Общая характеристика работы

Мировое сообщество обладает огромным объемом информации, лишь часть которой имеется в Интернет. Интернет развивается фантастическими темпами, развивается как экстенсивно, так и интенсивно. Все больше и больше видов человеческой деятельности включается в среду Интернет. Побочным эффектом этого расширения становится то, использование всего богатства имеющихся источников информации сопряжено с проблемами эффективного обнаружения требуемой информации. Это обусловлено и возрастанием объема информации, и тем, что информация слабо упорядочена, постоянно изменяется как она сама, так и ее положение. Это связано со способами выбора того, что следует индексировать, как обеспечить единое информационное пространство, с проблемами определения, в контексте каких запросов следует выдавать ту или иную информацию.

Специализированные системы, осуществляющие доступ к предметно-ориентированным ресурсам, обеспечивают значительно более высокое качество результатов, получаемых пользователями, предоставляют результаты поиска в удобной для чтения форме, тогда как результаты поиска в Интернет системах, представляются как “сырые” данные. Это определяется природой процесса каталогизации и индексации информации. Специализированные системы каталогизируют отобранные подготовленные информационные ресурсы, индексируют их, используя специально подготовленные описания ресурсов (метаданные), а Интернет системы осуществляют автоматическую индексацию Web-ресурсов на основе слов документов. Попытки централизованно создать каталоги ресурсов Интернет, использовать средства самокаталогизации пока не дали приемлемого результата.

Поскольку ни одна из поисковых машин не индексирует все содержимое Интернет, пользователи вынуждены обращаться к разным источникам информации, содержание которых скрывается за разнообразными интерфейсами и схемами запросов. Пользователь не имеет возможности исследовать “связанный контекст“ обнаруженного ресурса в ходе навигации в этом контексте. Он лишен возможности, которая позволяет находить связанную информацию с наибольшей полнотой и точностью при наименьших усилиях, для которой требуется осуществить *интеграцию* информационных ресурсов, то есть объединить их с целью обеспечения восприятия различной информации пользователями как единого информационного пространства.

Актуальность работы.

В силу сложности Интернет и его высокой динамичности, приводящих к вышеуказанным проблемам, необходимо создавать и использовать всевозможные средства, ограничивающие и систематизирующие информационную анархию в Интернет, облегчающие поиск необходимых ресурсов, делающие поиск значительно более управляемым, предметным и содержательным. Вышеизложенные проблемы особенно актуальны для научного информационного пространства, где источники информации, отличаются качеством и полнотой представления информации, где имеется острая необходимость в создании единого информационного

пространства - в объединении разрозненных научных данных в концептуально одну информационную систему, в обеспечении централизованного доступа к результатам научной деятельности, имеющим распределенный характер.

Необходимо решить вопросы, связанные с тем, как объединить разрозненные данные в концептуально одну информационную систему, как организовать работу с источниками информации и обработать запросы так, чтобы помочь пользователям находить и использовать информацию, в которой они нуждаются. Работа рассматривает проблемы интеграции существующих и вновь создаваемых информационных и вычислительных ресурсов вне зависимости от числа источников, проблемы обеспечения развитие информационной системы как единой структуры.

Все больше и больше пользователей согласны готовить сложные поисковые запросы, чтобы быстрее и точнее получить необходимые данные. Из-за огромных объемов информации невозможно обслужить запросы простыми способами, обращаясь ко всем известным источникам. Запросы пользователей могут потребовать доступа к источникам информации, существенно отличающимся по типам информационных объектов, которые они содержат, по схемам запросов и интерфейсам, которые они представляют пользователям. Одни источники содержат только текстовые документы и поддерживают простые модели запроса, например, только список ключевых слов. Другие содержат структурированные данные и поддерживают запросы в стиле реляционных интерфейсов базы данных. Следовательно, необходимо, имея дело с гетерогенными источниками информации, обеспечить унифицированные интерфейсы к совокупности источников информации и поисковых систем, дав пользователям иллюзию одного объединенного источника информации. Способность объединять ресурсы, разработанные независимо друг от друга является существенным свойством выживания технологии в распределенной информационной системе.

В связи с вышесказанным значительный интерес представляет разработка и реализация распределенной информационной системы интегрированных ресурсов, ключевыми направлениями которой являются

- интеграция разнотипных ресурсов и систем,
- идентификации ресурсов,
- использование метаданных,
- применение открытых стандартов взаимодействия систем, поиска, обмена и представления данных.

Цель диссертационной работы заключается в разработке принципов проектирования цифровых библиотек, позволяющих эффективно реализовывать “хранилища” разнородных ресурсов, обеспечить их интеграцию в единое информационное пространство. Этот базис должен обеспечить объединение в единое пространство всевозможных цифровых библиотек, информационных и вычислительных систем, основывающихся как собственных принципах организации, так и на предлагаемой архитектуре. В соответствии с этим в работе исследуются возможности и особенности существующих распределенных информационных систем, цифровых библиотек и предлагаются концептуальная модель цифровых библиотек

и соответствующая ей архитектура открытой распределенной информационной системы интегрированных ресурсов (ИСИР). Гибкая организация информации, ее интегрированное представление, открытая архитектура системы являлись ключевыми моментами в проектировании системы ИСИР для информационных и вычислительных ресурсов РАН, построенной на основе этих принципов. Реализация системы велась в соответствии со следующими требованиями:

- Логическая группировка данных – система должна позволять обрабатывать все запросы на логических группах баз данных, полностью скрывая тем самым физическое расположение последних.
- Абстрактная модель данных – информационная система должна строиться на основе абстрактной схемы данных, на которую должны быть отображены конкретные базы данных, что позволяет объединять данные из разнородных систем в одной логической группе.
- Абстрактная система запросов – система должна оперировать не конкретным синтаксисом запросов, а его логической сутью на основе абстрактных ресурсов и их атрибутов.
- Метаинформация – система должна владеть полной информацией о себе и обо всех своих ресурсах.
- Работа с распределенными данными – информационная система должна допускать возможность работы с данными, расположенными на разных физических серверах, различных аппаратно-программных платформах.
- Связь с другими системами – возможность системы интегрировать свои ресурсы с ресурсами одних информационных систем и взаимодействовать с другими при осуществлении поиска информации.
- Открытость – система должна легко расширяться и быть основана на открытых стандартах и протоколах.
- Разграничение доступа – система должна быть способна предоставлять различные уровни доступа к информации для различных пользователей.
- Легкость в общении – для пользователей система должна предоставлять простые, удобные интерфейсы поиска и доступа к информации, важнейшим среди которых является WEB-интерфейс.

