Система поддержки работы с удаленными XML-документами

А. Г. Марчук
Институт систем информатики им.А.П.Ершова СО РАН, Новосибирск
mag@iis.nsk.su

Аннотация

Понятие электронного документа требует определенной регламентации. Это касается его смысловых свойств: идентификации, метаинформации, регистрации, копирования и свойств копий, обеспечения жизненного цикла документов, и это касается регламентации технических решений, позволяющих повысить удобство работы с документами вне зависимости от их конкретной структуры. Первая группа вопросов рассматривается автором в текущем проекте. Результаты была изложена (совместно с А.Е.Осиповым) в докладе, представленном на первую конференцию по электронным библиотекам (Санкт-Петербург, 1999 г.), а также в работе [1]. Вторая, техническая, группа вопросов в настоящее время находится в центре внимания и рассмотрения Консорциумом "W3"в связи с обсуждением и принятием группы стандартов, касающихся XML-документов [2]. Настоящая работа описывает реализацию технологии работы с XML-документами, реализующей общие принципы, сформулированные в [1], обеспечивающей объектно-ориентированный стиль и интероперабельность в обработке.[1].

1 ДОКУМЕНТ КАК ОБЪЕКТ

Основная проблема, которая стоит перед автором, при проведении настоящего исследования заключается в построении корректной модели электронного документа, отражающей смысл документа как опубликованной информации в условиях новых технических возможностей опубликования в (виртуальном) информационном пространстве. С одной стороны, документ должен быть статическим (иначе он не документ), иметь неизменную ссылку и, по возможности, постоянное размещение. С другой стороны, электронные публикации существенно более динамичны, чем печатные, они могут изменяться, перемещаться, группироваться в новые группы. Электронные документы могут иметь разнообразные интерфейсы (как человеко-ориентированные, так и с ориентацией на машинную обработку), накапливать информацию (как в случаях баз данных, списков) и быть информационно зависимыми от других документов (напр. при изменении схем данных). В общем, документ должен вести себя как объект, и, как правило, модифицироваться и, в то же время, он должен вести себя как публикация, то есть - быть стабильным. Для разрешения этого противоречия автор предложил разделить два понятия: документ и экземпляр документа. При таком подходе экземпляр документа обладает всеми статическими свойствами, а документ - это множество своих экземпляров, которое в совокупности обладает объектными (в смысле объектно-ориентированного программирования) свойствами. Далее идет краткий пересказ основных положений концепции, изложенной в [1].

Документы являются объектами, "отвечающими"на присланное сообщение. При этом запускается некоторая процедура доступа, результат выполнения которой пересылается обратно запрашивающему агенту. В частном, но достаточно типовом случае, выполняется процедура создания и пересылки копии данного объекта. В другом частном случае, производится вычисление нового, измененного, значения (экземпляра) объекта и значение возвращается агенту для дальнейшего использования. В любом случае, экземпляр документа доступа остается неизменным.

Экземпляром документа назовем идентифицированный информационный набор, помещенный и постоянно присутствующий в информационном пространстве, не меняющий своего значения, в том числе идентификационного кода. Экземпляры документа, кроме того, расположены в некоторых обобщенных точках информационного пространства, точки имеют координаты. Зная координату, можно послать экземпляру сообщение и получить ответ.

Документом назовем множество экземпляров документа, имеющее единую идентификацию. Таким образом, множество всех экземпляров документов разбивается идентификационным кодом на непересекающиеся множества - документы.

[1]Работа выполнена при поддержке грантом РФФИ 98-07-91256 э.

Оригиналом документа назовем экземпляр документа, внешним (относительно содержимого документов) образом назначенным на эту роль. Статус оригинала документа может (со временем) переходить от одного экземпляра документа к другому (экземпляру того же документа). Для каждого документа, в каждый момент времени существования этого документа, есть один, и только один, оригинал. Два и более экземпляров документа называются копиями, если они являются одинаковыми (ведут себя одинаково, т.е. на одинаковые внешние запросы отвечают одинаковыми результатами), а различаются лишь их координаты.

Координату оригинала можно определить через объект (документ), который назовем регистратором или регистрационной коллекцией. Регистраторов в информационном пространстве может быть один и более. Каждый документ зарегистрирован только в одном регистраторе. По идентификатору документа однозначно определяется регистратор, в котором документ зарегистрирован. Регистраторы составляют иерархическую систему. Регистратор, как документ, зарегистрирован в своем родителе. Корневой регистратор зарегистрирован сам в себе. В итоге, зная координату корневого регистратора, мы всегда можем определить координату оригинала любого документа.

Ориентироваться в информационном пространстве документов помогают каталоги. Каталог является документом, хранящим произвольное число координат объектов в виде пары: идентификатор документа - координата экземпляра. Покрытие информационного пространства конкретным каталогом является произвольным, разные каталоги могут иметь пересекающиеся части.

Использовать документы предполагается в соответствии с двумя основными принципами. Первый принцип - информационное пространство неоднородно по доступу: условно, есть (могут быть) зоны более "близкие", более "далекие", в текущий момент недоступные. Это следует понимать так, что данные на "своем"компьютере более (напр. быстро) доступны, чем данные в локальной сети, те, в свою очередь, более доступны, чем данные в региональной сети и т.д. К тому же, возможна ситуация работы в условиях отключения от прямой связи с Интернет, не исключаются другие варианты.

Второй принцип заключается в том, что, как правило, использовать можно не оригинал, а одну из его предыдущих версий. Действительно, реальная правка документа может произойти несколько раньше или несколько позже. Эта недетерминированность может служить основанием использования копии вместо оригинала. Определение такой возможности осуществляется на основе критериев, сопровождающих запрос. Детальная структура критериев в статье обсуждаться не будет, но принципы вполне понятны, исходя из прагматики и традиций. Критерием может быть величина допустимой временной свежести, например "воспользоваться копией, если она последний раз была синхронизована с оригиналом не далее, как X дней (часов, минут, ...) назад". Или: "допустимо использовать версию N или более свежую".

Поддержанием (в основном, хранением) экземпляров документов занимаются объекты под названием "хранилища". Хранилище, в отличие от коллекции, не является документом, а является физическим подпространством для размещения документов. Хранилище обеспечивает выполнение (возможно, удаленных) запросов на доступы к документам. Хранилище может производить синхронизацию документов с оригиналами. Синхронизация документа выполняется следующим образом: в соответствующем регистраторе находится координата оригинала, если копии оригинала нет в хранилище, то она там порождается. После этого хранилище способно предоставлять обновленное значение документа.

Общая схема статической, т.е. не изменяющей значения, работы с документами такова: 1. Экземпляры и копии экземпляров документа "разбросаны"по информационному пространству. 2. Находится "ближайший"экземпляр и проверяется соответствие его атрибутов требованиям к доступу. Если соответствие имеет место, экземпляр используется вместо оригинала документа. 3. Если экземпляр не соответствует требованиям, то ищется более "удаленный"экземпляр объекта, соответствующий требованиям запроса на доступ. Оригинал (по определению) должен удовлетворять требованиям запроса, и хранилища копий могут произвести синхронизацию "своего"значения документа с оригиналом. Решение о синхронизации принимает хранилище, в соответствии с собственной стратегией "кеширования"данных.

Строгой синхронизацией назовем требование на предоставление доступа только к копии оригинала. Задача решается в контексте взаимодействия читателей и писателей. При строгой синхронизации запрос на чтение (использование) может быть задержан на период фиксации нового значения оригинала. Управляет синхронизацией доступа к объекту регистратор. В регистраторе фиксируется одно из возможных состояний объекта (документа): "ДОСТУПЕН"с параметром в виде времени, до которого документ гарантированно не будет изменен, и "БЛОКИРОВАН"с аналогичным параметром. Синхронизация читателя заключается в том, чтобы он успел использовать текущее значение документа до истечения указанного времени. Синхронизация писателя заключается в том, что он посылает заявку на обновление документа и ожидает от регистратора разрешения это сделать. Регистратор, в соответствии с политикой обслуживания читателей и писателей, в подходящий момент блокирует документ, возвращает писателю разрешение на модификацию и временной интервал, в течение которого надо успеть породить новое значение и сообщить об этом.

2 МОДЕЛЬ

Предложенная концепция электронных документов была реализована в виде технологического комплекса (набора спецификаций, средств и инструментов), ориентированного на интероперабельную работу с XML-документами. В настоящее время реализация имеет некоторые упрощения модели относительно [1], которые заключаются в следующем: объекты доступа и обработки - только XML-документы, идентификационная система реализует только URI, оригинал документа модифицируется владельцем документа, регистраторы документов отсутствуют, их функции реализуются полем doc-name, содержащимся в метаинформации о документе, роль хранилища исполняет административная система. Поскольку в систе-

ме стандартов XML термин документ имеет свое специальное значения, автор в дальнейшем тексте использует термин "целевой документ".

Целевым документом назовем XML-документ, "выставленный"в Интернетовское пространство и доступный в нем через регламентированные доступы. Доступы могут быть как на "чтение", т.е. без модификации, так и на "запись с модификацией. Целевой документ может существовать лишь в условиях его поддержки административной системой некоторого сервера.

Административная система - серверная система, обеспечивающая создание целевых документов, их поддержание под одним URL-адресом, модификацию целевых документов в соответствии с запросами и уничтожение целевого документа с дальнейшей поддержкой попыток доступа к нему. Запрос (к целевому документу) представляет собой сообщение регламентированного формата, характерного для CGI, инициирующее специальный доступ к документу, часто, модифицирующий содержимое документа. Для немодифицирующих доступов возможно использование существующих в стандартах XML ссылок (links) и вставок (entities), хотя есть и ограничения на их использование.

Целевой документ можно рассматривать как значение объекта некоторого класса. Соответственно класс этого объекта определяет доступы и конструкторы, и, частично, структуру документа. Целевой документ может быть произвольной структуры. Для реализации этого свойства и возможности однородной работы с целевыми документами со стороны административной системы, объект, как XML-элемент, помещается в (XML) капсулу, фиксирующую существенную мета-информацию о данном документе. К мета-информации относится имя документа (doc-name) и имя класса документа (doc-class). Другая мета-информация может помещаться в раздел meta. В соответствии с естественными особенностями XML-подхода, документ может использоваться по доступу без модификации напрямую (прочтением).

Модификация документа удаленным абонентом производится по следующей схеме: присылается запрос, содержащий данные для модификации. Этот запрос группируется с целевым документом и к ним применяется XSLT преобразование, определенное как доступ в классе документа. Результат преобразования заменяет старое значение целевого документа или возвращается посылавшему сообщение агенту. Кроме основного результата, может быть сформировано несколько сообщений, посылаемых другим документам (возможно, этому же). Эти новые сообщения посылаются от имени того же агента и могут иметь статус немедленного исполнения или быть поставлены в очередь. В процессе проведения преобразования может выясниться невозможность применения пришедшего запроса. В этом случае, цепочка действий насильно прерывается, а агенту возвращается код ошибки.

3 РЕАЛИЗАЦИЯ В СРЕДЕ XML-XMLT-SERVLETS

Оперативная часть административной системы реализована в виде сервлета quest. Сервлет принимает строку определенного вида и далее выполняет некоторые действия.

Первая фаза содержательных действий обработчика quest заключается в следующей последовательности:

- формируется XML-значение запроса inquiry;
- по заданному в запросе значению атрибута target-doc находится целевой документ;
- по заданному в целевом документе полю doc-class находится документ, определяющий класс целевого документа;
- по заданному в запросе значению атрибута action-variant в документе doc-class находится XSLT документ, определяющий требуемые преобразования;
- запрос и целевой документ объединяются в группу transform-group и подвергаются обработке с использованием найденного XSLT документа;
- используемый XSLT документ должен преобразовывать целевой документ в документ также формата target-doc;
- если сформирован элемент messages и его атрибут errors установлен в состояние "yes", то среди сообщений находится message с отсутствующим полем direction (предназначенный для возврата источнику) и внутренность этого сообщения возвращается источнику;
- если ошибок преобразования не зафиксировано, то:
- имеющиеся сообщения рассылаются по указанным направлениям;
- результирующий документ (в устраненными сообщениями?) либо замещает целевой документ, либо возвращается источнику в соответствии со значением атрибута direction в элементе access документа doc-class.

4 ВЫВОДЫ

Описанный технологический комплекс обладает целым рядом удобств и преимуществ. Удобства, в основном, связаны с тем, что поскольку трансформации описываются специализированным языком XSLT, для построения простых систем и подсистем, можно для значительного количества случаев обойтись совсем без программирования на алгоритмических языках. Это является существенным, когда специфические особенности доступа к данным должны задавать администраторы данных. Такой подход был опробован для построения диалогов для WWW страниц и показал свою эффективность.

Существенные преимущества выявляются при использовании технологического комплекса для построения распределенных и сложно устроенных систем коллекций и баз данных. Предложенные решения позволяют рассредоточить в пространстве и времени администрирование разными частями данных не теряя целостности данных и обеспечивая синхронизацию за счет базовых средств.

Для проверки этих положений был реализован административный интерфейс для общественной организации, имеющей территориальные отделения. База данных членов общественной организации теперь может поддерживаться как централизованно, так и силами администраторов из отделений, причем в произвольной смеси этих действий и в стиле систем типа Source Safe.

Другой группой преимуществ является встроенная возможность использования близких копий вместо удаленных оригиналов. Эти же синхронизационные свойства позволяют решать задачу однородного доступа к данным, расположенным на неоднородных носителях. Например, база данных или коллекция может поставляться на CD, изменения могут браться из Internet, а локальный кэш может быть организован на HDD. Программные средства, которые позволят создавать подобные авто-обновляемые информационные системы, сейчас находятся в разработке.

5 СПИСОК ЛИТЕРАТУРЫ

1. А.Г.Марчук, А.Е.Осипов К вопросу об идентификации электронных документов и коллекций// Программирование, N 3, 2000 г.
2. W3C Technical Reports and Publications // http://www.w3.org/TR