технических достижений последнего десятилетия XX века, основой целого ряда новых информационных технологий, имеющих весьма значимые социально-экономические последствия.

Идеи проекта, возникшего в стенах CERN (Европейский центр ядерных исследований, Женева) в конце 80-х годов, в короткие сроки воплотились в беспрецедентно интенсивно развивающуюся глобальную открытую бесконечно масштабируемую распределенную гипермедийную систему с прозрачными для пользователя распределением и неоднородностью ресурсов. Количество пользователей и объем представленных в ней информационных ресурсов продолжают чрезвычайно быстро наращиваться. При этом возможен свободный доступ к большинству информационных ресурсов Web в любой момент времени.

Вместе с тем, за несколько лет интенсивного развития потенциал качественного совершенствования технологий существующей версии Web (называемой далее Web-1) оказался в значительной мере исчерпанным. Сдерживающее влияние на дальнейшую эволюцию приложений Web-технологий стали оказывать, прежде всего, слабые стороны языка HTML - основного выразительного и структурообразующего средства представленных в Web гипермедийных информационных ресурсов, а также ограниченные функциональные возможности среды поддержки этого языка в Web. Эти слабые стороны и ограничения заключаются, главным образом, в следующем:

- Для HTML-документов не поддерживаются метаданные, которые бы описывали структурные, семантические и другие их свойства. Исключением являются введенные в HTML средства, позволяющие

[*] *Работа частично поддержана грантами РФФИ 01-07-90444 и РГНФ 00-02-12002.*

ассоциировать с ними ключевые слова или рубрики. Эти простейшие средства могут использоваться для неформального описания семантики документов. В значительной мере указанное ограничение является следствием ориентированности языка HTML не на структурную разметку документов, а на описание формата их представления на экране компьютера.

- HTML является закрытым языком, не позволяющим пользователю дополнять при необходимости новые теги для расширения функциональности языка.
- Информационные ресурсы HTML могут идентифицироваться только по их местоположению в распределенной среде Web (с помощью URL).

Отсутствие поддержки метаданных для HTML-документов не позволяет верифицировать целостность их структуры и содержания. По этой же причине невозможно осуществлять эффективный целенаправленный поиск необходимой пользователю информации в огромном накопленном объеме информационных ресурсов Web и выполнять другие операции обработки информационных ресурсов. Удовлетворение информационных потребностей пользователей с помощью навигационного поиска во многих случаях является крайне неэффективным или просто невозможным. Созданные для решения этой проблемы поисковые сервисы Web реализуют только технику контекстного поиска. Поиск документов в Web с учетом свойств их структурных компонентов является невозможным. Довольно высок уровень информационного шума.

Наконец, без поддержки метаданных в среде Web невозможна эффективная интеграция информационных ресурсов, поддерживаемых в этой среде и в других взаимодействующих с Web средах. Технически средства языка HTML позволяют, интегрировать в среду Web ресурсы баз данных, большие архивы текстовых документов, различные мультимедийные ресурсы. Но эти инородные для гипертекста ресурсы, хотя и становятся доступными пользователю, остаются, с точки зрения их семантики, для среды Web "черным ящиком". Такая интеграция сводится по существу лишь к обеспечению доступа к “внешним” ресурсам посредством Web.

Следствием закрытости языка HTML являются ограниченные возможности структурирования документов, адекватного потребностям пользователей и отражающего хотя бы простейшие аспекты семантики содержащихся в них данных. Закрытый характер языка приводит также к необходимости периодического пересмотра версий стандарта HTML для расширения его функциональности путем добавления новых тегов или атрибутов тегов.

Настоятельно необходимыми стали такие шаги в дальнейшем развитии информационной среды Web, которые позволили бы создать новые технологии, основанные на современных методах управления данными, прошедших испытание временем в технологиях баз данных и текстовых поисковых систем. Основу этих методов составляют модели данных, адекватные требованиям, предъявляемым к информационным ресурсам и к характеру их использования, явное представление и поддержка метаданных в системе, использование техники поиска документов на основе их содержания.

Решение указанных выше проблем стало важнейшей задачей развития Web-технологий.

Существо радикальных перемен в Web и используемые подходы

В последние годы консорциум W3C ведет активную деятельность, направленную на радикальный пересмотр основ Web-технологий и затрагивающую все три базовых элемента первоначального проекта WWW, на которых построена действующая сегодня его реализация (язык гипертекстовой разметки HTML, универсальный локатор ресурсов URL, протокол передачи гипертекстовых ресурсов HTTP).

Создано ядро и продолжается процесс формирования независимого от области приложений комплекса средств, базирующегося на расширяемом языке разметки XML [1-3] и служащего для описания и обработки информационных ресурсов Web, который мы называем здесь XML-платформой. Этот комплекс призван стать основой нового поколения Web, называемого далее для краткости Web-2. В указанном комплексе предусматривается использование более общего по сравнению с URL механизма идентификации информационных ресурсов - URI (Universal Resource Identifier) [4]. Кроме того, для него разрабатывается новый протокол обмена XML-ресурсами [5].

Новые технологии Web базируются на открытом для расширения концептуально и в большой части синтаксически едином комплексе стандартов, которые составляют XML-платформу и определяют многоаспектные функциональные возможности для представления информационных ресурсов Web и доступа к ним.

В разработке XML-платформы важное место занимает создание стандартов представления метаданных, описывающих структурные и семантические свойства XML-ресурсов, что позволяет вести речь о “семантическом Web”. Благодаря введению поддерживаемых в явном виде метаданных и стандартизации средств их описания открылись возможности для синтаксической и семантической интеграции информационных ресурсов XML и поддерживаемых средствами других информационных технологий. В последние годы в этой области интенсивно проводятся исследования.