Научная новизна работы связана с разработкой методики формирования и интеграции разнородных информационных и вычислительных ресурсов в единое информационное пространство, на основе которой предложена открытая архитектура информационной системы интегрированных ресурсов. Разработанная архитектура позволяет настраиваться на требуемую предметную область цифровой библиотеки. Предложена технология декларативной разработки интерактивных интегрированных с СУБД Web-приложений, поведение которых может зависеть от данных БД. Технология может быть интегрирована в существующие Web-системы.

Практическая значимость. На основе предложенных технологий реализована интегрированная информационная система РАН, позволяющая объединить

информационные ресурсы РАН в единое информационное пространство, к которой можно обратиться по адресу ‘http://isir.ras.ru/’. Система поддерживает такие ресурсы, как персона, организация, подразделение, публикация и проект. Система является многоязычной. Для поддержки иерархические связей, рубрикации ресурсов в системе реализованы специальные механизмы навигации, учитывающие специфику соответствующих типов ресурсов и настраиваемые по их описаниям. В реализации системы использован пакет программных средств декларативной разработки интерактивных Web-приложений, созданный в рамках диссертационной работы. Реализованные пакеты могут быть использованы в других системах для публикации в Интернет разнообразных информационных ресурсов.

Апробации работы. Научные результаты и основные положения работы докладывались на международном симпозиуме “Software technology” FUSST’99 (Tallin, Estonia, 1999), на 8-м научно-практическом семинаре “Современные технологии в информационно-библиотечном обеспечении научных исследований” (Тарусa, 1999), на международных семинарах проекта ESPRIT (EIS in CCE/NIS) (Vien, Budapest, 1999), на Всероссийской научной конференции "Научный сервис в сети Интернет" (Новороссийск,1999) и на международной научной конференции “Электронные библиотеки: перспективные методы и технологии, электронные коллекции” (Санкт-Петербург, 1999).

Публикации. По теме диссертации и выполненных исследований опубликовано 4 печатные работы.

Структура и объем работы. Диссертация состоит из введения, пяти глав, заключения, списка литературы, включающего 103 наименования, и 3 приложений. Общий объем работы - 150 страниц.

Краткое содержание работы.

Введение посвящено постановке проблемы исследований, обоснованию актуальности темы диссертации, новизны полученных результатов. Приводится общая структура работы.

Первая глава содержит обзор различных способов публикации информации, ее поиска и извлечения, организации систем их поддерживающих. Рассматриваемые способы имеют свои сильные и слабые стороны.

Основой большинства видов индексации и каталогизации в Интернет служат просто слова документов. Это могут быть все слова документа или некоторым образом отмеченные слова, например, слова из оглавлений, выделенные в документе фраз. С другой стороны, специализированные информационные системы для обеспечения автоматизации обработки информационных ресурсов используют не сами ресурсы, а некоторые их описания, которые представляют собой наборы именованных значений (*свойство*, *значение*), существенных для обработки этих ресурсов. Такие описания называют метаданными ресурсов, а иногда утверждениями о свойствах ресурсов (*ресурс* имеет *свойство* со *значением*). В разделе 1.1 анализируются несколько форматов метаданных с тем, чтобы сравнить их свойства, области применимости, определить текущие подходы к выбору метаданных,

рассмотреть технологии их применения и расширения. Основное внимание уделяется форматам, которые не требует серьезного предварительного обучения, как самому формату, так и правилам работы (RFC 1807, IAFA, SOIF, DC) и могут быть использованы неспециалистами. Рассматривается методика RDF, предлагаемая W3C в качестве стандартного базиса для определения и обработки метаданных Интернет ресурсов. RDF определяет удобный механизм описания ресурсов, не делающий никаких предположений относительно специфики предметной области. Он позволяет формулировать утверждения не только о свойствах ресурсах, но и о самих утверждениях. Семантика в RDF описаниях задается с помощью RDF схем, определяющих семантику свойств, позволяющих квалифицировать значения свойств, используя механизм формирования типов. Это дает возможность осуществить автоматическую обработку метаданных в формате RDF.

В разделе 1.2 рассматриваются подходы к организации распределенных сред информационных ресурсов. Файловые системы с глобальной областью действия (Andrew File System, Prospero, ALEX, Archie, Gopher) предоставляют мощные возможности для совместного использования больших совокупностей распределенных файлов, для организации коллекций информации, для ведения совместных работ. Простота интерфейсов файловых систем способствовала их широкому использованию. Эти системы предлагают только довольно примитивные средства управление данными, задания связей между данными, определения операций и т.п. Однако, свойства, лежащие в основе реализации файловых систем, такие, как кэширование, репликация, обеспечивают необходимую эффективность и могут послужить основой для реализации распределенных информационно-поисковых систем, реализуя обмен данными между серверами одной системы или разных систем.

