К вопросу об идентификации электронных документов и коллекций

А.Г.Марчук, А.Е.Осипов (Новосибирск)
E-mail: mag@iis.nsk.su
Работа выполнена при поддержке гранта РФФИ 98-07-91256э

1 Постановка задачи

При всем обилии существующих стандартов и технологий, определяющих пространство Internet, вопрос идентификации документов и фактов является одним из наименее проработанных. В самом деле, "прописать"факт достаточно произвольной природы можно только в некоторых специализированных системах. К сожалению, не существует общепринятого формата фиксации документов даже таких важных видов как персональные данные. В итоге, встретив в документах упоминание, например, "Иванов Иван Иванович", вы не имеете достоверного факта, связывающего конкретного Иванова с данным документом. Использование для идентификации гиперссылки с конкретным URL мало что меняет, в силу нестандартизованности информации, расположенной по данной ссылке.

Задача состоит в том, чтобы выработать систему идентификации электронных (возможно, не только) документов, предназначенную для работы (в том числе автоматической) с ними в пространстве Internet. Система должна поддерживать весь жизненный цикл документа и учитывать технологические особенности дистанционной работы с документами разных видов и размеров, и разнообразие серверных и пользовательских платформ.

Исследуется возможность применения идей объектно-ориентированного программирования, интероперабельных и клиент-серверных, CORBA, COM и Proxy систем для построения целостной системы идентификации и регистрации документов, и коллекций в пространстве Интернет. Основой подхода является взгляд на документ как на объект в объектно-ориентированном программировании [1].

2 Документ как объект

Исходя из целого ряда причин, имеет смысл взгляд на (электронные) документы как на объекты. Документом назовем уникальный целостный информационный объект (в смысле объектно-ориентированного программирования) фиксированного класса или типа, помещенный в информационное пространство и доступный в нем постоянно и однородно. Существенными в данном подходе являются следующие моменты.

Документ в общем случае не является статическим, он может (и должен) иметь мета-динамику - рождение, модернизацию и окончание своего существования. Он может иметь текущую динамику - изменение значения. Следующим моментом является видимость значения документа (объекта) через призму доступов и событий. Доступы и реакции на события определяются соответствующим классом и обычно выполняются удаленно. Естественно также говорить об уникальности документа. Очевидно, что наличие самостоятельных, "живущих своей жизнью"копий документа, противоречит сути понятия "документ". Копирование и синхронизация копий должны быть определенным образом регламентированы для поддержания корректной работы в распределенном информационном пространстве.

Возможно, нужно говорить не о документах, а об информационных ресурсах, информационных объектах или использовать новую терминологию. В работе понятие документа сознательно расширяется до понятия объекта, имеющего регламентированные доступы. Поэтому объект может быть не только "выдающим"информацию, но и "потребляющим"ее.

3 Идентификация документов

Достаточно легко поместить в информационное пространство новый объект, например, WWW страничку. Однако можно ли при этом говорить о порождении нового документа? Очевидно, что в общем случае - нет. Такой "безответственный"объект, строго говоря, не является также и публикацией. Дело здесь не в том, что автор или кто-то другой могут произвольно менять содержание информации, и не в том, что "приписка"документа к позиции в информационном пространстве (URL) достаточно условна и также может измениться. Принципиальными являются только два момента: идентификация и регламент. Всегда надо иметь возможность определить: об одном документе идет речь или о разных; этой цели служит идентификация. Однако, если допускается изменение информационного содержимого документа, то только при наличии некоторых регламентирующих соглашений и процедур можно говорить об объекте как о документе.

Технически вопрос об идентификационном коде решается достаточно просто. Уже предложено, или может быть предложено, много подходов к структуре кода и

Первая Всероссийская научная конференция
ЭЛЕКТРОННЫЕ БИБЛИОТЕКИ:
ПЕРСПЕКТИВНЫЕ МЕТОДЫ И ТЕХНОЛОГИИ,
ЭЛЕКТРОННЫЕ КОЛЛЕКЦИИ
19 - 21 октября 1999 г., Санкт-Петербург

его присутствии в документах (ссылки). Наиболее привлекательным выглядит система идентификации COM-объектов, разработанная фирмой Microsoft [2]. Однако вероятностный способ выработки уникального 128-битного кода представляется спорным и не гарантирующим основного требования. В любом случае, идентификационный код документа должен состоять как минимум из двух частей: уникального идентификатора и навигационной информации, по которой можно проложить маршрут к оригиналу документа. Добавление в идентификационный код информации о типе объекта и коллекции, хранящей документ, кажется, является избыточной и может отсутствовать.

Важна также организационная составляющая технологии идентификации документов. Хотелось бы иметь возможность получать свой "домен"кодов идентификации и использовать его, не обращаясь за каждым кодом в центральный распределитель. Доменная структура кода может также помочь в решении вопроса навигации, так как домен можно зарегистрировать в некоторой иерархии регистрации, и использовать технологию, аналогичную DNS.

