Объединение разнородных информационных ресурсов в электронном каталоге.

Палей Д.Э., Курчинский Д.Н., Смирнов В.Н., Русаков А.И.
Ярославский Государственный Университет им. П. Демидова
E-mail: paley@yars.free.net

1 Постановка задачи

Системы хранения и обработки данных (Data system) являются на сегодняшний день основой для большинства современных информационных систем. Принципы организации и хранения данных во многом определяют функциональные возможности всей системы в целом. С другой стороны одними из определяющих для проектирования и создания систем хранения данных являются специфика предметной области, которую описывают данные, и функциональные возможности, которые должна обеспечивать система.

На практике во многих случаях требуется сохранение и повторное использование данных, точная структура которых заранее не определена. При этом лишь априорные представления о характере информации и взаимосвязях ее частей. Организация словаря данных в этом случае ложится на конечного пользователя, который зачастую является специалистом в конкретной предметной области. Вместе с тем, всем пользователям таких систем необходим доступ к информации о структуре данных и их взаимосвязях (информация о метаданных).

Приведенное на рис.1 разбиение пользователей подобных систем на группы достаточно условно, но достаточно хорошо отражает большинство встречающихся на практике случаев. Информация о метаданных необходима потребителям данных, для навигации и извлечения данных. В равной степени она необходима и разработчикам информационных систем для использования сохраненных данных. Администраторы данных не только просматривают имеющуюся фактическую информацию, но могут изменять и метаданные.

При построении систем хранения и обработки данных возникают известные трудности и противоречия. В качестве типичного примера можно привести требования с одной стороны простоты и доступности конечным пользователям принципов организации данных, с другой стороны требования сохранения разнообразных данных со сложными взаимосвязями. Таким образом речь идет о проблеме создания информационной системы, предназначенной для хранения и повторного использования разнородной информации, структура которой определяется конечным пользователем, специалистом в конкретной предметной области. Далее будем называть такую систему хранилищем данных. Один из вариантов построения хранилища данных и рассматривается в докладе.

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

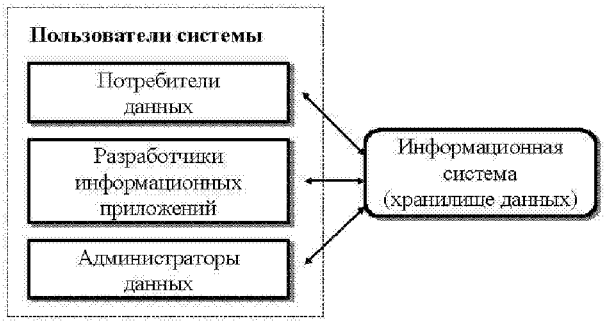

Рис. 1: Взаимодействие с хранилищем данных различных типов пользователей.

2 Структура хранилища данных

Общая функциональная схема предлагаемого хранилища данных приведена на рис 2. Основными его компонентами являются: репозиторий, электронный каталог, интерфейсы доступа к ним, система авторизации.

Репозиторий определяет типы и взаимодействия хранимых данных. Электронный каталог собственно содержит информацию. Интерфейсы доступа позволяют внешним приложениям оперировать структурой данных и самими данными.

Чтобы определить принципы организации репозитория и электронного каталога рассмотрим вначале характерные признаки информации, которую предполагается помещать в хранилище данных. Некоторую сущность - реальное явление, предмет, описание которого как единого целого должно быть помещено в хранилище будем называть *артефактом* К наиболее важным особенностям сохраняемых данных можно отнести следующее:

- достаточно большое количество артефактов различных по структуре описания;
- для большинства артефактов полное описание не закончено, полностью не определено, или подразумевается, что оно может быть изменено в дальней-

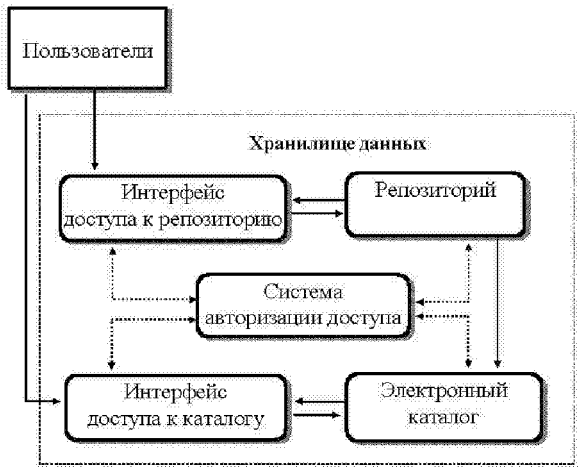

Рис. 2: Функциональная схема хранилища данных

шем; также возможна ситуация двух различных формальных описаний одного и того же артефакта;