Современные поисковые Интернет системы поддерживают автоматическое исследование пространства Интернет, используя программы ("роботы"/"пауки"), которые "обходят" Интернет, анализируя и/или сохраняя встречающиеся документы. Эти программы потенциально могут посетить и проиндексировать весь Интернет. Они используются для различных целей - находят устаревшие ссылки, обнаруживают новые серверы, оценивают размер сети, извлекают информацию из баз данных. Однако и этот подход имеет свои проблемы. Роботы, извлекая множество документов, значительно загружают сеть. Они находят все достижимые документы даже те, которые не следует индексировать. В результате каталог поискового сервера заполняется ненужными сведениями, становится очень большим, теряется структура документов. Ещё хуже обстоит дело с серверами, в которых страницы генерируются динамически в соответствии с запросами пользователей. При индексировании содержания таких серверов очень сложно выбрать все варианты генерируемых страниц. Если в параметрах запроса сохраняется контекст навигации, то будет формироваться огромное количество одинаковых страниц, но имеющих разные URL адреса.

Распределенные информационно-поисковые системы, именуемые цифровыми библиотеками (ALIWEB, Z39.50, WAIS, Harvest, Dienst, ROADS, Isaac и

FEDORA), предоставляют средства поиска и сопровождения ресурсов на основе метаданных, дают возможность поддерживать коллекции распределенных ресурсов. Концептуально они состоят из трех частей: репозитория метаданных, хранящего и предоставляющего доступ к ресурсам, индексной службы, передающей между узлами индексную информацию на основе метаданных, и службы поиска, с которой взаимодействуют пользователи. Системы существенно различаются по своей ориентации, наборам предоставляемых услуг, используют разнообразные протоколы взаимодействия. Они имеют дело только с так называемыми документо-подобными ресурсами, используют метаданные, описывающие и идентифицирующие только эти ресурсы, не отождествляя другие виды важных данных, например, персональные данные, сведения об организациях и т.п. В итоге, встретив упоминание персоны, невозможно идентифицировать ее, нельзя запросить документы, связанные именно с ней и т.д.

Раздел 1.3 посвящен рассмотрению способов интеграции Web-технологий с СУБД, возможностям организации цифровых библиотек на основе этих современных и эффективных механизмов управления данными. Интернет технологии предоставляют богатый набор удобных средства реализации распределенных информационных систем. С другой стороны СУБД предоставляют развитые и эффективные средства хранения, поиска и манипулирования данными. Современные Web-сервера, Интернет сервера приложений (Apache, Microsoft Internet Information Server, NetWare WebServer, Netscape Enterprise/FastTrack Servers, Oracle Web Server), обеспечивающие интеграцию Интернет технологий и СУБД, обладают собственными средствами разработки интерактивных Web-приложений (MS ASP, Apache PHP3, Netscape LiveWire, язык Perl), управляющих данными БД. Однако, отсутствует средство публикации данных БД, поддерживаемое всеми серверами. Все технологии одинаково требуют программирования вне зависимости от сложности логики приложений, хотя в большинстве случаев достаточно и желательно иметь высокоуровневый декларативный способ описания Web-приложений, использующих БД.

Вторая глава посвящена описанию и обоснованию предлагаемой концепция интегрированной цифровой библиотеки, принципам ее организации. Даются определения основных терминов и понятий цифровых библиотек.

В цифровой библиотеке логическую единицу хранения называют документом или ресурсом, иногда, цифровым объектом. Все множество единиц хранения библиотеки именуют репозиторием. Аналогично традиционным библиотекам, цифровые библиотеки, как правило, имеют дело с одним видом информационных ресурсов - документами или их обобщением, называемым документо-подобными объектами. Мы будем использовать термин *ресурс* для обозначения единицы, хранящейся в репозитории, а если речь будет идти о таком состоянии ресурса, в котором он обладает определенной активностью, то будем применять термин *объект-ресурса* или просто объект. Ресурс имеет две части - *метаданные* и, возможно, данные (*содержание* ресурса). Традиционные определения цифровой библиотеки считают метаданные центральным понятием, центральным звеном в работе с ре-

сурсами. Метаданные - это специально подготовленные, машинно-интерпретируемые, структурированные *сведения* о ресурсе, представляющие *свойства*, которые имеет ресурс, *услуги*, которые предоставляет ресурс. Это данные, на основе которых цифровая библиотека осуществляет поиск ресурсов, вывод результатов поиска, управление ресурсами, взаимодействие с ними и т.п. К метаданным относят и *уникальный идентификатор* ресурса, отличающий его от любых других ресурсов в соответствующей сетевой среде.

Анализируя форматы метаданных, используемые традиционными библиотеками, можно видеть, что их элементы имеют разную природу - одна часть описывают внутренние свойства ресурса, а другая с помощью примерного описания характеризуют другие ресурсы, связанные с рассматриваемым ресурсом. Всегда имеется несколько видов ресурсов, тесно связанных друг с другом, что обусловлено характеристиками одних другими или действиями одних на другие, например, `персона` является автором `публикации`, которую опубликовала издательская `организация`. В работе делается вывод о необходимости хранения и работы с метаданными не как с совокупностью сведений об одном виде ресурсов, а как с совокупностью сведений о *множестве взаимосвязанных разнотипных* ресурсов. Поскольку ресурсы неоднородные, то это должно проявляется в различии их свойств, предоставлении услуг. Каждый ресурс относится к некоторому *типу ресурсов*, который фиксирует это различие и определяет некоторую совокупность сведений, внутренне присущих ресурсам этого типа, но не рассматривающих взаимосвязи с другими ресурсами. Как образ некоторой сущности реального мира, которую он представляет, тип ресурсов косвенным образом характеризует множество потенциальных связей ресурсов этого типа с другими ресурсами, в частности, этого же типа. Определяемая типом ресурсов совокупность сведений является своеобразным форматом метаданных, отличие которого от традиционных форматов заключается в исключительной ориентации на сведения одного ресурса.