Одна из принципиальных установок рассматриваемой деятельности состоит в непременном обеспечении преемственности новой платформы с Web-1, что позволит сохранить возможность использования и в дальнейшем огромных информационных ресурсов, представленных средствами языка HTML.

Платформа XML

Наряду с созданием стандарта языка XML консорциум W3C, формирующий техническую политику развития Web и разрабатывающий стандартизованные спецификации для этой среды, на самом деле одновременно формирует новую открытую для расширения функциональности технологическую платформу, главным звеном которой является XML. Вероятно, не замечая этого обстоятельства, в некоторых публикациях языку XML ошибочно приписываются функциональные возможности, которые на самом деле обеспечиваются различными другими стандартами XML-платформы.

В отличие от Web-1, где все основные функции управления информационными ресурсами системы базируются на едином языке HTML, создатели XML-платформы избрали иной путь. Выделены “фундаментальные” стандарты, составляющие концептуальную и синтаксическую основу платформы. Их средствами определяется комплекс других стандартов, каждый из которых выполняет собственные специфические функции. И этот комплекс открыт для пополнения его в случае необходимости новыми стандартами. Именно такая “модульность” организации платформы обеспечивает ее открытый характер, возможности введения новых стандартов, не затрагивая уже существующих. Полная функциональность этой платформы определяется целым комплексом взаимосвязанных стандартов, часть из которых уже принята W3C, другие находятся в стадии разработки.

Функциональные возможности XML-платформы показывает приведенная ниже классификация составляющих ее стандартов (ниже приводятся только принятые стандарты и проекты стандартов, над которыми активно ведется работа):

- *Фундаментальные*: InfoSet, Namespace, XML

- *Структурообразующие*: XPointer, XLink
- *Форматирование и трансформация XML-документов*: XSL, XSLT, CSS
- *Представление метаданных*: XML DTD, XML Schema, RDF
- *Запросы*: XQuery
- *Интефейс прикладного программирования:* DOM
- *Преемственность с Web-1*: XHTML, XML Base
- *Транспорт данных*: XML-Protocol, XForm
- *Идентификация информационных ресурсов*: URI, URL, URN
- *Безопасность*: XML Signature
- *Вспомогательные*: XInclude, XFragment, XML Canonical, XPath
- *Вертикальная сфера*: MathML.

Рассмотрим кратко назначение перечисленных стандартов. Сведения о состоянии их разработки можно найти в наименованиях представляющих эти стандарты документов, на которые мы ссылаемся далее по тексту.

Прежде всего, о роли языка XML. В составе стандартов рассматриваемой платформы он выполняет две важные функции. Прежде всего, он обеспечивает содержательную (структурную) разметку информационных ресурсов, которые называют в рассматриваемой среде XML-документами, а также предоставляет средства (некоторый подъязык XML) для описания общей структуры документов интересующего пользователя типа. Такое описание называется Document Type Definition (DTD). Вместе с тем, как показывает приведенная классификация, язык XML служит одним из фундаментальных стандартов платформы XML. Другие стандарты платформы, которые дополняют его функции, связанные с управлением данными Web, определяются в терминах синтаксиса XML. В связи с этим их называют иногда приложениями XML.

Возвращаясь к выполняемой XML функции разметки, следует еще раз подчеркнуть, что он (в отличие от HTML) не является полнофункциональным языком, который должен решать все задачи представления, поддержки и обработки информационных ресурсов Web. Если проводить аналогию с технологиями баз данных, то XML можно квалифицировать как язык определения данных. Специфика XML как языка определения данных заключается в том, что в нем сочетаются возможности описания свойств экземпляров элементов XML-документов, составляющих содержание данного конкретного документа, с возможностями определения свойств типа XML-документов (DTD) в терминах типов элементов этих документов. Первая группа средств (теги разметки) используется по принципу самоописываемости, определяя некоторые свойства элементов конкретного документа с помощью встраиваемых в него тегов разметки. Что касается DTD, то оно описывает типовые свойства элементов документа и свойства типов документов в целом. Роль DTD аналогична роли схемы базы данных. При этом DTD отчуждается от описываемых документов и хранится где-либо в Web. Конкретные XML-документы ссылаются на это определение, хотя они могут и включать его непосредственно в явном виде.

Для определения других стандартов платформы служат наряду с XML также стандарты XML Information Set (InfoSet) [6] и Namespaces in XML (Namespace) [7]. Первый из них представляет абстрактный набор данных, используемых в XML-документах, содержит их определения, необходимые для спецификаций стандартов, имеющих дело с правильно построенными XML-документами. Можно сказать, что это своего рода онтологическое описание среды XML-документов для группы стандартов платформы

XML, их концептуальная основа. Стандарт Namespace определяет для заданного XML-документа или множества документов допустимые теги разметки и их атрибуты, ассоциируя с ними по умолчанию некоторую семантику. Зарезервированные W3C пространства имен используются в синтаксисе языка XML и других стандартов платформы.

Структурообразующие функции в среде информационных ресурсов Web-2 выполняют языки XPointer [8] и XLink [9], которые предусматривают значительно более богатые возможности по сравнению с HTML для определения гиперсвязей между XML-документами и/или их фрагментами, а также указателей на франгменты XML-документов.

