Математические вопросы кибернетики и вычислительной техники 22, 2001, 70–79.

Анализ архитектур электронных библиотек и моделей их взаимодействия

Тигран А. Шагинян

Институт проблем информатики и автоматизации НАН РА и ЕрГУ
Email: artsha@sci.am, тел.560116

Аннотация

В работе рассматриваются основные вопросы архитектеры электронных библиотек с точки зрения обеспечения механизмов для их взаимодействия. Теоретические рассуждения дополняются примерами из решений известных современных систем. Анализированы модели реализации электронных библиотек на примере таких систем, как Интегрированная Система Информационных Ресурсов(ИСИР) Российской Академии Наук, New Zealand Digital Library, разрабатываемой в университете Вайкато, Новая Зеландия, Stanford Digital Library, Стэнфордский университет и Illinois Digital Library, разрабатываемой в университете штата Иллинойс.

1 Введение

С развитием и распространением Интернета, в начале 90-х стали появляться(в основном академические) проекты электронных библиотек. Электронные библиотеки ориентированы на большое число удаленных пользователей и поэтому, практически всегда применяют Интернет как средство связи. По сравнению с ресурсами Интернета и поисковыми системами, электронные библиотеки обеспечивают более надежные, точные, полные результаты запросов пользователей. Известно по крайней мере десять определений электронных библиотек. Ниже перечислим некоторые из них [1-3].

- Интегрированные структуры, которые предоставляют пользователям интеллектуальный и физический доступ к огромным, растущим глобальным сетям информации в мульдимедийных форматах.
- Системы, обеспечивающие сообщество пользователей когерентным доступом к большому, организованному хранилищу информации и знаний, характеризующимся отсутствием предшествующего детального знания методов использования хранимой информации.
- Набор электронных ресурсов и связанных с ними технических средств создания, поиска и использования информации. Содержание электронных библиотек включает данные, метаданные, описывающие разные свойства данных (напр., представление, создатель, владелец, право на воспроизведение), и метаданные, состоящие из ссылок или отношений к другим данным или метаданным, внутренним или внешним к электронной библиотеке.

Для наиболее полного использования возможностей электронной библиотеки необходимо исследовать вопросы ее функциональности и архитектурых решений. В настоящее время на первый план выдвигаются также вопросы расширения электронных библиотек путем разработки методов и механизмов взаимодействия между ними. Для исследования указанных вопросов, в данной работе рассматриваются:

а) концептуальные модели решений функциональных и архитектурных задач;
б) их реализации в некоторых конкретных случаях;
в) модель взаимодействия электронных библиотек.

2 Функциональность электронных библиотек

Основной функцией электронных библиотек, с точки зрения конечного пользователя, является поиск информации в распределенной, неоднородной информационной системе. Однако, электронная библиотека отличается от обычных поисковых систем тем, что ее функциональность не ограничивается простым хранением информации и предоставлением поисковых служб. Исследование функциональности электронной библиотеки проведенное учеными из стэнфордского университета, путем изучения информационных потребностей и привычек работников некоторой большой компании, вывело пять основных задач, которые должна выполнять электронная библиотека [4].

Обнаружение и выбор источника. Обычно пользователь пытается извлечь все ресурсы содержащие информацию об интересующей его проблеме. Для этого необходимо обнаружить все доступные источники, содержащие такие ресурсы, что является одним из основных задач электронной библиотеки.

Однако, источник может содержать ресурсы, созданные для иных целей и не соответствующие требованиям данного пользователя. Это накладывает на электронную библиотеку дополнительное требование хранения ресурсов с представлениями для пользователей с разными намерениями.

Поиск. После идентификации источников информация должна быть эффективно извлечена. Основной проблемой здесь является неясность вопросов, ответы на которых должна дать система поиска.

Обычно пользователи формулируют запросы на понятном для человека языке, но неудобном для реализации автоматизированного поиска. Эта проблема требует от электронной библиотеки наличие средств уточняющих(улучшающих) запросы, а также возможности поиска нетекстовых ресурсов.

Перевод(осмысление). После извлечения нужного материала из источников, наступает стадия его «осмысления», то есть выявления интересующих пользователя тенденций в множестве извлеченных информационных ресурсов.