Разделим определяемую типом ресурсов совокупность сведений на два класса так, чтобы один представлял как бы “статические”, а другой “динамические” сведения о ресурсах. Первые будем именовать *атрибутами* ресурса. Они представляют то, что может использоваться пользователями библиотеки при поиске ресурсов (поисковые атрибуты), что должно показываться пользователям при просмотре описаний найденных ресурсов (просмотровые атрибуты), чтобы пользователи смогли определить, является ли рассматриваемый ресурс искомым или потенциально таковым, и т.п. Вторые будем называть *аспектами* ресурса. Это то, как может использоваться ресурс, в чем состоит работа с ним. Это сведения, характеризующие поведение ресурсов или библиотечной системы при работе с ресурсами этого типа. Это предоставляемые ресурсами услуги, способы организации поиска, представления информации, взаимодействия с ресурсами и т.д. Между этими двумя классами сведений нет ясной границы. Многие сведения можно отнести как к первому, так и ко второму классу. Четкое разделение возникает при выборе предметной области конкретной библиотеки и её предпочтений. Как правило, аспекты ресурса выражаются не функциями, а с помощью данных - параметров,

определяющих поведение соответствующих функций системы. Например, формат представления публикации (ps, pdf, sgml) может быть отнесен к аспектам, поскольку он определяет, какие действия должна исполнить система, показывая содержание ресурса.

Уникальный идентификатор ресурса не будет относить ни к одному из этих классов. Будем считать, что он является свойством иного рода, имеющимся у всех ресурсов, поскольку идентификатор ресурса, аналогично типу ресурса, требует иной поддержки, чем атрибуты и аспекты.

Описание взаимосвязей ресурсов, то есть сведений, которые затрагивают не один, а несколько ресурсов, переведем в отдельную категорию характеристик ресурсов. Необходимо обеспечить им соответствующую поддержку, чтобы из средства упоминания других ресурсов получить средство интеграции разнотипных ресурсов. Среди взаимосвязей ресурсов ключевыми являются те, что являются характерными для всех ресурсов некоторых типов. Назовем их *отношениями* между типами ресурсов. Связи, наличие которых определяется не типами ресурсов, а конкретным содержанием ресурсов, достаточно редки и специфичны. В случае необходимости их можно реализовать через программные интерфейсы, представляемые аспектами взаимодействия с ресурсом. Поддержка только отношений между типами ресурсов позволяет абстрагироваться от деталей отдельных ресурсов, их специфических связей, устранить влияние, которое могут оказать различные программные средства, обслуживающие ресурсы.

Разложение понятия метаданных, имеющего очень абстрактное определение “данные о данных”, на системы типов ресурсов и отношений между ними, введение уникальной идентификации ресурсов всех типов, а не только документов, существенно расширяет возможности цифровой библиотеки по предоставлению информационных сервисов. Осуществив необходимую обработку метаданных традиционных форматов, при том же самом объеме информации можно выполнять более точные и более сложные поисковые операции, получать в качестве результата поиска сведения не только о публикациях, но и о связанных с ними ресурсах, много другой информации, в частности, аналитического характера. Например, по автору публикации можно узнать именно его публикации, издательства, в которых он публиковался. Включение в это множество типов ресурсов, связанных с научными публикациями, таких, как проекты, конкурсы, гранты, конференции, спонсоры, организации, существенно расширяет возможности информационной системы. Например, по автору публикации можно узнать проекты, в каких он принимал участие; можно посмотреть материалы этих проектов; можно познакомиться с материалами тех секций конференций, на которых докладывались коллеги автора по проекту и т.п.

Поскольку основная цель цифровой библиотеки состоит в обеспечении распределенного поиска информации, в поддержке навигации по информационному пространству, в которое она входит, то существенным является объем метаданных, которыми должен обладать узел распределенной системы, чтобы обеспечить поддержку отношений между собственными ресурсами и ресурсами других узлов.

Поэтому объектами отношений между типами ресурсов, следует сделать часть метаданных ресурсов, а именно, ту часть, которая ориентирована на обеспечение поиска ресурсов (*поисковые атрибуты*). Это упрощает и делает более эффективным распределенный поиск. Содержание ресурса должен оставаться для поисковой части системы "черным ящиком", правила управления, и взаимодействия с которым определяются оставшимися метаданными.

С позиции распределенной системы важным является разделение отношений между типами ресурсов на *основные* и *второстепенные* отношения. Ими, соответственно, являются отношения, которые будут использоваться большинством пользователей, и отношения, которые существенны только для определенного рода специалистов, например, библиотекарей, причем, скорее всего, в рамках отдельного узла распределенной системы или небольшой группы узлов. Например, отношение *автор* между *персоной* и *публикацией* следует отнести к первому типу, а отношение *редактор* между этими же типами ресурсов - ко второму. Основные отношения должны иметь эффективную реализацию даже за счет некоторых ограничений, например, можно разрешить определять такие отношения только при генерации или модификации системы. Для вспомогательных отношений характерной должна быть возможность свободно вводить новые типы отношений, удалять старые, возможно, за счет некоторого понижения производительности при работе с ними.

Изложенная концептуальная модель, основывающаяся на понятиях типа ресурса, определяющего состав и структуру сведений, отношения между типами ресурсов и уникальной идентификации ресурсов позволяет:

- обеспечить эффективное связывание ресурсов в распределенной среде, предоставить набор стандартных механизмов связывания ресурсов;
- структурировать информационное пространство, организуя ресурсы иерархические, сетевые структуры, обеспечить предметную навигацию в нем;
- параметризовать и упростить реализацию операций поиска и навигации, структурировать предоставляемые услуги;
- обеспечить естественную и удобную навигацию в контексте найденного ресурса, упростить сопровождение и создание ресурсов.