Средства для форматной разметки XML-документов определяют стандарты каскадных таблиц стилей CSS [10] и расширяемого языка таблиц стилей XSL [11]. Заметим, что стандарт CSS используется и как дополнительный к HTML инструмент разметки страниц HTML. Вторая часть стандарта XSL, называемая XSLT [12], позволяет описывать форматные преобразования (трансформации) XML-документов.

Важное место в составе платформы XML занимают стандарты представления метаданных XML Schema [13-15] и RDF [16-17], которые позволяют описывать дополнительные (по отношению к DTD) синтаксические свойства XML-документов, а также их семантику.

Группа рабочих проектов W3C определяет активно разрабатываемый со второй половины 2000 г. стандарт языка запросов XQuery для платформы XML. Указанные документы описывают требования к разрабатываемому языку запросов [18], модель данных [19-20], на которой он базируется, примеры, иллюстрирующие его функциональные возможности [21], а также спецификации синтаксиса XQuery в BNF [22] и в XML [23].

Стандарт DOM [24] объектной модели XML- и HTML-документов определяет функции интерфейса прикладного программирования для их обработки.

Особое место в рассматриваемом комплексе стандартов занимает недавно принятый W3C стандарт XHTML 1.0 [25]. Он обеспечивает один из возможных путей сохранения преемственности развития среды Web, позволяя использовать на платформе XML информационные ресурсы, накопленные в рамках технологий HTML. Этот стандарт поддерживает средствами XML функциональность текущей версии языка HTML (HTML 4.01) на трех различных уровнях, различающихся степенью полноты ее поддержки. Следует упомянуть здесь также стандарт XML Base [26], который служит для поддержки средствами стандарта XLink некоторых видов гиперссылок, используемых в языке HTML.

Разрабатываемый XML-протокол [5] предназначен для стандартизации процедур обмена XML-данными в среде Web-2. К числу стандартов транспорта данных можно отнести также XForms [27] усовершенствованный и адаптированный к среде XML аналог механизма форм в языке HTML, обеспечивающий передачу данных, например запросов, от Web-клиента к Web-серверу.

В стандартах XML-платформы предусматривается возможность использования более общего по сравнению с URL вида идентификаторов ресурсов – Universal Resource Identifier [4]. Привычный для Web-1 способ идентификации с помощью URL, а также абстрактные имена ресурсов URN, являются частными случаями URI.

Предусматриваются средства обеспечения безопасности передачи XML-документов. Эту задачу решает разрабатываемый стандарт электронной подписи XML-Signature [28].

Комплекс стандартов платформы XML включает также целый ряд вспомогательных стандартов.

Стандарт XPath [29] определяет понятие фрагмента XML-документа, используемое в языках XPointer, XSLT, XQuery и в разработке новой версии DOM. В стандарте XML Inclusions (XInclude) [30] представлены модель и синтаксис для описания слияния XML-документов. Стандарт XML Fragment Interchange [31] позволяет описывать контекст фрагментов XML-документа и благодаря этому просматривать и редактировать их вне полного текста документа. К рассматриваемой группе относится также стандарт Canonical XML [32], который определяет метод, позволяющий устанавливать эквивалентность двух XML-документов с различным синтаксическим представлением. Эта возможность существенна, в частности, для стандарта цифровой подписи [28].

Отметим, наконец, что наряду с разработкой "горизонтальных" компонентов комплекс стандартов XML включает также и "вертикальные" компоненты. Первым из них является математический язык разметки [33].

Обеспечение расширяемости языка XML и XML-платформы

Принципиально важным свойством языка XML, обеспечивающим новые функциональные возможности среды Web, является его расширяемость. Достижение расширяемости XML основано на двух факторах. Прежде всего, он представляет собой язык метауровня, подмножество известного языка SGML [34], а не конкретный язык, подобный HTML. Благодаря этому XML выполняет функции языка определения данных. Используя его синтаксис, можно определять различные типы элементов, экземпляры которых образуют содержание конкретных XML-документов, и вводить тем самым адекватный потребностям набор тегов разметки документов. Второй фактор – это использование пространств имен - именованных множеств символов, используемых в качестве имен типов элементов и атрибутов элементов XML-документов. Пространство имен позволяет также явным или неявным образом ассоциировать нужную семантику с именуемыми элементами документов, их атрибутами и допустимыми ддя них значениями.

Таким образом, отдельные пользователи или сообщества пользователей могут порождать нужные им языки разметки для различных категорий документов с нужной семантикой тегов разметки. Поскольку в одном XML-документе можно ссылаться на разные пространства имен, допускается использование для его разметки комбинаций различных порожденных XML языков.

Важно подчеркнуть, что рассмотренные принципы обеспечивают также расширяемость функциональных возможностей всей XML-платформы. Однако для этой цели необходимо достижение консенсуса на уровне консорциума W3C. Основу каждого дополняющего XML стандарта платформы составляет некоторый набор таких типов элементов XML-документов, синтаксис которых может быть определен средствами XML и которые поддерживают требуемые новые функциональные возможности. Помимо этого вводится пространство имен с зарезервированным именем, включающее имена новых типов элементов XML-документов и их атрибутов. Семантика элементов этих типов, их атрибутов и значений, которые они могут принимать, определяются в спецификациях данного дополнительного стандарта.

Примером использования рассмотренного механизма расширения XML-платформы может служить стандарт XLink [9], который позволяет использовать в XML-документах специального вида связующие элементы, обеспечивающие различного рода связи между XML-документами и/или их фрагментами. В самом языке XML концепция гиперссылки не поддерживается.