После осмысления извлеченной информации выделяются нужные информационные ресурсы или их части и сохраняются для дальнейшего использования.

Локальное управление информацией. Материалы собранные для конкретной задачи должны быть сохранены для дальнейшего краткосрочного или долгосрочного использования.

После решения задачи пользователя, еще некоторое время могут возникать новые проблемы, для решения которых может быть использована уже собранная информация.

При долгосрочной совместной работе групп(ы) пользователей может возникать двухслойная информационная среда, где на одном слое находятся разделенные

ресурсы, а на другом локальные копии этих ресурсов. Второй (локальный) слой необходим для эффективной работы при частом использовании ресурсов. На этом слое часто образуются новые информационные ресурсы(«композиции») , составленные из отдельных частей извлеченных ресурсов и новой информации.

Сложность представляется в разработке механизмов, позволяющих пользователям составлять композиции таким образом, чтобы отдельные части оставались актуальными при изменении базовых данных, либо оставались неизменными в течении времени. В любом случае, локально составленная информационная композиция может быть действительно полезной, только если она будет разделена с другими пользователями.

Разделение. Проблема двухслойной среды состоит в том, что лучше всего поддерживаемые и обновляемые ресурсы находятся на втором(локальном) слое и распределены, но не разделены.

Сложность переноса локальных ресурсов в разделенную область заключается в том, что возникает необходимость использования разных представлений для разных пользователей.Подобно тому, как информация должна быть сохранена с возможностью ее использования в разных целях, должна быть обеспечена также и возможность ее преобразования в вид, удобный для разных пользователей.

На рис.1 показан цикл этапов функционирования электронной библиотеки, а также некоторые известные технологии используемые для реализации каждого из них.

Рис. 1. Цикл этапов функционирования электронной библиотеки, где SDI-Selective Dissemination of Information, SOAP-Seal Of APproval, OCR-Optical Character Recognition.

Согласно представленной схеме, электронная библиотека должна обеспечивать возможность свободного передвижения по представленному кругу, то есть переключение с одного этапа на другой [4].

3 Отличительные черты электронных библиотек на этапе их проектирования

Важно определить цели, для достижения которых проектируются те или иные электронные библиотеки. Ниже приведены основные цели создания рассмотренных в статье систем.

Объединение гетерогенных ресурсов. Основной целью проекта ИСИР РАН является разработка концептуальной основы и инфраструктуры для интеграции разнородных информационных и вычислительных ресурсов РАН в единое информационное пространство [5].

Борьба со сложностью. С увеличением возможностей распределенных электронных библиотек борьба с организационной и программной сложностью становится ключевым вопросом. Система должа иметь программную структуру справляющуюся с этой сложностью.

Основной целью в случае новозеландской системы является создание реально работающей, легко переносимой, легко поддерживаемой системы с автоматическими процессами [6].

Автономность частей системы. Использование компонентных технологий (например CORBA, DCOM, EJB) дает коллосальные возможности расширения и облегчения обновляемости частей системы. Такой подход является еще одним способом борьбы со сложностью системы.

Разработчики стэнфордской системы исходят с точки зрения, что электронная библиотека состоит из распределенных ресурсов, которые могут быть поддержаны автономно разными организациями и не будут требовать соблюдения единых интерфейсов [7]. Архитектура этой системы, основанная на CORBA, будет рассмотрена позже.

Автоматизация поддержки. Минимизация работы, требуемой для поддержки системы и расширения ее возможностей в новозеландской системе рассматривается, как основная цель проекта [6]. Система поддерживает в частности, автоматическое обновление коллекций и индексов без воздействий на текущие запросы.

Работа с полным текстом. Для высокой эффективности работы с информационными ресурсами со стороны конечных пользователей, работа с метаданными, описывающими ресурс может быть недостаточно. В этом случае необходимо организовать механизмы полнотекстовой работы с ресурсами. Хорошим примером реализации возможностей полнотекстовой обработки является иллинойсская электронная библиотека [8].