В третьей главе описывается инфраструктура и архитектура информационной системы интегрированных ресурсов (ИСИР), в основе которой лежит вышеизложенная концептуальная модель, исходящая из того, что ресурс характеризуется набором присущих ему атрибутов и совокупностью взаимоотношений с другими ресурсами. Это позволяет воспользоваться хорошо себя зарекомендовавшими принципами информационного моделирования. В рамках этой модели ресурсы следует рассматривать как *сущности уровня модели данных бизнеса*, то есть сущности, которые составляют скелет информационной модели, вокруг которых сконцентрирована логика приложения. Ресурсы не должны соответствовать сущностям системной модели данных, ориентируемым на поддержку функциональности конкретных приложений, вносящих в модель данных массу зависящих от при-

ложений деталей, девуалирующие информационные свойства предметной области. Атрибуты ресурсов, которые могут иметь сложную структуру, должны описывать внутренние свойства ресурсов, а не детали предметной области, для выражения которых в информационном моделировании используются сущности системной модели данных. Описывая предметную область цифровой библиотеки, следует думать только о том, что характеризует ресурс, что должно использоваться при работе с ресурсом. То в чем состоит работа с ресурсом, как он будет использоваться, фиксируется самим понятием цифровой библиотеки, в базовый набор функциональности которой входят такие операции, как хранение, поиск, просмотр, извлечение, обмен, загрузка, выгрузка и, возможно, редактирование.

Уменьшая возможности по определению свойств ресурсов, предоставляя высокоуровневые понятия, предполагая в связи с этим введение соответствующих ограничений на данные, мы упрощаем и облегчаем описание информационной модели предметной области библиотеки. Зная, какого рода функциональностью должно обладать приложение, можно автоматизировать введение в модель данных сущностей системной модели, необходимых для некоторой стандартной реализации этой функциональности цифровой библиотеки. Следовательно, по высокоуровневому описанию предметной области библиотеки можно обеспечить автоматическое построение и корректной схемы базы данных, и приложения, выполняющего предопределенные операции с ней. Чтобы иметь определенную свободу управления функциональность библиотеки, можно осуществить некоторую параметризацию операций библиотеки, а в описание предметной области библиотеки включить средства управления этими параметрами функций.

Способ реализации функций цифровой библиотеки зависит от их вида. Чтобы осуществить эффективную реализацию функций, выполняющих *групповые* операции над ресурсами (загрузка, выгрузка, поиск, вывод результатов поиска), необходимо реализовывать их с помощью системных компонент библиотеки, базирующихся на соответствующих современных технологиях. В качестве хранилища ресурсов следует использовать современные реляционные БД, позволяющие обеспечить эффективный доступ к данным, целостность и защиту данных, упрощающие управление хранилищами. Функции, поведение которых зависит от *конкретного* ресурса (просмотр, редактирование), следует реализовывать с использованием объектно-ориентированного подхода, предоставляющего наиболее адекватные средства для моделирования поведения объектов реального мира и поддерживаемого современными технологиями. Для этого каждому типу ресурса сопоставляется соответствующий ему класс *объектов-ресурсов*, методы которого обеспечивают обслуживание соответствующих ресурсов. Все классы объектов-ресурсов должны быть производными от предопределенного корневого класса *Resource*. Виртуальные методы этого класса должны иметь базовую реализацию всех сервисов, требуемых от объектов-ресурсов стандартными функциями цифровой библиотеки. Чтобы исключить необходимость программирования при введении каждого нового ресурса, эти реализации должны быть параметризованы некоторым описанием ресурса, то есть они должны уметь настраиваться на реализацию

стандартной функциональности, соответствующей некоторому типу ресурса, по его описанию. При этом необходимо учитывать то, что данные ресурсов имеют реляционное представление, а поведение объектов-ресурсов реализуется в объектной концепции, поэтому возникает необходимость в преобразовании из реляционного представления в объектное и обратно. В рассматриваемом случае высокоуровневое описание информационной модели, позволяющее иметь сущности с атрибутами не только элементарных типов, и фиксированная функциональность приложения позволяют средствами системных компонент библиотеки эффективно осуществлять активизацию объектов-ресурсов, преобразование данным между моделями, поддержку деятельности объектов-ресурсов, их взаимодействия.

Такая архитектура системы позволяет автоматизировать создание цифровых библиотек. Информацию о типах ресурсов системы, их атрибутах и аспектах, отношениях между типами ресурсов можно сохранять в метарепозитории системы. Это отдельная схема (подсхема) базы данных, содержащая описание информационной модели предметной области, параметров настройки стандартных функций библиотеки, в которой может храниться описание как одной, так и нескольких предметных областей. По информации из метарепозитория осуществляется генерация и модификация схемы базы данных цифровой библиотеки, заполнение служебной базы данных, хранящей данные, обеспечивающие поддержку стандартных функций библиотеки, динамически определяемые отношения между ресурсами и т.п.

Для описания информационной модели цифровой библиотеки используются специальные конструкторы типов ENTITY и RESOURCE. Вводимые ими типы определяют последовательности пар: имя атрибута объекта данных и тип его значений. Эти последовательности характеризуются разными информационными и функциональными свойствами. Атрибуты объектов данных могут быть атрибутами с одним значением и атрибутами с множеством значений (rfbr[]: RFBR_SPECIALITY), число которых обычно не ограничивается, но может быть ограничено с помощью перечислимого типа (title[LANG]: TITLE). Конструкторы типов помещены в такие рамки, что практически позволяют осуществлять ER-моделирование, используя язык спецификаций. Квалификаторы атрибутов объектов данных управляют обязательностью связей.

Для обеспечения поиска ресурсов и обмена данными между узлами распределенной цифровой библиотеки предлагается использовать два вида языков запросов. Языки не привязываются к конкретному набору типов ресурсов и отношений между ними. Настройка языков на ту или иную предметную область, конвертирование поисковых запросов в команды языка SQL должна осуществляться на основе описания предметной области, хранящегося в метарепозитории. Один язык запросов представляет некоторую форму языка SQL, в которой вместо понятий таблиц, ключей и элементарных данных используются понятия ресурса, отношения между ресурсами, составных типов данных и способов выделения элементов составных типов данных. Второй язык запросов предназначен для формулировки поисковых запросов. В нем выделены понятия двух видов подвыражений. Первые

(F-выражения) задают ограничения на атрибуты ресурсов (фильтрует ресурсы некоторого типа), а вторые (J-выражения) специфицируют условия на наличие отношений между ресурсами (соединяет ресурсы).

- F-выражение позволяют описать условия, которым должны удовлетворять атрибуты ресурса определенного типа, и объединить эти условия операциями OR, AND и NOT.
- J-выражение - это выражение, содержащие условия, объединяемые операциями OR, AND и ANDNOT, требующие наличия или отсутствия отношений между ресурсом, представляемым F-выражением, и другими ресурсами, указываемыми J- или F-выражениями. Чтобы иметь возможность несколько раз сослаться на ресурс, удовлетворяющий некоторому выражению, при указании его связей с другими ресурсами, с выражением можно связать идентификатор и использовать его в дальнейшем в формулировке запроса.

Строгость базовой формы языка обеспечивают эффективную обработку запросов, позволяет создать единый механизм доступа к ресурсам, используемый разнообразными поисковыми протоколами системы. Наряду со строгостью языка, требующей указания того, какой атрибут должен принимать то или иное значение, обязывающей явно разделять ограничения на атрибуты и отношения, в поисковом языке имеется ряд правил, дающих пользователям возможность достаточно просто формулировать сложные поисковые запросы и позволяющих охватить большинство возможностей, предлагаемых современными поисковыми системами.

Это вид языка запросов используется в качестве внутреннего представления поисковых запросов, обеспечивающего управляемый перевод поисковых запросов разнообразных протоколов в последовательность SQL запросов. Такой вид реализации, кроме унификации поддержки разных видов поиска и серверов баз данных, необходим для того, чтобы обеспечить приемлемое преобразование поисковых запросов. Преобразование исходных поисковых запросов во внутренний формат, генерация SQL запросов производится на основе метаинформации из метарепозитория.

В архитектуре ИСИР ключевыми понятиями организации распределенного хранения и доступа к интегрированной информации являются: ресурс, отношение между ресурсами, поисковые метаданные и уникальный идентификатор ресурса. Каждому серверу назначается собственный сегмент именования в зарегистрированном пространстве имен. В рамках этого сегмента организация отвечает за уникальность имен своих ресурсов. Вышестоящая организация может выделить часть своего сегмента своим подчиненным организациям. Для обеспечения распределенного поиска и хранения используются следующие соглашения:

- Каждый ресурс обладает уникальным идентификатором.
- Каждому ресурсу сопоставлен один и только один сервер, *отвечающий* за него - его метаданные и содержание. Изменение, редактирование ресурса ведется только этим сервером. Другие сервера могут иметь копии ресурса. Они должны следовать одной из дисциплин поддержки актуальности копии ресурса (временной

период, сигнал об изменениях). Копии создаются и используются только для обеспечения эффективного доступа к содержанию ресурса.

- Каждый сервер имеет некоторое множество (возможно, пустое) ресурсов, за которые он отвечает. Кроме метаданных собственных ресурсов сервер обязан иметь копии *поисковых* метаданных ресурсов, которые *непосредственно связаны* с его собственными ресурсами, в соответствии с отношениями между типами ресурсов. Например, если публикация, ведомая некоторым сервером, имеет персону соавтора, за которую отвечает другой сервер, то первый сервер обязан иметь копию поисковых метаданных персоны, а второй - копию поисковых метаданных публикации.
- Для организации эффективного поиска и обеспечения отказоустойчивости, особенно, при малопроизводительных сетях и компьютерах, один сервер может аккумулировать поисковые метаданные подчиненных ему серверов. Это обеспечивает более эффективную обработку поисковых запросов и не требует больших объемов памяти для хранения информации. Такой способ индексирования информации при аккумулировании метаданных на основе некоторого критерия дает возможность создавать коллекции информации.
- Сервера могут образовывать сегменты, в которых обеспечивается “широковещательная” рассылка поисковых запросов. Ответом такого сегмента системы на некоторый запрос является сумма ответов об собственных ресурсах отдельных серверов системы. Технологии маршрутизации запросов на основе предварительной информации позволяют оптимизировать прохождение распределенных запросов при такой организации взаимодействия.

В четвертой главе рассматривается высокоуровневый механизм интеграции Web-технологий с СУБД, использовавшийся в реализации системы ИСИР и называемый Web+SQL технологией.

На текущий момент имеется масса инструментальных средств для создания статических Web страниц, которые в дальнейшем не будут модифицироваться или потребуют не значительных изменений. Имеются специализированные средств для создания и сопровождения динамических Web страниц, использующих информацию баз данных. Каждое из таких средств предлагает свой способ, свой язык для программирования динамических Web страниц. Отсутствие средства публикации данных из БД, поддерживаемого всеми Web серверами, необходимость программирования Web-приложений, хотя большинство динамических Web страниц не нуждается в сложном управлении данными БД, потребовали реализации для системы ИСИР более высокоуровневых, использующих декларативные спецификации средств доступа к данным БД и управления ими. Web+SQL технология позволяет легко создавать интерактивные, манипулирующие данными из БД приложения, поведение которых может зависеть от данных БД. Эта технология относится к среднему звену трехзвенной клиент/серверной архитектуры Web приложений. Но она не ориентирована на поддержку сложной прикладной логики, для чего должны использоваться соответствующие программные средства с мощной выразительностью такие, как C++ и Java, обеспечивающие и приемлемый

уровень переносимости. Web+SQL определяет класс электронных объектов, называемых *Web+SQL страницами*, поведение программ, их обрабатывающих, интерфейсы для связи с внешними компонентами.

Web+SQL страница - это обычная HTML страница, в которой имеется ряд вполне обычных HTML комментариев, однако, имеющих вполне определенные синтаксис и семантику. Как статическая HTML страница, она представляет собой образец, демонстрирующий внешний вид и поведение прототипа Web узла. В случае ее интерпретации специальными программными средствами Web+SQL страница выступает в качестве сценария формирования динамически генерируемых HTML страниц. Специальные комментарии специфицируют правила формирования, определяют структуру динамически генерируемых HTML страниц.

Поскольку Web+SQL страницы служат для публикации данные из БД, то их разбор осуществляется заранее на этапе подготовки, а не в процессе генерации HTML страниц. Полученные в результате разбора структуры сохраняются в БД, а извлекаются и интерпретируются уже в процессе генерации HTML страниц. Это позволяет исключить из времени генерации время, затрачиваемое на синтаксический разбор, контекстный анализ. Web+SQL приложения и весь Web сайт могут полностью содержаться в БД. Конечно, некоторые статические страницы и изображения могут находиться в файловой системе. Сохранение Web сайта в БД дает возможность сосредоточить средства развертывания и сопровождения в одном месте. Такая архитектура означает, что все данные, прикладные программы и Web сайты копируются, когда копируется БД. Это позволяет уменьшить сложность, понизить стоимость эксплуатации Web-системы.

Для поддержки Web+SQL технологии используется ряд программных компонент. Загрузчик страниц считывает содержимое файла, содержащего Web+SQL страницу, осуществляет разбор страницы, помещает внутреннюю структуру страницы, ее HTML и SQL код в соответствующие таблицы БД. Эти данные используются интерпретатором при генерации HTML страниц в ответ на запросы браузера, а компилятор позволяет сформировать по ним “родные” для Web сервера приложения, соответствующие таким технологиям, как Microsoft ASP, Apache PHP, Netscape LiveWire Javascipt, Oracle PL/SQL, Perl.

Связь Web+SQL страницы с окружающей средой осуществляется с помощью параметров. Параметры - это переменные, которым перед началом интерпретации страницы сопоставляются значения. Обычно, это значения, указанные в поисковой цепочке URL ссылки. Параметры и переменные Web+SQL страницы предоставляют простую возможность управлять формированием соответствующих HTML страниц и в основном используются для задания связи между Web+SQL конструкциями. Например, переменные дают возможность связать запрос с подзапросами, то есть сформировать SQL оператор одного запроса на основе данных и в процессе исполнения другого родительского запроса.

Единственным средством вычисления значений в Web+SQL является ассоциативное выражение. Оператор присваивания дает возможность определить значения параметров и переменных через значения других переменных, полей ре-

зультата запроса, ассоциативных выражений и т.п. Внести изменения в HTML код можно только с помощью оператора подстановки. В процессе генерации HTML страницы операторы подстановки заменяются соответствующими значениями переменных, столбцов запроса, ассоциативных выражений. Условный оператор позволяет управлять включением фрагментов Web+SQL страницы в процесс формирования HTML страницы в зависимости значений переменных или результата запроса.

Для обеспечения взаимодействия с БД служат SQL-команды:

- **SELECT** - запрос к БД для получения данных, подлежащих включению в генерируемую страницу,
- **INSERT** - вставка новых данных в БД, используя значения переменных страницы,
- **UPDATE** - модификация данных в БД на основе значений переменных страницы,
- **SQL** - может включать любой SQL оператор, который можно сформировать, используя значения переменных страницы.

SQL-команды состоят из условия исполнения и кода SQL оператора. SQL оператор выполняется только, если условие исполнения истинно, в частности, опущено. Текст SQL оператора можно формировать с помощью операции подстановки значений переменных. Кроме того, поддерживается классическая для динамического SQL возможность - привязка (binding) значений переменных к откомпилированному SQL оператору. Во время исполнения SQL оператора вместо динамических параметров подставляются текущие значения переменных. Такая возможность может быть использована при реализации вложенных запросов, чтобы сэкономить на времени, затрачиваемом на перекомпиляцию подзапросов. При этом следует учитывать, что все переменные Web+SQL страницы относятся к символьному типу. Этот механизм нельзя использовать для получения данных из БД. Единственной возможностью присвоить переменным данные из БД является обработка записей результата запроса команды SELECT.

Среди SQL-команд команда SELECT имеет самую сложную структуру, поскольку она определяет правила получения формируемых оператором SELECT данных и указывает, как использовать эти данные БД. Команда состоит из последовательности предложений: **`SELECT`**, **`NULL`**, **`ROW1`**, **`ROWS`** и **`TAIL`**, определяющих запрос и правила обработки возвращаемых им данных. В предложении **`SELECT`** указывается текст оператора SELECT, типы столбцов возвращаемого запросом результата, максимальный уровень управляющих прерываний, если этот механизм используется. Типы выбираются из предопределенного множества. Предложения **`NULL`**, **`ROW1`**, **`ROWS`**, могут содержать SQL команды. В частности, это могут быть SELECT команды, что позволяет описывать подзапросы, формируемые и управляемые в рамках родительских запросов. Например, два запроса можно связать соотношением “мастер-деталь”. Предложения **`ROW1`** и **`ROWS`** указывают, как должны быть обработаны соответственно первая и последующие за-

писи результата запроса. Предложения **NULL** выделяет конструкции, которые должны быть использованы генератором, если нет данных, удовлетворяющих условиям запроса.

Команда SELECT поддерживает механизм “управляющих прерываний” (control breaks), применяемый в генераторах отчетов и позволяющий осуществлять группировку выводимых данных, подставлять фактические значения управляющих столбцов только при возникновении прерываний. Если результат запроса слишком велик, то его можно выдавать по частям. Команда имеет две возможности определить виртуальный запрос, позволяющий повторить обработку фрагмента кода необходимое число раз в соответствии с заданным числом или последовательностью символьных цепочек.

Оператор вызова “подпрограмм” (Web+SQL страниц) позволяет при формировании одной страницы использовать HTML код, генерируемый другой страницей. Это дает возможность конструировать HTML страницы по модульному принципу. Поведением вызываемой страницы можно управлять посредством ее параметров так же, как при ее вызове по URL ссылке. Имя вызываемой страницы, значения ее параметров может быть получено в результате подстановки значений переменных, результатов запросов.

Предопределенные Web+SQL “подпрограммы” предоставляют ряд системных сервисов. Например, Web+SQL страниц ***#MIME*** вставляет в заголовок HTTP пакета MIME тип, который сообщается ей с помощью параметра. Web+SQL страница ***#EXIT*** прерывает обработку всей текущей иерархии обслуживаемых страниц. При этом управление передается другой Web+SQL странице, имя которой указывается ***#EXIT*** с помощью параметра. Такое же прерывание может быть инициировано внешними компонентами посредством соответствующего оператора создания исключительных ситуаций.

Для обращения к процедурам внешних для Web+SQL компонент используется оператор, во многом аналогичный по оператору вызова Web+SQL страниц, например, <!---@@pkg.&ProcName{&1, '&ln', '&env.uk-a', &oi, -&kind}---> Внешним процедурам могут передаваться текстовые и числовые значения. Процедуры могут вставлять HTML код, используя средства соответствующий программной платформы.

Кроме рассмотренных конструкции в Web+SQL имеется еще ряд простых и полезных способов обработки HTML кода.

В пятой главе рассматривается текущая реализация ИСИР РАН, реализованная на вышеизложенных принципах и механизмах, к которой можно обратиться по адресу ‘http://isir.ras.ru/’. Система поддерживает ресурсы: персона, организация, подразделение, публикация и проект. Реализована совокупность стандартных типов, упрощающих описание цифровой библиотеки. Система является многоязычной, но на данный момент имеются данные только на русском и английским языках. Для поддержки иерархические связей, рубрикации ресурсов в системе реализованы специальные механизмы навигации, учитывающие специфику соответствующих типов ресурсов и настраиваемые по их описаниям.

Используя Web-интерфейс, можно создавать новые ресурсы, редактировать их свойства, искать ресурсы по значениям их атрибутов. Можно осуществлять навигацию в пространстве ресурсов, просматривая сведения о них, загружая их содержание. В системе обеспечивается несколько уровней прав доступа к ресурсам. Метаданные ресурсов содержатся в реляционной БД, содержимое ресурсов, если оно имеется, хранится либо в реляционной БД, либо в файловой системе в зависимости от вида ресурса. Ресурс может иметь несколько экземпляров содержания, возможно, представленных в разных форматах.

Система реализована на платформах UNIX (Solaris) и MS Windows/NT, использует RDBMS Oracle и Oracle Web-сервер.

Заключение содержит основные выводы по данной работе. В нем определены возможные направления развития архитектуры ИСИР, реализации ИСИР РАН.

Приложения к диссертации содержат материал, иллюстрирующий основное содержание работы конкретными примерами.

Основные результаты

Основные результаты, полученные в диссертационной работе, состоят в следующем:

1. Разработана методика формирования и интеграции разнородных информационных и вычислительных ресурсов в единое информационное пространство. Методика ориентирована на использование специально подготавливаемых описаний ресурсов (метаданных), на обслуживание отобранных информационных ресурсов, оставляя вне рассмотрения некачественные, не представляющие какого-либо интереса данные.
2. На основе предложенной методики разработана открытая архитектура цифровой распределенной информационной системы интегрированных ресурсов. Архитектура дает возможность создавать и вести независимые, распределенные репозитории информационных ресурсов, интегрировать репозитории и ресурсы как на основе логической, так и тематической направленности.
3. Предложен язык описания ресурсов цифровой библиотеки, их связей, служб цифровой библиотеки, позволяющий настраивать систему на необходимую предметную область. Разработан механизм его реализации.
4. Разработана и реализована технология (Web+SQL) создания по декларативной спецификации интерактивных, управляющих данными в базе данных Web-приложений, поведение которых может зависеть от данных БД.
5. Осуществлена интеграция технологии Web+SQL с рядом Web систем.
6. На основе предложенных технологий реализована интегрированная информационная система РАН, позволяющая объединить информационные ресурсы РАН в единое информационное пространство.

Основные публикации по теме диссертации:

[1] Агошков С.В., Бездушный А.Н., Галочкин М.П., Кулагин М.В., Меденников А.М., Серебряков В.А., “Интегрированная Система Информационных Ресурсов (ИСИР) РАН – подход к созданию интегрированных цифровых библиотек”, международная научная конференция “Электронные библиотеки: перспективные методы и технологии, электронные коллекции”, Санкт-Петербург, 1999.

[2] Агошков С.В., Бездушный А.Н., Галочкин М.П., Кулагин М.В., Меденников А.М., Серебряков В.А., “Подход к созданию интегрированных цифровых библиотек”, Всероссийская научная конференция "Научный сервис в сети Интернет", Новороссийск, 1999.

[3] S.Agoshkov, A.Bezdushny, M.Galochkin, A.Medennikov, M.Koulagin, V.Serebiakov, “The Integrated System of Information Resources of the Russian Academy of Sciences - an Approach to Digital Library Design”, международный симпозиум “Software technology” FUSST’99, Tallin, Estonia, 1999.

[4] Бездушный А.Н., Кулагин М.В., Серебряков В.А., “Интегрированная Система Информационных Ресурсов РАН”, материалы 8-го научно-практического семинара “Современные технологии в информационно-библиотечном обеспечении научных исследований”, Таруса, 1999.