Важно здесь подчеркнуть, что стандарты платформы XML синтаксически однородны. Все дополняющие XML стандарты используют синтаксис этого языка. Именно в силу этого они

квалифицируются авторами как приложения XML. Указанное обстоятельство имеет существенное значение, поскольку информационные ресурсы, представленные в среде XML с расширенной средствами тех или иных стандартов платформы функциональностью, остаются XML-документами и могут обрабатываться и транспортироваться в поддерживающей XML среде как “чистые” XML-документы.

Моделелирование данных XML

Хотя понятие модели данных упоминалось в прошлые годы в спецификациях ряда стандартов XML-платформы, проблемы моделирования данных не были здесь основательно проработаны. Единой функционально полной, охватывающей как структурные, так и операционные возможности, специфицированной в явном виде модели данных, на которой бы базировались все стандарты платформы, не существует до сих пор. Однако, с нашей точки зрения, она чрезвычайно необходима. Только на ее основе может быть обеспечено функциональное единство платформы наряду с синтаксическим единством ее стандартов, обеспечиваемым языком XML.

В настоящее время ситуация с моделированием данных в рассматриваемой области такова. Вопросы моделирования данных обсуждаются лишь автономно в рамках спецификаций некоторых стандартов. При этом авторы имеют в виду только структурные аспекты моделирования данных. Исключение составляет стандарт DOM [24], определяющий API для репозиториев XML- и HTML-документов. Заметим, что хотя DOM может применяться к XML-данным, он не является стандартом XML-платформы (приложением XML).

Такая ситуация, похоже, начинает изменяться в связи с активизацией работ над языком запросов платформы. Возможно, причина заключается в активной роли, которую играет в коллективе разработчиков Д. Чамберлин (IBM), один из создателей реляционного языка запросов SQL В рамках проекта языка запросов XQuery опубликовано несколько документов. Среди них документы [19, 20] посвящены спецификации модели данных. Судя по наименованиям этих документов, авторы полагают, однако, что к модели данных имеет отношение лишь первый из этих документов, с чем нельзя согласиться.

Более детально вопросы моделирования данных в стандартах XML-платформы обсуждаются в нашей работе [35].

Метаданные и семантика XML-документов

Одной из важнейших целей создания платформы XML является привнесение в среду Web метаданных, описывающих свойства поддерживаемых в ней информационных ресурсов, прежде всего, структуры XML-документов и их смыслового содержания (семантики). Благодаря этому обеспечиваются возможности автоматической проверки правильности структуры XML-документов и снижения уровня информационного шума при поиске информационных ресурсов в Web с помощью различных поисковых машин. Явное описание семантики XML-документов необходимо также для разнообразных продвинутых Web-приложений. В частности, становится возможным создание принципиально новых приложений высокого уровня, основанных на интеграции информационных технологий и обеспечивающих интеграцию неоднородных информационных ресурсов. Это направление активно развивается во многих научных центрах разных стран и связано с созданием информационных систем нового класса, функционирующих в среде Web и называемых электронными библиотеками [36].

В стандартах платформы XML предусмотрено несколько средств определения метаданных. Для определения структуры XML-документов специальные синтаксические конструкции предусмотрены в

языке XML. Представленные их средствами метаданные называются определением типа документов (Document Type Definition, DTD). В DTD XML-документы данного типа описываются как иерархические структуры, состоящие из их элементов. Это описание может быть встроено в XML-документ или оно хранится где-либо в Web, и в документе дается на него ссылка. Для более утонченного описания структуры XML-документов могут использоваться средства стандарта XML Schema [14-15]. По сравнению с DTD, этот стандарт предоставляет для описания XML-документов дополнительные возможности, в частности более развитую систему типов значений атрибутов элементов.

Семантика XML-документа может быть определена явным или неявным образом (по умолчанию). Явное определение может быть формализовано в различной степени. Простейший способ задания семантики – использование пространства имен. Механизм пространства имен может, как уже отмечалось, определять явным или неявным образом семантику элементов XML-документов различных типов, их атрибутов, а также принимаемых атрибутами значений.

В последнее время начали создаваться сервисы регистрации и поддержки пространств имен в интересах различных сообществ разработчиков и пользователей. Зарегистрированное пространство имен становится своего рода стандартом для сообщества пользователей. В качестве такого согласованного пространства имен может использоваться, например, набор элементов метаданных Дублинского ядра (Dublin Core, DC). Его поддержкой и развитием занимается учрежденная для этих целей организация - Dublin Core Metadata Initiative (DCMI). Текущая версия спецификаций Дублинского ядра - DC 1.1 [37] была принята в июле 1999 г. Она включает 15 элементов метаданных. DCMI опубликовал также спецификации рекомендованных квалификаторов [38], уточняющих смысл элементов метаданных DC и интерпретацию их значений В настоящее время на основе DC 1.1 ведется разработка официального стандарта ANSI/NISO Z39.85 [39].

Более формализованный способ явного описания семантики XML-документов обеспечивается средствами стандарта W3C - Resource Definition Framework (RDF) [16, 17]. Такое описание, называемое RDF-спецификацией, аналогично по своим возможностям концептуальной схеме в системах баз данных. По сравнению с рассмотренными выше средствами, оно представляет собой более высокий уровень семантического описания информационных ресурсов, приблизительно эквивалентный ER-модели.

