Цифровая библиотека Ярославского региона. Итоги работы, перспективы развития♣

Палей Д.Э.
ЯрГУ, Ярославль, Россия
paley@yars.free.net

Курчинский Д.Н.
ЯрГУ, Ярославль, Россия
reno@econom.uniyar.ac.ru

Смирнов В.Н.
ЯрГУ, Ярославль, Россия
smirnov@yars.free.net

Аннотация

Доклад посвящен оценке итогов разработки и перспективам развития Цифровой библиотеки Ярославского региона. Работы по ее созданию начинались при поддержке РФФИ (грант 98-07-92152) и РГНФ (грант 97-04-12016) и уже на протяжении ряда лет ведутся в Ярославском государственно университете им. П.Г. Демидова.

1. Первоначальная постановка задачи, общее описание системы.

Проект был ориентирован на решение информационных задач учреждений культуры, образования и науки. Предполагалось, что в цифровой библиотеке (ЦБ) будет описано достаточно большое количество артефактов - физических объектов (предметов, документов, картин, персон, памятников архитектуры и т.д.) реального мира. Все объекты связаны друг с другом различными отношениями. При этом одним из основных условий было то, что описание артефактов изначально имеет нечеткую структуру. Т.е. формально полное описание структуры большинства объектов неизвестно или может меняться с течением времени. Более того, описание структуры может меняться уже после занесения некоторой части данных из предметной области. То же самое можно сказать и про связи между объектами.

В силу поставленной задачи были выбраны следующие принципы построения ЦБ и организации хранения данных. Цифровая библиотека содержит метаописания артефактов в виде классов. Каждый класс определяется набором атрибутов и методов. Возможно наследование классов, объединяющее их в иерархии отношение “родитель->наследник”. Данные размещаются в экземплярах классов – объектах электронного каталога (ЭК). Объект содержит информацию о некоторой реальной сущности, согласно правилам, заданным описанием его класса. Объединение объектов электронного каталога в единую логическую структуру осуществляется при помощи связей. Связи могут иметь различный тип и определяться на различных уровнях (объект-объект, атрибут-объект, атрибут – атрибут и т.д.). Также на системном уровне поддерживается набор различных сервисов - интерфейсы доступа к данным, система авторизации и т.д. [1].

Технологической основой выполнения проекта служила RDBMS Sybase ASE 10.0, а позднее Sybase ASE 11.2.

На основе этой цифровой библиотеки был выполнен ряд проектов. В качестве примера можно привести [2,3,4]. В настоящее время цифровая библиотека региона находится в постоянном развитии и модернизации. Вместе с тем уже можно (на наш взгляд) подвести первые итоги практической реализации и наметить основные пути дальнейшего развития.

2. Логическая структура каталога.

2.1. От сети объектов к иерархии.

Специфика задачи создания ЦБ предполагала (и предполагает) наличие между объектами большого количества связей. Разработчикам очень импонировали (и до сих пор впечатляют) идеи организации хранения слабоструктурированных данных, заложенные в проекте Lore [5,6] и подобных ему. Поэтому на начальном этапе реализации проекта была выбрана следующая схема построения каталога.

Каждый объект ЭК мог иметь связь с любым другим объектом. Причем эти связи (между объектами) определялись не на уровне метаданных, а на уровне каталога объектов. Для обеспечения универсальных методов доступа к данным и уменьшения накладных расходов на поддержку ЭК представилось разумным наложить некоторые ограничения на возможные определения классов. Основное из таких ограничений - невозможность атрибутов объектного типа. Предлагалось заменить

Труды 5ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2003, Санкт-Петербург, Россия, 2003.

их отношениями объектов типа “главный-подчиненный”.

Таким образом, ЭК каталог был реализован в виде некоторой сети равноправных объектов. Точкой входа в общем случае мог служить произвольный объект сети. Несомненным преимуществом такого подхода являлся большой универсализм и широкие возможности по моделированию данных произвольной структуры. Но уже, в процессе опытной эксплуатации выяснился существенный недостаток. Несмотря на наличие разнообразных сервисов, пользователи системы испытывали значительные трудности при навигации по сети объектов. Также, весьма существенными оказались трудности по администрированию объектов и определению и модификации связей объекта при его создании или изменении его атрибутов. Особенно это стало заметно, когда количество объектов в каталоге превысило значение 500. Фактически ЭК становился неуправляем. Т.е. исходная схема, несмотря на свою универсальность, к сожалению оказалась мало пригодной для практической реализации и использования.