- большинство артефактов связано согласно той или иной системе с другими, причем эти связи разнообразны, заранее четко не определены, могут изменяться

Информация подобного плана возникает во многих предметных областях например, при описании музейных коллекций, хранилищ редких книг, описании архитектуры и т.д. Для авторов доклада задача сохранения подобной информации возникла в ходе работ по грантам РГНФ, РФФИ по сохранению культурного наследия учреждений культуры Ярославской области. Это работы по сохранению фресковой живописи храмов города Ярославля, сохранению коллекции редких книг Ярославской областной библиотеки и библиотеки ЯГПУ и т.д.

Указанные условия во многом определяются спецификой предметной области и связаны с различными подходами в описании и систематизации применяемыми научными работниками не только различных учреждений культуры, но зачастую и одного музея, библиотеки, архива.

Оптимальным для сохранения данных с такими свойствами является объектный подход. Определим далее некоторые термины: *Классом* будем называть описание некоторого артефакта, сохраненное в репозитории системы в виде набора атрибутов и правил их обработки. Единицей хранения информации является *объект хранения* (или просто *объект*). Объект хранения представляет собой экземпляр класса, сохраняющий информацию о конкретном артефакте.

В соответствии с вышесказанным, определим далее функциональные требования к хранилищу данных

- описание и хранение объектов различной структуры, которая заранее не определена;
- хранение и администрирование различных типов описательной информации (текст, числовые данные, изображения , звук, видео);
- обеспечение доступа пользователей хранилища к репозиторию системы с возможностью изменения классов и их взаимосвязей;
- систематизация объектов хранения путем объединения их в иерархическую структуру;
- обеспечение системы ссылок между объектами хранения и их атрибутами, как на уровне определения классов, так и на уровне ввода и администрирования объектов;
- встроенные механизмы, обеспечивающие возможность повторного использования

Большинство перечисленных требований в настоящее время являются достаточно распространенными. Различные варианты реализующих их информационных систем и хранилищ данных достаточно хорошо описаны в литературе. Вместе с тем предметная область и практические стороны использования системы накладывают определенную специфику на принципы построения системы.

3 Принципы моделирования и описания данных и объединения объектов в каталог

Как уже было указано наиболее оправданным подходом организации хранения данных является объектный. Вместе с тем разумным, по мнению авторов, является введение некоторых ограничений на объектную модель. Они обусловлены прежде всего спецификой использования хранилища данных и требованиями к пользователям системы.

Определения классов хранятся в репозитории системы. Класс определяется набором атрибутов и методов. Типы атрибутов класса определяются в репозитории. Атрибуты могут иметь и стандартные скалярные типы, и типы, определяемые пользователем в том числе содержащие BLOB данные. В качестве атрибута может выступать массив. Методы (механизмы обработки) объектов, согласно принятому подходу также являются атрибутами классов и хранятся в репозитории для каждого класса. Вместе с тем атрибуты не могут иметь объектный тип.

Это ограничение на наш взгляд является оправданным. При объектно-ориентированном подходе описания данных фактически используются те же абстракции, что и при семантическом моделировании и анализе. Определение классов объектов основывается абстракции агрегации, формирование иерархии классов соответствует абстракция специализации-обобщения и т.д.

Определение составных классов (атрибуты которых также являются классами) основывается на абстракции агрегации. Чтобы определять составные классы, необходимо выделить некоторые элементарные сущности, которые описываются классами и объединены отношением "часть-целое". Такой анализ подразумевает с одной стороны достаточно полную априорную информацию о структуре данных, с другой стороны, подразумевает что отношения "часть-целое" являются постоянными. В данном случае эти требования часто не выполняются. Структура сохраняемых данных заранее четко не определена, отношения "часть-целое" не фиксированы, более того один и тот же артефакт (набор артефактов) при различных подходах к анализу данных может быть описан различными составными классами. (Рис.3)

Если учитывать эти особенности, то применение составных классов во многих случаях приводит к неоправданному усложнению и избыточности модели данных.