В RDF-спецификации объявляется некоторое множество ресурсов, для каждого из которых определяются пары "свойство-значение". Информационные ресурсы в RDF - это ресурсы Web, идентифицируемые уникальным образом с помощью их URI. Они могут также представлять собой коллекции других информационных ресурсов или литералов, называемые контейнерами. Допускаются контейнеры типа мультимножества, последовательности и альтернативы. Значения свойств задаются литерально либо ссылками на другие ресурсы, которые представляются, в свою очередь, их свойствами. Таким образом, свойства могут определять и связи между ресурсами. Описание семантики свойств называется схемой. В стандарте RDF не регламентируется способ задания схемы для RDF-спецификации. Достаточно лишь представить ее как некоторый ресурс в WWW, и использовать URI этого ресурса для ссылки на нее в RDF-спецификации. В документации стандарта RDF рассматривается, например, вариант использования для этих целей упоминавшегося выше Дублинского ядра. Во второй части стандарта, называемой Schema Specification [17], предлагается значительно более богатый способ задания схемы. Этот способ основан на объектной модели, в которой используются концепции классов, свойств и ограничений, ассоциируемых с классами и свойствами, поддерживается иерархическое отношение "класс-подкласс".

Заметим, что для приложений, нуждающихся в более формальном описании семантики данных, схема в RDF-спецификациях является той “открытой точкой”, которая позволяет интегрировать в среду XML онтологические спецификации предметной области или иные описания семантических свойств информационных ресурсов на уровне систем представления знаний.

В настоящее время уже создано значительное количество свободно распространяемых и коммерческих инструментальных средств для поддержки RDF-спецификаций [40] - синтаксических анализаторов, программного обеспечения репозиториев, реализаций языков запросов RDF и т.д.

Сферы применения стандартов XML

Хотя язык XML и базирующаяся на нем платформа стандартов W3C создавались как средство представления информационных ресурсов Web, они, тем не менее, уже находят значительно более широкие применения в различных областях информационных технологий. Это обстоятельство, по нашему мнению, обусловлено прежде всего развитыми возможностями средств платформы для представления информационных ресурсов, их адаптируемостью к условиям применения. Второй фактор заключается в возможности метаописания информационных ресурсов с нужной степенью формализованности используемого для этого инструментария, в открытом характере стандартов, позволяющем интегрировать средства пользователя в определяемую ими среду. Наконец, важную роль играют возможности XML как языка, поддерживаемого в глобальной коммуникационной среде Web. Обмен XML-сообщениями через Web позволяет обеспечить взаимодействие различного рода систем.

Назовем несколько конкретных направлений использования XML-платформы. В настоящее время создано и продолжает создаваться большое количество конкретизаций языка XML для разметки документов в различных предметных областях и создания DTD, согласованных различными профессиональными сообществами. Известны, в частности, версии DTD для применения в химии, географии, астрономии, истории, библиографии, издательском деле и др. Важная сферой применения становится e-Business.

В последнее время активно развиваются технологии баз данных XML, в которых XML используется в качестве языка определения данных (см. следующий раздел).

XML применяется также в системах управления документами, аналогичных тем, которые основаны на стандарте SGML и уже много лет используются на практике. Преимущество использования языка XML в этой сфере состоит в том, что становится возможной интеграция указанных систем в среду Web.

Следует далее упомянуть о применениях XML в стандартах других информационных технологий, где он используется как язык-посредник для обмена информацией между различного рода системами с помощью Web. В качестве примеров можно назвать созданный консорциумом OMG стандарт XMI (XML Metadata Interchange) [41] обменного формата метаданных для CASE, а также стандарт OIM (Open Information Model) [42, 43] консорциума Meta Data Coalition и созданный OMG на его основе стандарт CWMI (Common Warehouse Metadata Interchange) [44], определяющие формат представления метаданных и обмена метаданными для хранилищ данных. Планируется использовать XML для кодирования сообщений, которыми обмениваются клиент и сервер в известном стандарте ISO/IEC RDA/SQL (Remote Database Access for SQL) [45] удаленного доступа к системам SQL баз данных. В разрабатываемом консорциумом Workflow Management Coalition (WfMC) стандарте потоков работ [46] определяются спецификации XML

DTD, позволяющие осуществлять обмен сообщениями на языке XML между программными средствами потоков работ для поддержки их интероперабельности.

В связи с успешным продвижением платформы XML в практику, начались работы над новым ранее не планировавшимся компонентом SQL/XML [47] следующей версии стандарта языка SQL - SQL:200n [48]. По замыслу разработчиков, он будет определять возможности совместного использования ресурсов SQL и XML. В частности, будут определяться представление схем и данных SQL в форме XML-документов и наоборот.

Еще одно важное направление применения стандартов платформы XML, актуальное для создания электронных библиотек, интеграция неоднородных информационных ресурсов.

XML и технологии баз данных

Еще на ранних этапах развития Web-технологий сложилась тесная их связь с технологиями баз данных. Она сводилась к обеспечению теледоступа к системам баз данных через среду Web. В настоящее время создано и функционирует огромное количество приложений такого рода в самых различных областях деятельности. При этом не обеспечивается, однако, интеграции информационных ресурсов Web и баз данных. Система базы данных выступает здесь по отношению к Web как “черный ящик”.

На последующих этапах стали проявляться более глубокие связи между этими двумя направлениями информационных технологий.

Стремление к обеспечению в Web полноценных возможностей управления данными, поддерживаемыми в этой среде, объективно привело к необходимости использования подходов и принципов, аналогичных тем, которые на протяжении десятилетий прошли испытание временем в технологиях баз данных. Действительно, в стандартах Web-2 можно усмотреть целый ряд аналогий с идеями, воплощенными в технологиях баз данных.