Перед разработчиками этого проекта были поставлены следующие наиболее важные цели: построить и протестировать мульти-издательскую (состоящую из публикаций множества изданий) электронную библиотеку, которая использует гибкие возможности поиска и подачи пользователю, и предлагает соединения к внутренним и внешним ресурсам, с источниками в формате SGML; интегрировать систему с другими полнотекстовыми хранилищами в сплошную среду (континуум) информационных ресурсов предлагаемых конечному пользователю; определить эффективность полнотекстного поиска статей по сравнению с суррогатным (по метаданным) поиском и изучить поведение конечного пользователя при полнотекстовом поиске, с целью установления его потребностей; установить модели для эффективного поиска и публикации полнотекстовых статей в среде Интернета и использовать эти модели в проектировании и разработке системы.

4 Концептульные модели для электронных библиотек

Для достижения рассмотренной выше функциональности электронной библиотеки необходимы соответствующие архитектурные решения, которые должны быть смоделированы и реализованы.

Для обоснования конкретных реализаций архитектурных решений, а также проведения их сравнительного анализа, лучше всего к этим вопросам подойти через рассмотрение концептульных моделей, которые в совокупности представляют общую архитектурную модель системы.

Рассмотрим только некоторые, наиболее важные с нашей точки зрения концептуальные модели. Прежде всего их можно разбить на две группы: моделей данных и моделей действий.

Рассмотрим группу моделей данных. В ней можно выделить модели относящиеся к онтологии (типы моделирумых объектов и их свойства), представлению (структуры данных, описывающие объекты) и хранения (методы хранения и управления записями представлений) [9].

Рассмотрим некоторые основные вопросы онтологии.

Типы ресурсов: базовые ресурсы - основной тип ресурсов, интересующий конечного пользователя напр., документы; библиотечные ресурсы - информационные структуры, необходимые для организации доступа к базовым ресурсам и управления ими, напр., списки коллекций, индексы, каталоги, базы данных; метаресурсы - ресурсы, содержащие информацию о других ресурсах, напр. о базовых ресурсах; вычислительные ресурсы — компьютеры, программы, процессы.

Модель идентификации ресурсов, напр., по содержанию, по уникальным указателям предоставляемым ресурс при поступлении в систему.

Структура ресурсов, которая может быть заранее определенная (напр., HTML), опредеяемая тэгами(напр., SGML,XML,),...

Владение ресурсом в системе, которое может иметь разные значения: право передачи доступа; ответственность за содержание; получение оплаты от пользователей ресурса.

Вопросы представления включают следующие:

Модель отношения между цифровыми представлениями и представляемыми объектами. В отличие от обычных информационно-поисковых систем, в библиотечных системах цифровое представление есть частичное описание реального объекта. Ресурсы могут быть представлены как объекты или как структурированные потоки данных(напр., текст). Ресурсы-объекты представляются их аттрибутами и методами.

Управление данными - Независимо от видов представления ресурсов должны быть выработаны механизмы для управления ими (хранение, копирование, поиск, и т.д.). Должны быть продуманы:

модель распределения — напр., единственный источник (большинство старых систем), координированная репликация(напр., Lotus Notes), множество несогласованных источников,

модель согласованности(консистентности) — абсолютная, динамическая(при обнаружении несогласованности решение принимают эвристические механизмы), квалифицированная (несогласованность рассматривается как часть системы и данные маркируются мета-информацией об их источнике , что помогает разрешить проблемы с несогласованностью),

модель управления доступом — методы распределения прав на доступ,

модель управления записями — обеспечение устойчивости(persistence), журнализация.

Модели действий можно разбить на следующие подгруппы:

Основные модели действий: *определение действий* – стандартные действия, определенные приложениями и инструментальными средствами, специализированные действия по типу объекта, *модель автоматизации*, *экономическая модель*.

Модели подготовки ресурсов, их изменения, обновления, индексации управления доступом.

Модели представления, включающие модели управления представлением, возможностей клиента, приложения, управления временем.

5 Общая архитектура

Рассмотрим варианты реализаций представленных выше концептуальных моделей на примерах некоторых известных открытых электронных библиотек.

Информационные ресурсы (документы) в новозеладской электронной библиотеке(NZDL) хранятся в коллекциях, каждая из которых предоставляет унифицированный интерфейс ко всем документам, содержащимся в ней и имеет свой набор поисковых метаданных, объявленных при создании коллекции. Документ, при попадании в коллекцию проходит автоматическую обработку - приводится в стандартный формат, определяются значения метаданных.