Это побудило разработчиков пересмотреть подходы к организации связей между объектами. Было принято решение объединить все объекты в жесткую иерархию. Электронный каталог при этом представляет собой дерево, в узлах которого находятся объекты, связанные между собой отношениями типа “главный-подчиненный”. Все остальные связи организуются при помощи ссылок, которые могут содержать атрибуты каждого объекта каталога на другие объекты и/или на их атрибуты. В таком виде ЭК каталог функционирует и до сих пор.

По итогам практического использования можно сказать, что выбранная структура каталога оказалась вполне жизнеспособной. На сегодняшний день он содержит порядка 15000 объектов и постоянно расширяется. Достоинства такого подхода достаточно очевидны и многократно подтверждены на практике. К основным из них можно отнести: логическую стройность хранения и описания данных, простоту навигации по иерархии объектов, простоту создания WWW сервисов, для доступа к структуре и объектам каталога и т.д.

Оправданным на наш взгляд также оказался отказ от атрибутов классов объектного типа. Это сильно упростило процесс моделирования данных и позволило эффективно повторно использовать уже введенную информацию.

2.2. От иерархии – к лесу объектов.

К сожалению, применяющийся подход имеет ограничения связанные с жесткой иерархией расположения объектов. Практика использования ЦБ показала, что для различных пользователей и для различных задач необходимо различное преставление данных, в виде различных иерархических структур объектов, при этом использование связей на уровне ссылок атрибутов явно недостаточно.

В качестве примера может выступать работа исследователя живописи, которому при изучении творчества художников удобно работать с деревом, в узлах которого расположены художники. От этих объектов отходят ветви – картины, периоды творчества и т.д. С другой стороны, при исследованиях художественных стилей, направлений или художественных школ пользователю необходима иерархия другой структуры: “стиль->картина->художник” или “стиль->художник->картина”. При работе с содержимым коллекция может понадобиться иерархия “коллекция->картина->художник” и т.д. Следует отметить, что в общем случае заранее неизвестно, какой тип иерархии будет необходим конечному пользователю, т.е. невозможно однозначное выделение связей “главный-подчиненный” для всех объектов каталога. Таким образом, исходная задача создания сети объектов достаточно актуальна до сих пор.

Решение этой задачи и дальнейшее развитие ЭК предполагается в проекте с помощью модификации и расширения иерархической схемы. Прелагается ввести некоторую “основную” иерархию объектов электронного каталога (фактически она сейчас уже существует). “Основанная“ иерархия будет определяться и модифицироваться администратором системы. В нее будет встраиваться каждый вновь создаваемый объект. Это позволит отделить этап ввода данных от этапа моделирования связей между объектами. Также это позволит избежать потерь информации при ошибках моделирования связей.

Вместе с тем, решено создавать “виртуальные” каталоги объектов для конкретного пользователя (групп пользователей). Т.е. каждый объект может иметь некоторый набор связей с другими объектами, которые не отражаются в “основной” иерархии и описываются при его создании или модификации. Эти связи в общем случае могут быть как произвольны, так и типизированы и организуют собственно “сеть” объектов. Представление данных для конечного пользователя определяется выбранным им из существующего набора некоторым правилом (типом связей, набором и типом исходных объектов для просмотра и т.д.), по которому из исходного набора связанных сущностей формируется иерархический каталог объектов. При этом пользователь получает дерево объектов, оптимальное для решения его конкретных задач. Отношения, не укладывающиеся в полученную иерархию, выступают в качестве дополнительных ссылок между объектами. Таким образом, появляется возможность совместить максимальную гибкость и полноту описания достаточно произвольных отношений между сущностями со всеми преимуществами хранения и обработки данных в виде иерархических структур.

Задача представления данных в виде различных иерархий будет решаться поэтапно. На первом этапе предполагается реализовать несколько стандартных иерархических представлений электронного каталога. Далее сделать возможным их динамическую генерацию по условиям, определяемым пользователем информационной системы.

3. Повторное использование данных

Важной задачей, стоявшей в начале разработки проекта, являлось обеспечение эффективного повторного использования данных, внесенных в объекты ЭК. Это связано с тем, что в ЦБ часто возникают задачи создания объектов, значения некоторого набора атрибутов которых совпадали бы со значениями соответствующих атрибутов другого объекта. Для решения этой проблемы была предложена концепция “наследования данных”. Поясним ее на простейшем примере.

Положим, есть набор объектов, являющихся экземплярами класса “персона”. Атрибуты этих объектов содержат полную информацию о персональных данных. Положим далее, что при реализации какого-либо проекта понадобилось определить классы “сотрудник”, “пользователь библиотеки”, “исследователь”, порожденные от класса “персона” и отличающиеся от него наборами дополнительных атрибутов. Положим далее, что экземпляры порожденных классов будут описывать некоторое подмножество людей уже описанных объектами класса “персона”. В этом случае для создания новых объектов разумно уже использовать имеющуюся информацию.

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.

Формально “наследование по данным” выглядит следующим образом. Пусть имеется класс типа C_a и порожденный от него класс C_b. Обозначим соответственно объекты, являющиеся экземплярами этих классов, O_{aj} , O_{bj} ($j=1..\infty$). Механизм наследования предполагает, что при объявлении некоторого экземпляра O_{bn} наследником по данным O_{am}, соответствующие атрибуты экземпляра O_{bm} будут иметь значения атрибутов экземпляра O_{an} (и соответственно изменяться при изменении атрибутов O_{an}). Отметим следующий принципиальный момент: “наследование данных” обеспечивает непосредственно ЭК, а не разработчики информационных систем на основе цифровой библиотеки и не ее пользователи.

Опыт использования функции “наследования данных” показал, что она является эффективным инструментом повторного использования уже имеющейся в каталоге объектов информации, более того в ряде случаев это сильно упрощает администрирование данных и поддержку их актуальности. Вместе с тем стало ясно, что “наследования данных” должно иметь возможность поддержки на уровне метаданных. Имеется ввиду следующее: при создании произвольного экземпляра O_{bm} класса C_b не наследующего данные от какого либо объекта класса C_a , автоматически должен создаваться экземпляр O_{aj} класса C_a . При этом O_{bm} должен наследовать данные от O_{aj}.

4. Поиск данных.

Важнейшим элементом каждой информационной системы является модуль поиска информации. На первых этапах построения ЦБ было реализовано несколько алгоритмов поиска. Основными из них является поиск по атрибутам объектов (или по всем атрибутам всех объектов), поиск по объектам заданного типа и т.д..

Эффективное решение этих задач тесно связано с физическим способом хранения данных, т.к. логическое представление информации имеет объектный вид, а в качестве сервера данных используется реляционная СУБД. Информация в нашей системе хранится подобно тому, как это описано в [7]. Такой способ хранения данных позволил в реляционной структуре построить эффективные индексы по атрибутам объектов. Также поиск осуществляется с помощью набора словарей (поисковых таблиц) по множеству значений атрибутов и/или наборов атрибутов объектов определенного типа.

В качестве дальнейшего развития поисковой машины предполагается создать систему поиска по предметным терминам. Для решения этой задачи необходимо создание справочников предметных терминов и/или различных тезаурусов. В качестве основы для них могут служить уже существующие поисковые таблицы для атрибутов объектов соответствующих классов. Следует отметить, что для выполнения этой работы необходимо участие специалистов в конкретной предметной области. Только в этом случае может быть обеспечен качественный поиск по синонимам, по связанным терминам и т.д.

С переходом от одной иерархии объектов к их множеству представляется возможным использовать в качестве основы для своеобразного тезауруса набор связей между объектами. При этом тип связи может выступать в качестве типа отношения терминов. В качестве терминов может выступать атрибут или атрибуты заданного типа (в простейшем случае это наименования объектов).

Другим направлением развития ЭК должны стать работы по реализации поиска информации “близкой” к заданному объекту. Под понятием

“близость” в данном случае понимается связь объекта (объектов) или его атрибута с другим объектом или его атрибутом через тезаурус или по определенному правилу через определенное количество связей в электронном каталоге. При этом могут учитываться связи только определенного типа, связи через объекты только заданного класса и т.д. Простейшим примером поиска близкой информации может служить поиск данных близких на один шаг к какому либо объекту класса “художник”. Результатом поиска в этом случае будет множество объектов, с которыми исходный объект связан непосредственно. Отметим, что для этого типа поиска основным “тезаурусом” выступает сам электронный каталог (совокупность объектов и связей между ними). Организация эффективного поиска на количестве шагов более одного в такой постановке задачи является, вообще говоря, нетривиальным делом. В рамках работ по гранту для решения этой задачи предполагается использовать идеи описанные в [8] по поиску в слабоструктурированной информации. Основу алгоритма такого поиска является построение индекса по связям “объект-объект” для каждого объекта каталога. Для оптимизации и снижения объема индекса предполагается анализ и выделение групп объектов (модулей), связанных небольшим количеством связей, с последующим индексированием связей между модулями и внутри модулей.

5. Заключение

Проект создания цифровой библиотеки Ярославского региона находится в постоянном развитии и модернизации. В ходе работ по созданию ЦБ накоплен достаточный опыт и определены пути дальнейшего развития системы. Это прежде всего модернизация и расширение структуры хранения объектов электронного каталога, модернизация поисковой системы, расширений функция обработки данных. Дальнейшие работы по развитию цифровой библиотеки предполагается вести в рамках выполнения проектов поддержанных РФФИ (грант 03-07-90178) и РГНФ (грант 03-04-12019в).

Литература

[1] Палей Д. Э. Объединение разнородных информационных электронных ресурсов в электронном каталоге. / Палей Д. Э., Курчинский Д. Н., Смирнов В. Н., Русаков А.Н. //Труды первой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Санк-Петербург, 19 - 22 октября 1999 г. - С. 70-74.

[2] Смирнов В. Н. Информационная система по культурному наследию Л. В. Собинова. / Смирнов В. Н., Смирнова Н. Е., Аносовская А. В. // Труды Всероссийской ежегодной научной конференции "Научный сервис в сети Интернет". г. Новороссийск, 18 - 23 сентября 2000 г. с. 150-151 (http://www.sobinov.yar.ru)

[3] Палей Д.Э. Вариант интеграции цифровой библиотеки и библиографического каталога / Палей Д.Э., Курчинский Д.Н., Смирнов В.Н. //Труды третьей Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Петрозаводск, 11 - 8 сентября 2001 г. - С. 8-14. (http://lib.yar.ru)

[4] Смирнов В. Н. Библиотека, музей, архив: создания единого информационного ресурса. Проблемы и решения. / Смирнов В.Н., Палей Д.Э., Курчинский Д.Н., Смирнова Н.Е., Грязнова Н. А. // Труды восьмой Международной конференции “Крым 2001” "Библиотеки и ассоциации в меняющемся мире: новые технологии и новые формы сотрудничества". г. Судак, 09 - 17 июня 2001 г. с. 294-298. (http://culture.yar.ru/yarmus/)

[5] J. McHugh, S. Abteboul, R. Goldman, D. Quass, J. Widom. Lore: A Database Management System for Semistructured Data. (http://www-db.stanford.edu/lore/pubs/)

[6] S. Abteboul, D. Quass, J. McHugh, J. Widom J.L. Wiener. The Lorel Query Language for Semistructured Data. (http://www.ds.standford.edu/lore)

[7] Палей Д. Э. Моделирование квазиструктурированных данных // Открытые системы. 2002 № 9.

[8] R. Goldman and J. Widom. DataGuides: Enabling Query Formulation and Optimization in Semistructured Databases. Proceedings of the Twenty-Third International Conference on Very Large Data Bases, pages 436-445, Athens, Greece, August 1997. (http://www-db.stanford.edu/lore/pubs)

Yaroslavl Region Digital Library. Work Results, Prospect Of Development

Paley D. (paley@yars.free.net),
Kurchinsky D. (reno@econom.uniyar.ac.ru),
Smirnov V. (smirnov@uniyar.ac.ru)
Yaroslavl State Universityhors

This paper is devoted to results of development and evolution of Yaroslavl Region Digital Library.

These works were begun in the Yaroslavl State University named after Demidov P.G. with financial support of RFFI (grant 98-07-92152) and with support РГНФ (grant 97-04-12016). They have been performed for several years.

Digital library design methods, which were used by authors, are analysed in this work. Different methods of digital catalogue building are considered, for example: net of objects, hierarchy of objects and forest of objects.

Idea of repeated data using is described in the paper. This method uses in digital library work. It made a good showing in the practice.

The search engine development ways are considered in the last part of this paper.

♣ Работа выполнена при поддержке РФФИ – 03-07-90178, РГНФ – 03-04-12019в.