Нетрудно обнаружить, в частности, воплощение в рассматриваемой среде концепций многоуровневого представления данных. Действительно, поддерживаются “логическое” и “физическое” представление данных XML-документов, обеспечиваемое языком XML. К сожалению, они сосуществуют пока в едином представлении XML-документа, что не позволяет полноценно реализовать в среде XML концепцию независимости данных. Поддерживается также несколько уровней представления метаданных. В лексиконе спецификаций стандартов платформы XML появились такие ключевые термины технологий баз данных, как модель данных, схема, ограничение целостности, язык запросов.

Во второй половине 90-х годов взаимодействие рассматриваемых двух направлений в информационных системах стало проявляться и на уровне развития практических технологий. Стали активно разрабатываться системы баз данных XML. Некоторые компании, выпустили для этих целей коммерческие программные продукты. Такие системы следовало бы называть точнее системами баз данных XML-документов. Хранимые в таких базах данных документы являются независимыми друг от друга, и никаких связей между ними не поддерживается.

В качестве схемы базы данных при этом используется DTD документов хранимых типов. Эту функцию может выполнять и описание XML-документов средствами стандарта XML Schema. Здесь может идти речь и об аналоге концептуальной схемы базы данных, роль которой способна играть RDF-спецификация. Для доступа к XML-документам разрабатываются языки запросов, как и в системах баз данных. Один из языков этого рода – XQL [49] - используется в продукте Tamino компании Software AG. В имеющихся проектах таких языков информационные ресурсы рассматриваются как множества

независимых XML-документов. Гиперссылки, определяемые стандартами XLink и XPointer, во внимание не принимаются.

В последнее время совместными усилиями специалистов в области XML-технологий и технологий баз данных активно ведутся работы по созданию стандарта языка запросов XQuery [18-23] для XML-платформы. Этот проект имеет чрезвычайно важное значение. Мы полагаем, что он не только обеспечит решение непосредственно инициировавшей его задачи - создания развитого языка запросов, но и будет иметь весьма существенные побочные эффекты.

Прежде всего, можно ожидать, что под его влиянием сформируется адекватный взгляд на моделирование данными в среде XML и будет осознана необходимость в единой базовой полнофункциональной модели данных платформы, основы которой и создаются в рамках проекта XQuery.

Далее, если проанализировать функциональные возможности прототипа XQuery, роль которого играет созданный основными авторами этого проекта язык Quilt [50], а также опубликованную W3C рабочую версию спецификаций создаваемого языка запросов, то складывается впечатление, что авторы стремятся уже изначально обеспечить язык такой функциональностью, которая выходит за узкие рамки потребностей работы только с XML-ресурсами. А именно, обеспечиваются возможности интеграции неоднородных информационных ресурсов, таких как XML-документы, данные иерархической и реляционной структуры.

Злободневная проблема интеграции неоднородных информационных ресурсов нашла встречный отклик и в технологическом “крыле” реляционных баз данных, где, как уже отмечалось выше, также предпринимаются попытки создания стандартных средств интеграции SQL и XML данных [47].

В контексте обсуждения баз данных XML важно обратить внимание также на разработанный и развиваемый W3C стандарт Document Object Model (DOM) объектной модели для XML-документов, на основе которого могут строиться интерфейсы прикладного программирования для систем баз данных XML.

С базами данных XML связано еще одно направление в технологиях баз данных. Выполнен ряд исследований, связанных с отображением XML-данных в среды реляционных [51, 52], объектно-реляционных [53] и объектных баз данных [54]. Эти направление разработок имеет практическую направленность, и их результаты вполне могут быть востребованы практикой.

Перспективы развития и применения XML-платформы

Создание XML-платформы положило начало новому более наукоемкому и технологически более совершенному этапу в развитии Web. Язык XML и некоторые другие стандарты основанной на нем платформы уже, несомненно, стали стандартами де-факто. Все ведущие поставщики программного обеспечения не только Web, но и систем баз данных, включают в свои программные продукты поддержку языка XML или даже создают специализированные основанные на нем системы. Продвижением XML-технологий в практику наряду с W3C занимается консорциум OASIS [55].

Распространению стандартов XML-платформы существенным образом способствует политика W3C, направленная на обеспечение доступности их спецификаций, создание ряда свободно распространяемых синтаксических анализаторов для языка и другого свободно распространяемого программного обеспечения, то большое внимание, которые создатели стандартов XML уделяют обеспечению преемственности для существующей HTML-платформы и накопленных на ее основе ресурсов.

При оценке перспектив языка XML нельзя также не учитывать, что он начинает играть существенную роль в других широко распространенных технологиях - CASE-технологиях, технологиях хранилищ данных, потоков работ, в технологиях баз данных, становится основой интеграции информационных ресурсов Web и реляционных баз данных. Предпринимаются также шаги, направленные на интеграцию XML-среды с объектными средами.

Что касается применения технологий XML в области электронных библиотек, то вполне уместный здесь оптимизм базируется прежде всего на таких важных в рассматриваемых приложениях свойствах XML-среды, как: наличие средств поддержки метаданных, описывающих свойства информационных ресурсов, в том числе и семантические свойства; способность технологий XML к интеграции с другими технологиями информационных систем, возможность обеспечивать совместно с ними интеграцию неоднородных информационных ресурсов; возможность транспортировки XML-данных в глобальной коммуникационной среде Web.