Обычно, несколько коллекций находятся на одном сервере, в одной директории файловой системы. Коллекции являются активными объектами, получающими и обрабатывающими сообщения, напр., поисковые запросы. Поэтому от них может быть потребовано найти другие коллекции на своем сервере, что позволяет обойтись без трудноподдерживаемых централизованных списков колекций. Добавление новой коллекции к электронной библиотеке может быть осуществлено путем включения его в директорию с уже известной коллекцией [6].

Коллекции NZDL совмещают возможности полнотекстового поиска и навигации по индексам, основанным на разных метаданных.

Основным элементом архитектуры стэнфордской электронной библиотеки (SDLP) является инфошина (Infobus), которая позволяет интегрировать хранилища данных, службы обработки информации и пользовательские интерфейсы(рис.2). Система поддерживает 10 служб обработки информации(могут быть добавлены и новые): резюмирования документов; библиографического преобразования; нахождения ресурса; он-лайн оплаты; управления правами; выборочного распространения информации; анализа результатов; перевода запроса; аннотирования документов; управления метаданными [3].

Инфошина разработана как распределенная объектная система, основанная на реализации ILU(Inter-Language Unification) Xerox PARC архитектуры CORBA.

Использование распределенных объектов приводит к намного более ясной конструкции, чем например использование CGI-bin скриптов. CORBA включает понятие интерфейса, который формально специфицирует каждый класс объектов и его методы. Спецификация интерфейса фиксирует каждый объект используя язык IDL. Тот же интерфейс может быть реализован множество раз, что имеет принципиальную важность для инфошины: каждый из LSP (прокси библиотечного сервиса) объектов является реализацией единого интерфейса, независимо от того, какого типа ресурсы или сервисы представляет данный LSP. Поскольку интерфейс ясно описывает какие методы могут использовать клиенты для взаимодействия с LSP объектами, могут существовать множество раличных клиентов.

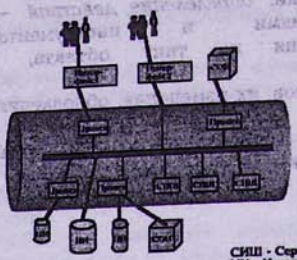

Рис. 2: Архитектура инфошины стэнфордской электронной библиотеки

6 Архитектура ресурсов

NZDL содержит документы на разных языках, включая английский, французский, немецкий, арабский, маори, португальский и суахили. Некоторые коллекции, например «Local Oral History Collection», содержат не только текст, но и графические и звуковые данные.

Исходные документы поступают в NZDL в разных форматах: ACSII, PostScript, PDF, HTML,SGML и Microsoft Word для текстовых документов; Refer, BibTex и USMarc для библиографической информации и в различных графических и звуковых форматах.

Документы непременно проходят процесс обработки, чтобы сделать их удобными для поиска и отображения на экран, что часто включает в себя конвертирование документов в другой формат и установление частей, требующих собсвенных индексов[6].

В SDLP документы моделируются как объекты. Их переменные реализации содержат поля документов, такие как, автор(*author*) и заглавие(*title*).

В IDLP поддерживаются полнотесктовые документы в формате SGML, ассоциированные с метаданными и изображениями для статей научных журналов. Использование SGML для представления структуры документа и синтеза его метаданных, что гомогенизирует гетерогенные SGML, является критическим элементом проекта.

Документы в формате SGML могут рассматриваться как представляемые, манипулируемые объекты. Превосходство SGML при поиске документов заключается в возможности выявления компонентной структуры документа.

DTD(Document Type Definition), которые сопровождают каждый SGML файл, определяют семантику и синтаксис тэгов SGML. Одной из сложнейших проблем является обработка гетерогенных DTD.

7 Хранение ресурсов и метаинформации

В NZDL – ресурсы хранятся в коллекциях, для каждой из которых есть отдельная поддиректория в общей директории на сервере:

а) *import* -для хранения исходного(импортированного) материала,

б) *archives* -GML файлы, переведенные из них,

в) *index* - окончательный вариант коллекции,

г) *building* - директория для использования во время процесса построения,

д) *etc* - директория для любых вспомогательных файлов, включая конфигурационный файл, контролирующий процедуру создания коллекции.