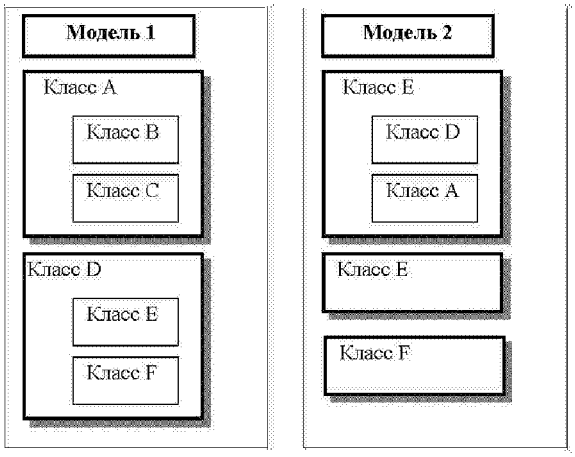

Рис. 3: Различные структуры описания данных с одним и тем же набором объектов

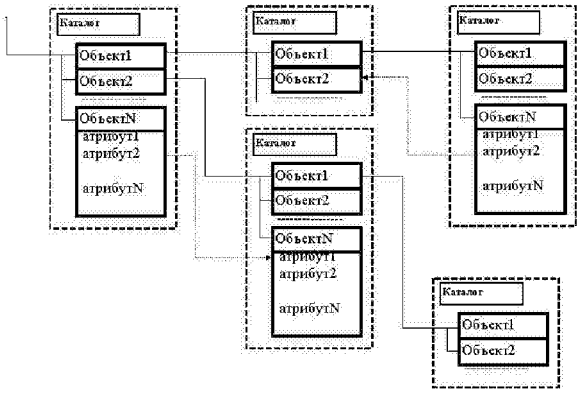

Рис. 4: Пример каталога и взаимосвязей объектов

Также следует отметить, что сложности с использованием составных классов возникают и при модификации структуры данных, т.к. введенная в хранилище информация должна быть полностью преобразована в соответствии с новым описанием. Таким образом, модель данных при наличии уже введенной информации теряет необходимую гибкость и становится трудно модифицируемой.

Описанную проблему предлагается решить следующим образом:

- Созданием механизма ссылок типа "атрибут-объект", "атрибут-атрибут". Это решение является достаточно стандартным. Особенность в данном случае состоит в том, что реализация этих функций возложена на электронный каталог. Отделение ссылок от репозитория позволяет практически формировать произвольные связи между объектами электронного каталога. Значения всех атрибутов (кроме специально определенных) могут указывать на другие объекты или атрибуты других объектов каталога. При этом ссылка может быть синхронной (значение ссылающегося атрибута автоматически обновляется по мере обновления предмета ссылки), так и асинхронной (значение атрибута обновляется не автоматически). В таком виде ссылки являются частью данных и определяются пользователями системы при вводе и администрировании объектов.

- Объединением всех объектов отношением главный-подчиненный в некоторую иерархическую структуру - собственно каталог. По мнению авторов доклада, такое размещение объектов является достаточно естественным. Подобная иерархическая структура расположения объектов является естественной при анализе фактического материала во многих предметных областях. Также она обеспечивает естественную форму доступа к требуемым данным. Система поддерживает многовариантность иерархических каталожных структур на одном множестве объектов, что значительно облегчает повторное использование данных.

Единицей хранения набора объектов в хранилище данных случае является каталог. Вместе с тем можно сказать, что каталог представляет собой не что иное, как поддерживаемую системой предопределенную систему ссылок. На рис.4 в качестве примера показана часть иерархической структуры электронного каталога и возможные ссылки между объектами и их атрибутами (пунктирная линия).

4 Повторное использование информации

Обратимся теперь проблему повторного использования данных. Рассмотрим следующий ее аспект. Часто возникают задачи создания объектов, значения некоторого набора атрибутов которых совпадали бы со значениями соответствующих атрибутов другого объекта. В качестве примера можно привести следующую ситуацию. Положим, есть набор объектов, являющихся экземплярами класса "персона". Атрибуты этих объектов содержат полную информацию о человеке. У класса "персона" есть порожденные классы, в том числе класс "сотрудник". Положим далее, что при реализации какого-либо проекта понадобилось создать экземпляры класса "сотрудник", и что объекты "сотрудник" содержат персональную информацию некого подмножества объектов "персона".

Подобную проблему на первый взгляд легко решить с помощью системы ссылок. Достаточно определить значениями выбранных атрибутов объектов "сотрудник" ссылки на соответствующие значения атрибутов объектов "персона". Вместе с тем при больших объемах информации и ее частом использовании это является достаточно трудоемкой операцией. То есть появляется необходимость поддержки механизма ссылок по данным между объектами зависимых классов самой системой. Итак предполагается следующее.

Пусть имеется класс типа C0 и некоторое количество порожденных от него классов Ci (i=1...N). Обозначим соответственно объекты, являющиеся экземплярами этих классов, Oij (i=0..N, j=0..?). Механизм ссылок по данным предполагает, что при объявлении некоторого экземпляра Omn наследником по данным Rbk (при условии, что тип Ca производный от Cb), соответствующие атрибуты экземпляра Rmn будут иметь значения атрибутов экземпляра Rbk. Назовем эту возможность "наследованием данных" - на наш взгляд этот термин наиболее удачно

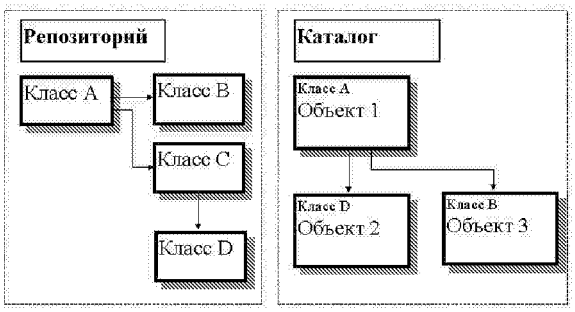

Рис. 5: Пример наследования по данным.

описывает ситуацию (Рис.5). Принципиальным является то, что эти функции должно обеспечивать хранилище, а не разработчики информационных систем на основе хранилища и не его пользователи.

5 Поведение классов

Как известно объектный подход моделирования данных предполагает описание информационной и поведенческой составляющей объектов. Информационная составляющая определяется набором атрибутов объекта, поведенческая набором методов.

В рассматриваемой системе, как хранилище данных, поведенческие функции на наш взгляд, прежде всего должны решать две задачи: представление объектов для пользователей (и разных групп пользователей) в различной форме и форматах, а также поиск и извлечение необходимой информации по запросам пользователей. В равной мере это относится к объектам хранения и атрибутам различного типа, прежде всего содержащим BLOB данные. Реализация этих функций ложится как на репозиторий, так и на интерфейсы доступа к электронному каталогу. Важнейшей средой доступа к хранилищу данных является WWW. Рассмотрим предлагаемый вариант решения этих задач на ее примере.

Простейшим типом получения информации из объектов является просмотр списка значений атрибутов. Но данный вариант годится лишь в немногих случаях, т.к. в зависимости от конкретной задачи требуется просмотр некого подмножества атрибутов, причем зачастую в достаточно специфичном виде. Тип представления информации и набор атрибутов может меняться как в зависимости от пользователя хранилища данных, так и в зависимости от задачи, которая решается при помощи хранилища. Все это приводит к необходимости создания средств, позволяющих программировать просмотр объектов. В предлагаемом варианте системы предусмотрено следующее:

- Набор стандартных средств по просмотру электронного каталога через WWW. Сюда входит просмотр определений классов, структуры каталога, просмотр объектов, поиск и навигация по электронному каталогу.

- Набор стандартных средств для работы с BLOB данными различного типа. Сюда входит просмотр изображений, проигрывание звука, видео. Просмотр документов подготовленных в разных системах.

- Создание настраиваемых шаблонов просмотра объектов и каталогов. Шаблон представляет собой HTML файл, в который включены дополнительные тэги, позволяющие манипулировать атрибутами. Такой подход позволит пользователям даже с небольшой подготовкой создавать собственные варианты просмотра содержимого электронного каталога. Обработка шаблонов осуществляется CGI скриптами или при помощи JAVA программ, которые являются частью интерфейса доступа к данным. Шаблоны просмотра объектов хранятся в репозитории, как составные части определений классов. Шаблоны просмотра каталогов хранятся в репозитории как часть описания структуры электронного каталога. При формировании шаблона доступна возможность пользоваться стандартными средствами просмотра.

- Набор вызовов API интерфейсов доступа к каталогу и репозиторию, используя которые можно создавать специализированные программные средства для доступа к информации хранилища.

Таким образом, для извлечения информации из хранилища данных пользователь может пользоваться только стандартными средствами. Вместе с тем у него существует возможность настройки форм представления и отображения данных для придания им необходимых свойств. Разработчики информационных систем на основе хранилища данных могут придавать объектам необходимые поведенческие свойства, разрабатывая приложения, основанные на вызовах API хранилища. Заметим, что отличительной особенностью такого подхода является то, что собственно выполнение методов классов в общем случае происходит не непосредственно в хранилище, а в некоторой внешней среде. (JAVA программа, CGI-скрипт, интерпретирующий HTML шаблон, и т.д.)

Аналогичным образом решена задача поиска требуемой информации. Наибольшие проблемы возникли при организации поиска по атрибутам, содержащим документы различных типов. Для каждого такого атрибута создается набор стандартных средств анализа и поиска информации в данных соответствующего формата. Сложность в данном случае состоит в том, что программирование таких средств требует высокой квалификации и специальных знаний. На данном этапе эти работы осуществляются разработчиками хранилища данных. Внешним пользователям доступны лишь API вызовы соответствующих средств. В качестве примера можно привести поиск по документу в формате Word For Windows, RTF и т.д.

6 Система авторизации доступа

Система авторизации является важнейшим компонентом любой современной информационной системы. В хранилище данных она реализована на уровне репозитория и на уровне электронного каталога.

Согласно существующим стандартам все пользователи хранилища объединены в группы и могут получать доступ к ресурсам, как в составе группы, так и в качестве пользователя.

Функции по изменению структуры данных определяются следующим образом. Для каждого пользователя определено множество классов, описание которых доступно пользователю. Это множество состоит из описаний классов доступных всем пользователям (глобальных классов) и классов, определенных самим пользователем. Для каждого глобального класса определены права модификации класса. Пользовательские классы, могут изменяться произвольным образом.

Репозиторий также содержит информацию о наборе классов, к экземплярам которых (объектам) разрешен доступ, права этого доступа. Для каждого класса из этого набора определено множество атрибутов, значения которых можно просматривать и изменять. Заметим, что пользователь может иметь доступ к объектам тех классов, описание которых ему недоступно.

На уровне электронного каталога каждый пользователь (группа пользователей) имеет права на просмотр, изменение каталога (добавление, удаление объектов), просмотр, изменение объектов (для тех классов и атрибутов, которые определены в репозитории).

Таким образом, каждый пользователь хранилища может пользоваться глобальным набором описаний классов, создавать собственные классы, в зависимости от предметной области вносимых данных, просматривать и изменять доступную ему часть электронного каталога, вносить в нее новые объекты.

7 Практическая реализация

В данный момент подобное хранилище данных реализовано в Internet центре Ярославского университета. Основу его составляет RDBMS Sybase System 11. Логика обработки электронного каталога и репозитория реализована в виде набора хранимых процедур. Доступ через WWW организован про помощи пакета Sybase Power Dynamo. Для доступа по локальной сети к хранилищу создано клиентское приложение, которое позволяет осуществлять функции администрирования пользователей, просмотра и изменения репозитория, а также ввода и просмотра объектов.

Библиография

[1] Буч Г.
Объектно-ориентированное проектирование с примерами применения.,М., Конкорд, 1992

[2] Ищмухаметов А.З.,Лукин В.В.
Организация словаря данных в предметно-ориентированных программных оболочках.,СУБД N 1-2,1998

[3] Калянов Г.Н.
CASE-структурный системный анализ.,М.,"Лори", 1996

[4] Липаев В.В.
Управление разработкой программных комплексов.,М.,Финансы и статистика, 1993

[5] Прижняковский В.В.
Абстракции в проектировании БД., СУБД N 1-2, 199

[6] Сахаров А.А.
Концепции построения и реализации информационных систем, ориентированных на анализ данных.,СУБД N4, 1996

[7] Цаленко М.Ш.
Моделирование семантики в базах данных.,М.Наука, 1989.

[8] Чень П.
Модель "Сущность - связь" - шаг к единому представлению данных.,СУБД N3, 1995