Вместе с тем, все еще существуют факторы, которые сдерживают энергичное массовое распространение XML в среде Web. Главных из них - два. Прежде всего, это - естественная инерционность столь масштабной среды, какой является сегодняшний Web. Эта инерция может преодолеваться только постепенно. Второй фактор - пока еще не завершена работа над двумя важнейшими стандартами платформы XML - XPointer и XLink, которые позволяют строить из отдельных XML-документов и их компонентов гипермедийную распределенную среду. В самом языке XML нет средств для определения гиперссылок. Эту задачу решают указанные стандарты. Учитывая тот факт, что стандарт XLink принят наконец в конце июня 2001 г., и состояние работы над проектом XPointer, можно полагать, что указанная проблема будет разрешена в ближайшее время.

Более подробное обсуждение некоторых возможностей платформы XML читатель может найти, в частности, в наших работах [35, 56-58]. Разработанный нами глоссарий русских терминов XML-технологий представлен на Web-странице [59].

Литература

1. Extensible Markup Language (XML) 1.0 (Second Edition). W3C Recommendation. 6-October-2000. http://www.w3.org/TR/2000/REC-xml-20001006.
2. Питтс Н. XML за рекордное время /Пер. с англ. - М.: Мир, 2000.
3. Эдди С.Э. XML: справочник. - СПб.: Питер, 1999. - 480 с.
4. Berners-Li T., Fielding R., Irvine U.C., Masinter L. Uniform Resource Identifiers (URI): General Syntax. RFC 2396. August 1998.
5. XML Protocol (XMLP) Requirements. W3C Working Draft 19 March 2001. http://www.w3.org/TR/xmlp-reqs/.
6. XML Information Set. W3C Candidate Recommendation, 14 May 2001. http://www.w3.org/TR/2001/CR-xml-infoset-20010514.
7. Namespaces in XML. W3C Recomendation, 14 January 1999. http://www.w3.org/TR/1999/REC-xml-names-19990114.
8. XML Pointer Language (XPointer). Version 1.0. W3C Working Draft. 8 January 20010. http://www.w3.org/TR/2001/WD-xptr-20010108.

9. XML Linking Language (XLink) Version 1.0. W3C Recommendation. 27 June 2001. http://www.w3.org/2001/REC-xlink-20010627

10. Style Sheet, level 2. CSS2 Specification. W3C Recomendation 12-May-1998. http://www.w3.org/TR/1998/REC-CSS2-19980512.

11. Extensible Stylesheet Language (XSL). Version 1.0. W3C Working Draft. 18 October 2000. http://www.w3.org/TR/2000/WD-xsl-20001018.

12. XSL Transformations (XSLT). Version 1.0. W3C Recommendation 16 November 1999. http://www.w3.org/TR/1999/REC-xslt-19991116.

13. XML Schema Part 0: Primer. W3C Recommendation. 2 May 2001. http://www.w3.org/TR/2001/REC-xmlschema-0-20010502.

14. XML Schema Part 1: Structures. W3C Recommendation. 2 May 2001. http://www.w3.org/TR/2001/REC-xmlschema-1-20010502.

15. XML Schema Part 2: Datatypes. W3C Recommendation. 2 May 2001. http://www.w3.org/TR/2001/REC-xmlschema-2-20010502.

16. Resource Description Framework (RDF). Model and Syntax Specification. W3C Recommendation. 22 February 1999. http://www.w3.org/TR/REC-rdf-syntax/.

17. Resource Description Framework (RDF). Schema Specification 1.0. W3C Candidate Recommendation 27 March 2000. http://www.w3.org/TR/2000/CR-rdf-schema-20000327.

18. XML Query 1.0 Requirements. W3c Working Draft 15 February 2001. http://www.w3.org/TR/2001/WD-xmlquery-req-20010215.

19. XQuery 1.0 and XPath 2.0 Data Model. W3C Working Draft 07 June 2001. http://www.w3.org/TR/2001/WD-query-datamodel-20010607.

20. XQuery 1.0 Formal Semantics. W3C Working Draft 07 June 2001. http://www.w3.org/TR/2001/WD-query-semantics-20010607.

21. XML Query Use Cases. W3C Working Draft 08 June 2001. http://www.w3.org/TR/2001/WD-xmlquery-use-cases-20010608.

22. XQuery 1.0: An XML Query Language. W3C Working Draft 07 June 2001. http://www.w3.org/TR/2001/WD-xquery-20010607.

23. XML Syntax for XQuery 1.0 (XQueryX). W3C Working Draft 07 June 2001. http://www.w3.org/TR/2001/WD-xqueryx-20010607.

24. Document Object Model (DOM) Level 2 Specification. Version 1.0. W3C Recommendation. 13 November 2000. http://www.w3.org/TR/2000/REC-DOM-Level-2-20001113.

25. XHTML 1.0: The Extensible Hypertext Markup Language. Reformulation of HTML 4 in XML 1.0. W3C Recomendation 26 January 2000. http://www.w3.org/TR/REC-xtml1-20000126.

26. XML Base. W3C Recommendation 27 June 2001. http://www.w3.org/TR/2001/REC-xmlbase-20010627.

27. XForms 1.0. W3C Working Draft 16 February 2001. http://www.w3.org/TR/2001/WD-xforms-20010216.

28. XML-Signature Syntax and Processing. W3C Candidate Recommendation 31 January 2001. http://www.w3.org/TR/2001/CR-xmldsig-core-20010131.

29. XML Path Language (XPath). Version 1.0. W3C Recomendation, 16 November 1999. http://www.w3.org/TR/1999/REC-xpath-19991116.

30. XML Inclusions (XInclude) Version 1.0. W3C Working Draft 26 October 2000. http://www.w3.org/TR/2000/WD-xinclude-20001026.

31. XML Fragment Interchange. W3C Candidate Recommendation 12 April 2001. http://www.w3.org/2001/TR/CR-xml-fragment-20010412.html.

32. Canonical XML. Version 1.0. W3C Recommendation. 15 March 2001. http://www.w3.org/TR/200/REC-xml-c14n-20010315.

33. Mathematical Markup Language (MathML) Version 2.0. W3C Recommendation 21 February 2001. http://www.w3.org/TR/2001/REC-MathML2-20010221.

34. ISO 8879:1986. Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML), 1986.

35. Когаловский М.Р. Энциклопедия технологий баз данных. – М.: “Финансы и статистика”, 2001 (в печати).

36. Когаловский М.Р., Новиков Б.А. Электронные библиотеки - новый класс информационных систем. Российская Академия наук. - М.: "Наука", МАИК "Наука/Интерпериодика", Программирование, 3, 2000. - С. 3-8.

37. Dublin Core Metadata Element Set Reference Description, Version 1.1, 1999-07-02. http://purl.org/dc/documents/proposed_recommendations/pr-dces-19990702.htm.

38. Dublin Core Qualifiers. Dublin Core Meta Data Initiative Recommendation. http://purl.org/dc/documents/rec/dcmes-qualifiers-20000711.htm.

39. ANSI/NISO Z39.85-2000x. The Dublin Core Metadata Element Set. Draft Standard. Modified September 30, 2000. National Information Standards Organization, 2000.

40. Semantic Web Activity: Resource Description Framework (RDF). http://www.w3.org/RDF/.

41. XML Metadata Interchange (XMI). Version 1.1. OMG Document ad/99-10-02.

42. Open Information Model. XML Encoding. Version 1.0. Review Draft 2. Meta Data Coalition. December 1999.

43. Open Information Model. Proposed XML Document Type Definitions. Meta Data Coalition. http://www.mdcinfo.com/OIM/xmldtds.html.

44. Common Warehouse Metamodel (CWM) Specification. Volume 2. XML, IDL and DTD. Proposal to the OMG ADTF RFP: Common Warehouse Metadata Interchange (CWMI). OMG Document ad/00-01-02. February 11, 2000.

45. ISO/IEC 9579:2000. Information technology - Remote Database Access for SQL (RDA/SQL).

46. Workflow Management Coalition. Workflow Standard - Interoperability. Wf-XML Binding. Document Number WFMC TC-1023. Draft 1.0. 20 April 1999.

47. Melton J. Subproject: “XML-Related Specs (SQL/XML)”. Project ANSI:1234D-ISO:1.32.3.4. 29 August, 2000.

48. Айзенберг Э., Мелтон Дж. Стандартизация SQL: следующие шаги. Пер. с англ. Открытые системы, 11-12, 1999, с. 80-84.

49. Robie J., Lapp J., Schash D. XML Query Language (XQL). The W3C Query Languages Workshop. December 3-4, 1998. Boston, Massachusets. http://www.w3.org/TendS/QL/QL98/pp/xql.html.

50. Chamberlin D., Robie J., Florescu D. Quilt: An XML Query Language for Heterogeneous Data Sources. Proc. of the WebDB 2000 Intern. Conf. Dallas, May 2000.

51. Malaika S. Using XML in Relation Database Applications. ICDE 1999.

52. Shanmugasundaram J., Tufle K., Zhang C., He G., DeWitt D.J., Naughton J.F. Relation databases for Quering XML Documents: Limitations and Opportunities. VLDB 1999, p. 302-314.

53. Klettke M., Meyer H. XML and Object-Relational Database Systems – Enhancing Structural Mappings Based on Statistics. WebDB 2000, p. 63-68.

54. Renner A. XML Data and Object Databases: A Perfect Couple? ICDE 2001.

55. OASIS homepage. http://www.oasis-open.org/.

56. Когаловский М.Р. XML: возможности и перспективы. Часть 1. Платформа XML и составляющие ее стандарты. – М.: Изд. “Открытые системы”. Директору информационной службы. Январь 2001. – С. 24-28.

57. Когаловский М.Р. XML: возможности и перспективы. Часть 2. Базы данных XML, семантика XML-документов, перспективы. – М.: Изд. “Открытые системы”. Директору информационной службы. Февраль 2001. - С. 16-20.

58. Когаловский М.Р. XML: сферы применений. – М.: Изд. “Открытые системы”. Директору информационной службы. Апрель 2001. – С. 10-12.

59. Когаловский М.Р. Глоссарий по технологиям XML. http://www.libweb.ru/resource/docs/xml/xml-gloss.html.ru.

XML PLATFORM STANDARDS AND DATABASES

M.R. Kogalovsky

Market Economy Institute

Russian Academy of Sciences

Moscow, 117418

Nakhimov Prospect, 47

Russia

e-mail: kogalov@cemi.rssi.ru

The paper is dedicated to analysis of the emerging XML technological platform foundation and functionality. The new platform is based on XML standard and intends to be a basis of second Web generation and means for heterogeneous information resources integration. The reasons of transition to XML technologies, the essence of radical Web changes, as well as XML platform organization and basic principles are discussed. The roles, classification, interrelationships of XML platform standards and their developing status are described. We consider also the approaches used to metadata representation and languages for the XML-document semantics description. Important and actual application domains of XML platform are presented. We discuss also the issues of the database and XML technologies convergence and mutual influence.