Для идентификации документов внутри системы, каждому документу присваивается уникальный идентификатор объекта OID (сформированный путем хэширования его содержания) при импортировании. После импортирования каждый документ хранится в своей поддиректории директории archives, вместе с ассоциированными с ним файлами, напр., изображениями [10].

Для полнотекстового поиска используется система MG[11], которая эффективно сохраняет текст и индесы в сжатом виде [6].

С каждой коллекцией связана база данных в формате GDBM(Gnu database manager), которая содержит по одному вхождению для каждого документа с его OID-идентификатором, внутренним MG номером документа и метаданными [10].

Хранилища в SDLP гетерогенны и представляются (are wrapped by) объектами прокси LSP(Library Service Proxy) [7].

В иллинойсской электронной библиотеке(IDLP) разработана архитектура расределенных хранилищ, объединяющая отдельные хранилища издателей полнотекстовых документов [8].

8 Модели взаимодейсвия между электронными библиотеками

Важнейшей задачей с точки зрения расширяемости электронной библиотеки является обеспечение механизмов взаимодействия с другими электронными библиотеками, информационными и поисковыми системами. При этом надо учитывать следующие проблемы неоднородности систем.

а) Документы хранятся в разных форматах.

б) Поиск в коллекциях осуществляется на разных языках запросов.

в) Доступ к поисковым сервисам осуществляется по несовместимым протоколам.

г) Схемы владения и доступа отличаются в разных системах .

д) Извлеченная информация возращается в разных представлениях и упорядочивается по-разному.

Разработка и реализация этой модели направлены на решение перечисленных проблем.

Одной из моделей взаимодействия является модель посредника. Посредник является автономным компонентом, который обычно выполняет следующие действия: получает запрос, переводит его для разных электронных библиотек, посылает их, получает ответы, переводит в вид, понятный для пользователя и возвращает ему.

Модель посредника, разработанная в стэнфордском университете[12] представлена ниже. В этой модели посредник состоит из компонентов, выполняющих отдельные задачи посредника, напр., перевод протоколов.

На рис.3 представлен пример работы посредника. Выбраны системы с разнымы протоколами для демонстрации работы посредника. Он принимает запросы по протоколу SDLIP(Simple Digital Library Interoperability Protocol)[13]. Один из серверов на примере предоставляет доступ к коллекциям NCSTRL по протоколу Dienst, а другой реализует протокол Z39.50. Клиент SDLIP кодирует запросы в формате XML по спецификации DASL и ожидает асинхронное получение результатов поиска. Сервер NCSTRL реализует модель запросов-ответов без состояний(stateless) и принимает URL запросы. Z39.50 принимает запросы в обратной польской записи и

возвращает число результатов, позволяя клиенту извлекать их подмножества отдельными запросами.

Упаковщики в некоторой степени решают проблему неоднородности, в частности, будучи соединенными с клиентом и серверами по их родным протоколам, они предоставляют относительно однородную среду передачи сообщений.

Посредник принимает запросы клиента, переводит их в запросы для серверов, обрабатывает полученные результаты и возвращает клиенту. Несмотря на присутствие упаковщиков, посредник выполняет основную работу: обеспечение семантической интероперабельности между несовместимыми представлениями информации, предоставляемой упаковщиками. Посредник выполняет три основных действия: перевод протоколов, запросов и данных.

9 Заключение

Анализ некоторых известных электронных библиотек, проведенной в настоящей работе, показал их принципиальные отличия, связанные с архитектурой и функциональностью.

Вопросами, имеющими принципиальное значение при разработке механизмов взаимодействия электронных библиотек являются следующие:

а) разные модели информационных ресуров,
б) разные модели метаданных,
г) разные модели поисковых запросов,
д) разные внутренние протоколы,
е) разные представления.

Был приведен пример реализации взаимодействий с помощью посредника, обеспечивающего интероперабельность на уровне протоколов, языков запросов и представлений данных. Однако, несомненно могут быть разработаны механизмы обеспечивающие взаимодействие также на других уровнях, при которых необходимо учитывать более детальные архитектурные решения разных систем.