Вряд ли целесообразно повторное использование идентифицирующих кодов в случаях, когда документ заведомо прекратил существование. Лучше, если система получит по старой ссылке информацию типа "документ отменен", чем будет пытаться использовать случайно попавший в данный контекст объект.

4 Взаимодействие агента с документом

Взаимодействие удаленного агента и объекта (документа) осуществляется в общем случае дистанционно, выполнением соответствующего доступа (в смысле ООП). Результат доступа "присылается"агенту и используется локально. Результат является также объектом или ссылкой на него, т.е. его идентификатором. Как уже указывалось, ссылка должна позволять не только идентифицировать объект, но и устанавливать его местонахождение и выполнять доступ к нему.

В силу технических и технологических особенностей пространства Интернет, необходимо иметь возможность "приблизить"объект полностью (зеркальная копия) или частично (рабочая копия) к агенту - потребителю/производителю информации. Это создает разнообразные задачи синхронизации оригинала документа и его копий. Предлагается, всю группу вопросов, связанных с копиями документов, рассматривать в рамках единого подхода, основные элементы которого изложены далее.

Документ помещен в некоторую среду, назовем ее коллекцией, которая не только обеспечивает сохранность документа, но и обеспечивает (удаленные) доступы к документу как объекту определенного класса. Такой взгляд позволяет говорить о коллекции как о платформе.

В простом случае идентификатор документа содержит прямое указание о локализации платформы (коллекции), соответственно, контакт между агентом и удаленным объектом осуществляется традиционным образом, например, в технологии CORBA. Однако желание обеспечить длительное существование документа, например, дольше времени существования сервера или домена, или по другим причинам, приводит к необходимости обеспечивать доступ к документам косвенно, через регистрационные коллекции (базы данных). При этом документ регистрируется в некоторой регистрационной коллекции, и навигационная информация определяет не коллекцию, охватывающую оригинал документа, а коллекцию- регистратор. Теперь документ выглядит помещенным в другую коллекцию, а его реальное расположение может меняться.

Документ как информационный объект обязан иметь метаинформацию о себе. Метаинформация включает в себя статусную информацию, такую как: статичность документа, возможность создания копий, наличие авторского и имущественных прав, возможный терминатор (VALID, BROKEN, TIMED_OUT, etc.), наличие резервных копий и др. Кроме того, в метаинформацию входит структурная информация - класс/тип и информация о копии, если это копия документа.

Метаинформация хранится при оригинале документа (в коллекции оригинала), часть метаинформации хранится в регистрационной записи, часть - при копиях. Собственно, поскольку метаинформация "привязана"к соответствующему документу, она является ничем иным, как полями информационной записи, т.е. свойствами (properties - в терминологии COM) объекта.

5 Классы и типы

Взгляд на документ как на объект, естественным образом диктует наличие у объекта класса, его свойств и доступов (методов). Собственно, базовая технология реализации объектов в Internet уже, в основном, сформировалась в различных подходах к интероперабельным системам. Класс документа понимается в выработанном для интероперабельных систем представлении. Это, в частности, определяет технологию удаленного доступа к объектам через "заглушки", IIOP, CORBA [3]. Такие технологии как CORBA, COM, RMI, Java Beans и др. дают основные необходимые средства. Однако предлагаемый подход к существованию в информационном пространстве электронных документов предполагает дополнительную регламентацию и спецификацию.

Очевидно, конкретные классы, определяющие те или иные документы, должны расширять базовый класс (напр., Document). Базовый класс содержит метаинформационные поля и ряд абстрактных доступов и интерфейсов, необходимых для реализации целостной концепции документа. В остальном - класс может иметь в значительной мере произвольную структуру и методы.

Коллекции и регистрационные коллекции являются документами определенного вида и регламентированного класса. Как правило, коллекции не являются статическими образованиями и могут модифицироваться (пополняться и редактироваться) определенными действиями. Важным свойством коллекции является возможность принимать участие в формировании специализированного информационного пространства в рамках распределенной системы поддержки информационных ресурсов (документов) определенного вида. Например, библиографические базы данных могут быть объединены в объединенную коллекцию первичных библиографических документов. Это можно сделать созданием коллекции коллекций или построением специализированной коллекции, обеспечивающей (возможно оптимизирующей) доступ к документам этого специализированного подпространства.

Другим подходом для "облегченного"решения типовых задач по доступу является механизм типизации данных. Он дает еще более простую реализацию доступа к простым информационным ресурсам типа записи. Использование механизма рекурсивной типизации позволяет также экономно решить проблему компактной "сери-

ализации"данных при их передаче через среду передачи данных.

6 Копии документов, виртуальные документы и коллекции

Документ должен быть уникальным, поэтому все копии документа должны иметь вспомогательную роль и специальный статус. Рассмотрим вопрос создания и поддержания копий документов и коллекций. Возможность порождения копии документа определяется статусной информацией о документе и наличием соответствующих доступов, например доступа типа .clone(). При этом агент получает копию документа (файл) либо по "проводам", либо другим доступным ему способом, например, в виде пластинки CD-ROM. Документ имеет (в данной миниупаковке) соответствующие атрибуты, необходимые, в частности, для синхронизации (напр., дата копирования). Документ может быть размещен в некоторой коллекции, если коллекция имеет средства, поддерживающие работу с документами нужного класса. И после этого может быть инициирован. Данная коллекция может обслуживать доступы агентов к документу, синхронизируясь с оригиналом (возможно, через вышестоящую копию) по необходимости.

Документ может иметь такие "неприятные" специфические особенности: он может (пока) отсутствовать, быть неполным или недостоверным. Или просто не найденным в пространстве (возможно, не зарегистрированным). При этом такой документ, тем не менее, может быть необходим для работы информационной системы или коллекции. Например, при анализе текста (документа) мы встретили упоминание о человеке (книге, статье, городе, и т.д.), причем контекст документа недвусмысленно подтверждает, что соответствующий электронный документ, понятно какого класса, может существовать или даже должен существовать, но пока его нет или он не найден. В таких случаях полезно заводить "заменитель"документа в виде виртуального документа и формировать содержимое такого документа косвенно, исходя из имеющейся информации. Такие lost+found документы могут существенно упорядочить и ускорить процесс расширения тематических фактографических "пространств"данных, хотя при этом концепция усложняется за счет необходимости решать дополнительные задачи идентификации и синхронизации.

Существенными "пластами"информационного содержимого Internet являются однородные или родственные информационные наборы сложившейся (для потребителя) структуры. Такими наборами документов являются электронные публикации, библиографическая информация, персональная информация, электронные "визитные карточки"фирм, организаций, территориальных и прочих образований, распространяемое программное обеспечение и т.д., и т.п. В большинстве случаев трудно представить возможную централизацию таких наборов в виде, напр., глобальных коллекций. Однако однородный доступ к распределенной информационной базе документов такого вида необходим для многочисленных приложений. Решение данной проблемы видится в формировании общего информационного поля из отдельных коллекций документов вокруг согласованной системы классов или типов. Такие поля естественно назвать виртуальными коллекциями. Для их создания необходимо породить соответствующие регистраторы (для случая, когда удается договориться об общем регистрационном механизме) или поисковые системы, работающие в соответствии с той или иной стратегией (для случая, когда отсутствует общая система поддержания виртуальной коллекции).

Проблемой для виртуальных коллекций, как правило, является согласование классов, определяющих входящие документы. Практика показывает, что представление о структуре документов различается у разных групп разработчиков. Кроме того, классы и их реализация со временем изменяются, что не представляет большой проблемы для однородных коллекций, но трудно преодолеваемо для виртуальных. Видимо до тех пор, пока специалисты не научатся договариваться о форматах данных, подобные коллекции, включаемые в виртуальные коллекции, нужно будет сопровождать дополнительными преобразователями из одних представлений в другие и сложной системой организационных мероприятий по внесению изменений в конкретные коллекции при изменении базовых стандартов.

7 Заключение и выводы

Предлагаемый подход затрагивает весьма существенный для создателей информационных систем и электронных библиотек круг вопросов. Для библиотечных и библиографических систем - это такие вопросы как: статус и реализация электронного документа в информационном пространстве, обеспечение жизненного цикла документов, коллекций и систем, формирование первичных и вторичных коллекций, формирование универсальных и специализированных поисковых систем и баз данных, создание и поддержание виртуальных коллекций документов, копирование и синхронизация. Авторы далеки от мнения, что работа по созданию концепции завершена, более того, для получения приемлемого результата необходимо объединение усилий разных групп разработчиков и специалистов. Однако, в любом случае, предполагается создать прототипную систему, максимально использующую сложившиеся стандарты и технологии. Особый, пусть несколько локальный интерес, представляет создаваемая по данной технологии система распространения коллекций и баз данных на CD-ROM. Система позволит автоматически поддерживать актуальность пользовательской копии базы данных, без частых закупок обновленных версий оптических дисков.

Библиография

[1] Марчук А.Г., Осипов А.Е. *Обеспечение унифицированного доступа к разнородным коллекциям и информационным ресурсам на основе технологии CORBA*//Тезисы докладов семинара-совещания "Электронные библиотеки: перспективные методы и технологии, электронные коллекции".– М.:1998.– С.2–3.

[2] Box, Don. *Essential COM.* Addison Wesley Longman, Inc., 1998, 440p.

[3] Object Management Group *The Common Object Request Broker: Architecture and Specification.* – 1995.