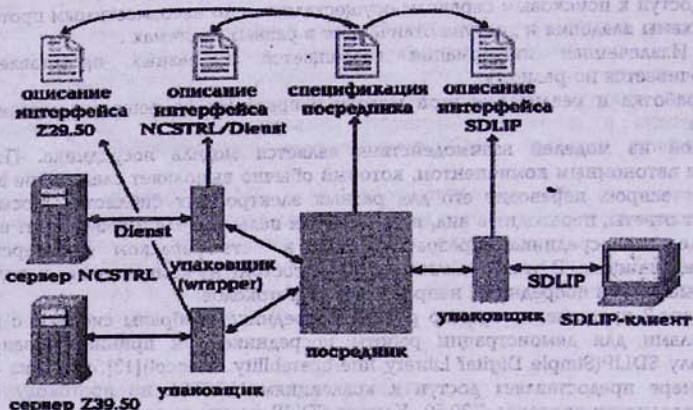

Рис. 3. Пример работы посредника

Литература

[1] E. Fox, «Digital library definitions», 1998, ei.cs.vt.edu/~fox/dlib/def.html
[2] Clifford Lynch, Hector Garcia-Molina, Interoperability, Scaling, and the Digital Libraries Research Agenda,A Report on the May 18-19, 1995, IITA Digital Libraries Workshop,diglib.stanford.edu/diglib/pub/reports/iita-dlw/main.html
[3] UCLA-NSF Social Aspects of Digital Libraries Workshop, www-lis.gseis.ucla.edu/DL/UCLA_DL_Report.html
[4] Andreas Paecke «Digital Libraries: Searching Is Not Enough» D-Lib Magazine, May 1996
[5] С.В. Агошков, А.Н. Бездушный, М.П. Галочкин, М.В. Кулагин, А.М. Меденников, В.А. Серебряков, «Интегрированная Система Информационных Ресурсов(ИСИР) — подход к созданию интегрированных цифровых библиотек», международная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции», Санкт-Петербург, 1999
[6] Ian H. Witten, Rodger J. McNab, Steve Jones, Mark Apperley, David Bainbride, Sally Jo Cunningham «Managing Complexity in a Distributed Digital Library» IEEE Computer 1999-32(2)
[7] Andreas Paepcke, Michelle Q. Baldonado, Chen-Chuan K. Chang, Steve Cousins, Hector Garcia-Molina "Using Distributed Objects to Build the Stanford Digital Library Infobus" IEEE Computer 1999-32(2)
[8] Bruce Schatz, William Mischo, Timothy Cole, Ann Bishop, Susan Harum, Eric Johnson, Laura Neumann, Hsichun Chen, Dorbin Ng « Federated Search of Scientific Literature» IEEE Computer 1999-32(2)
[9] Terry Winograd «Conceptual models for comparison of digital library systems and approaches», www-diglib.stanford.edu/cgi-bin/get/SIDL-1995-0001
[10] Ian H. Witten, Rodger J. McNab, Stefan J. Boddie, David Bainbridge, «Greenstone: A Comprehensive Open-Source Digital Library Software System», ACM Digital Libraries 2000,113-121
[11] I.H. Witten, A. Moffat, and T. Bell, «Managing Gygabytes: compressing and indexing documents and images», Morgan Kaufmann, second edition,1999.
[12] Sergey Melnik, Hector Garcia-Molina, Andreas Paepcke «A Mediation Infrastructure for Digital Libraries», ACM Digital Libraries 2000,123-132
[13] Simple Digital Library Interoperability Protocol, www.diglib.stanford.edu/~testbed/doc2/SDLIP,1999
[14] B.M. Leiner, «The NCSTRL Approach to Open Architecture for the Confederated Digital Library», D-Lib Magazine, December 1998.

Էլեկտրոնային գրադարանների ճարտարապետությունների եւ նրանց փոխազդեցության մոդելների վերլուծություն

Տ.Ա. Շահինյան

Ամփոփում

Կատարված է էլեկտրոնային գրադարանների և նրանց ճարտարապետության հիմնական հարցերի հարաբերական վերլուծություն նրանց փոխազդեցության մեխանիզմների ապահովման տեսակետից: Հիմնական հարցերի համար բերված են իրագործման օրինակներ: