2002 | Том 5 | Выпуск 1

Архитектура Сервисов Интегрированной Системы Информационных Ресурсов (ИСИР)

А.Н.Бездушный, Д. А. Ковалев, В.А.Серебряков
Вычислительный центр Российской академии наук

1. Введение

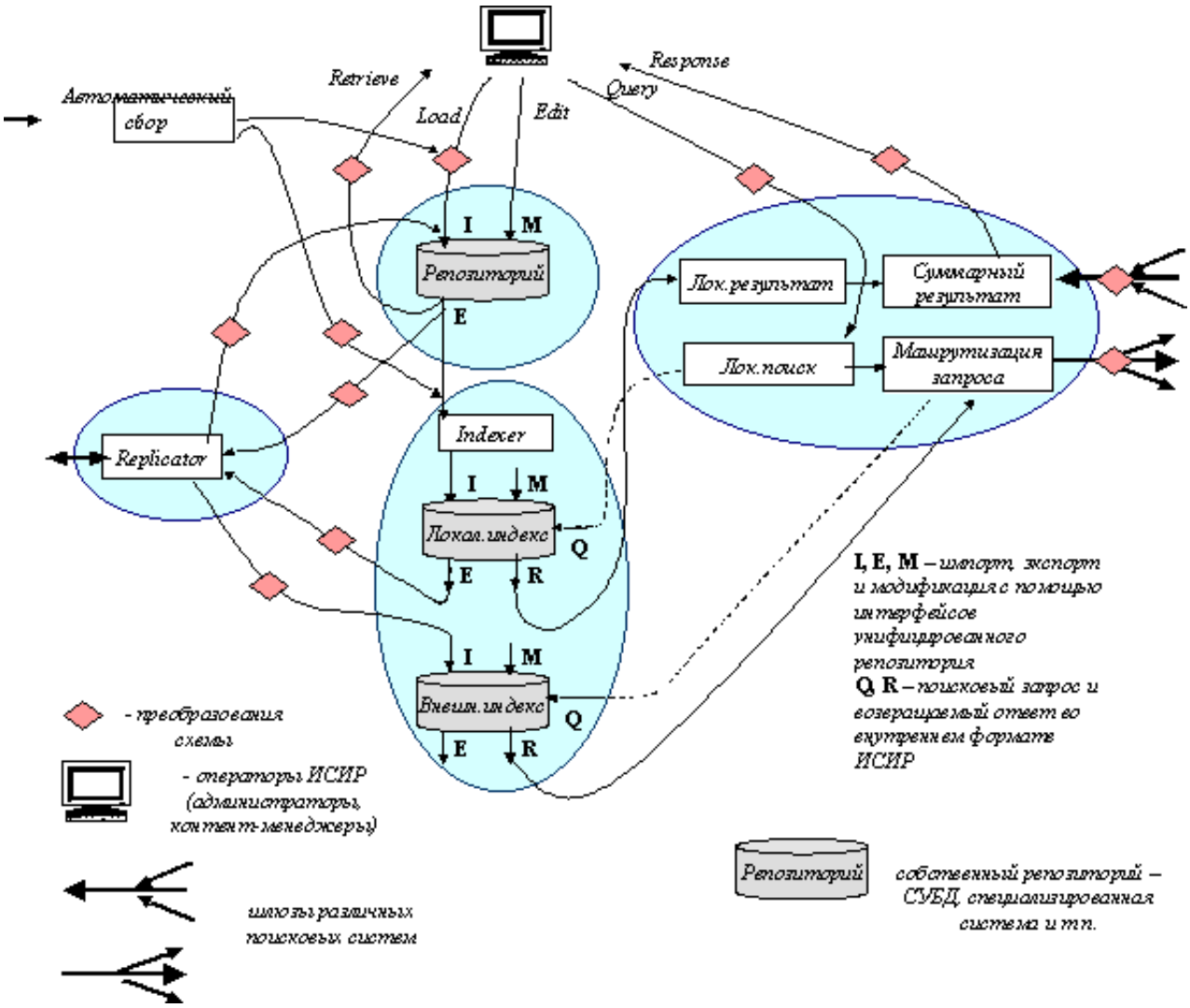
Работы, ведущиеся в рамках проекта ИСИР, ориентируются на внесение посильного вклада в решение проблемы организации единого информационного пространства РАН. Это требует решения большой совокупности задач - обеспечение извлечения и структуризации метаданных, поддержки их ввода в структурированном виде, предоставление средств интеграции информации разнообразных информационных источников (репозиториев) и т.д.

Под интеграцией мы понимаем следующее. Распределенная система ориентируется на объединение организаций, каждая из которых поддерживает коллекцию информации, представляющей общий интерес (например, научные публикации, сведения о сотрудниках и т.д.). Для хранения коллекции организации используют репозитории, представляемые некими «локальными» системами. Репозитории, в общем случае, используют различные модели представления данных, способы доступа к ним и т.д. В задачу подсистемы интеграции информации, выделяемой в рамках распределенной среды, входит обеспечение следующих уровней взаимодействия между отдельными репозиториями:

1. **обмен данными**; подсистема должна предоставлять средства, облегчающие и автоматизирующие импорт и экспорт данных, обмен данными между репозиториями;
2. **совместный поиск**; подсистема должна обеспечивать средства маршрутизации поисковых запросов, обслуживания их результатов, предоставления информации о способах доступа к найденным ресурсам;
3. **единообразный доступ**; подсистема должна обеспечивать унифицированный механизм доступа к найденным ресурсам, вне зависимости от конкретных репозиториев, в которых они располагаются, и базовых протоколов доступа, используемых внутри этих репозиториев.

В каждом конкретном случае количество поддерживаемых подсистемой уровней может варьироваться. Это зависит от возможностей и целей участия в формировании распределенной среды каждой «локальной» системы.

Поставленная задача интеграции решается с использованием сервисной архитектуры. В этом разделе дается краткое описание основных задач каждого из сервисов, более подробное описание приводится в соответствующих разделах. Концептуальное деление системы на сервисы, их взаимосвязь представлена на следующем рисунке.



2. Основные компоненты

2.1. Сервис**управления доступом**

Репозитории могут контролировать доступ к своим ресурсам, как с целью простого ограничения доступа, так и для обеспечения оплаты доступа, выполнения требований, на основе которых им были предоставлены ресурсы, например, контроль использования интеллектуальной собственности. Последнее процедуры связаны с *управление правами*, составляющими часть процесса *управления доступом*, который наряду с контролем доступа, поступления оплаты, использования интеллектуальной собственности, обеспечивает идентификацию пользователей и, возможно, ресурсов, может включать шифрование данных и т.п. В задачу сервиса управления доступом входит организация процесса управления доступом.

При взаимодействии пользователя с системой различаются разные уровни доступа, например, публичный и авторизованный обращения. В ходе публичного обращения любой пользователь может обратиться к информации полностью открытой для доступа. В процессе авторизованного доступа, после прохождения процедурыидентификации, пользователь дополнительно получает возможность взаимодействовать с информацией, предполагающей ограниченное

использование, например, репозиторий может содержать лицензированные материалы или материалы, являющиеся объектами особых условий доступа. Доступ пользователей, их групп выражается в терминах разрешенных действий.

Политика доступа к собственным ресурсам репозитория определяется его администраторами информации (информационными менеджерами). Она связывает на определенных условиях те или иные ресурсы репозитория с некоторыми группами пользователей и должна основываться на соответствующих законодательных актах, соглашениях с третьими сторонами, например, лицензиях предоставляемых держателями авторских прав. С этой целью пользователи и их обращения должны быть идентифицированы, их роль в процессе доступа четко определена. Аналогично, ресурсы репозитория, к которым осуществляется обращение, тоже должны быть идентифицированы, возможно, с установлением их аутентичности.

2.1.1. Сервис аутентификации

Каждое обращение пользователя к системе проходит определенный процесс управления доступом. При обращении к закрытым ресурсам системы осуществляются процесс идентификации пользователя - отождествления пользователя с одним из известных системе пользователей в ходе процедуры аутентификации, то есть опознавания и подтверждение подлинности. Затем процедура авторизации, проверяя полномочия пользователя, разрешает или отказывает ему выполнении определенной операции, например, в доступе к информации.Для обеспечения управления доступом к ресурсам они идентифицируются, то есть каждому ресурсу сопоставляется глобально уникальное имя (уникальный идентификатор - URN).

Аналогично аутентификации пользователей может проводиться аутентификация цифровых материалов, гарантирующая пользователям, что материалы не были подвержены изменениям. В большинстве случаев формальная аутентификация ресурсов не нужна и не проводится. Все базируется на высоком уровне доверия к репозиторию,его административным процедурами средствам доставки материалов, уверенности в том, что они обеспечивают адекватный уровень безопасности для хранения и предоставления доступа к ценной информации.

Преднамеренные изменения информации бывает очень редко, ошибки обычно легко отождествляются и не приводят к существенным проблемам. Однако, для таких материалов, как административные указания, решения искажение материалов или их подмена ошибки могут привести к серьезным последствиям. Для таких ресурсов должны использоваться методы аутентификации материалов. Для подтверждения истинности материала он обычно снабжается цифровой подписью, подтверждающей, что ресурс не изменялся с того момента формирования цифровой подписи. Для обеспечения безопасности пересылаемых данных они могут подвергатьсяшифрованию. В таких случаях используются стандартныеSSL-технологии.

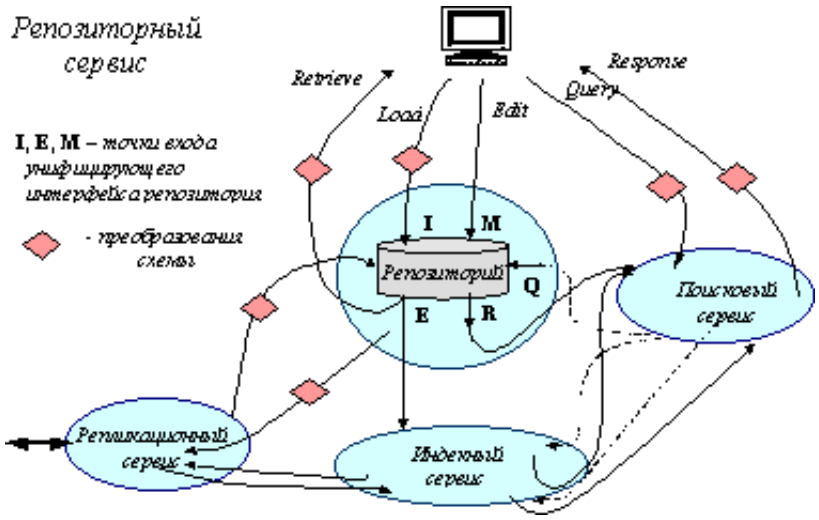
2.2. Сервис глобальной идентификации

Для организации распределенного функционирования, обеспечения управления доступом необходима глобальная система идентификации ресурсов всех репозиториев, независимая от схем идентификации в каждом из репозиториев. Проблема глобальной идентификации ресурсов решается в предположении, что в задачи каждого из интегрируемых репозиториев входит поддержка идентификаторов собственных ресурсов, уникальных в пределах репозитория. Для назначения ресурсам глобальных идентификаторов [URI] ИСИР использует внешнюю службуименования (Handle System [HDL]).

Служба именования используется сервисами для получения информации о возможностях доступа к ресурсу и другой метаинформации о ресурсе. Информация о возможностях доступа включает местонахождение ресурса, методы доступа к репозиторию-владельцу, идентификатор ресурса в рамках этого репозитория и т.д.

2.3. Репозиторный сервис

Каждый из репозиториев распределенной среды представляет собой некоторую «локальную» систему, содержащую предоставляемые данные. Локальные системы функционируют на различных платформах, используют различные технологии хранения и доступа, предоставляют различные возможности по работе с данными, и т.д. ИСИР предоставляет для этого собственное решение – подсистему ведения репозиториев, обеспечивающую ввод/вывод, экспорт/импорт, сопровождение, поиск структурированных данных. Подсистема предоставляется организациям РАН, однако ее применение не является необходимым для подсистемы интеграции.



2.3.1. Унифицирующий интерфейс репозитория

Задача унифицирующего интерфейса репозитория – «экранировать» остальные сервисы ИСИР от разнообразия используемых собственными репозиториями систем хранения, протоколов доступа и т.д. Это достигается за счет

предоставления минимума операций с данными, представленными в виде атрибутированных ресурсов.

Имеющейся огромный и постоянно увеличивающийся объем электронных данных сильно различается по степени структурированности данных. С одной стороны, это данные, хранящиеся в базах данных и имеющие строгую и правильную структуру. С одной стороны, это полностью неструктурированные данные, например, аудио- и видео- данные, планарный текст. Промежуточное положение между ними занимают, слабоструктурированные данные такие, как HTML страницы, форматированный текст, данные в XML формате и т.п.

Рассматриваемые нами «локальные» системы являются либо просто информационными системами, манипулирующими структурированными данными, либо Web-сайтами, которые можно рассматривать в качестве хранилищ «структурированных» данных, то есть поддерживающими меньшую степень гранулированности, чем HTML-страницы.

В исследованиях по интеграции информации сложилась архитектура, основанная на понятиях адаптера (wrapper) и посредника (mediator) [Wie92]. Первые используется для организации доступа к источнику данных, осуществляют выборку данных источника и их перевод в форму, позволяющую осуществлять дальнейшую их обработку средствами системы интеграции данных. Вторые осуществляют интеграцию данных различных источников, предоставляемых адаптерами. Они обеспечивают преобразование запросов пользователя в терминах промежуточной (mediate) виртуальной схемы в такие запросы, которые обращены непосредственно к источнику, его адаптеру. Для преобразования запроса используется описание источника информации, определяющее, какие имеются ресурсы у источника, а у ресурсов атрибуты, какие ограничения на содержание, возможности обработки запросов и т.п.

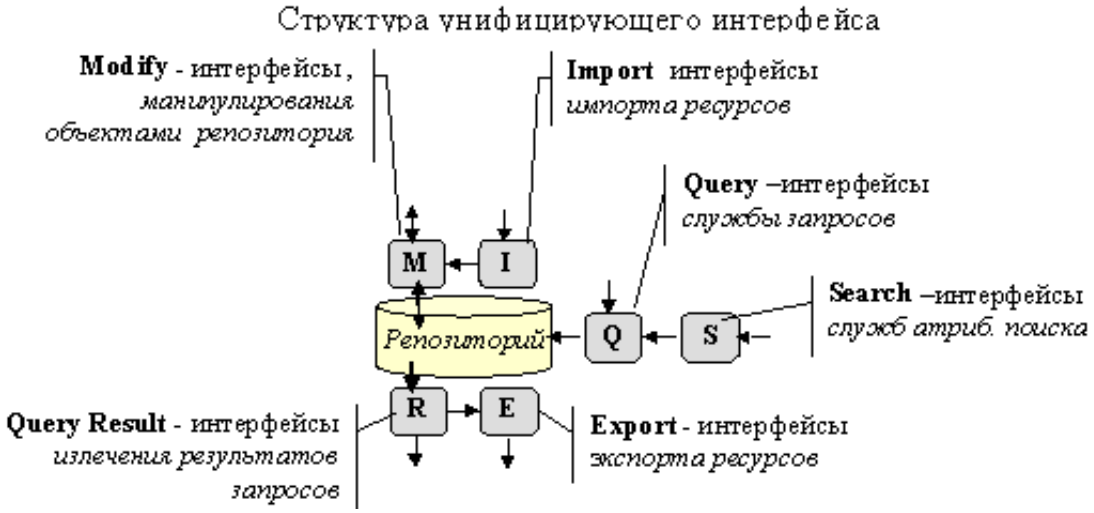
Такая архитектура возникла и широко применяется для интеграции гетерогенных реляционных баз данных, слабоструктурированных данных. Разработаны разные подходы к определению описаний источников и их использования при преобразовании запросов - GAV, LAV или их комбинация. Если источник обладает структурированным языком запросов, например, SQL и OQL, то система интеграции данных обеспечивает трансляцию поступающих запросов, выраженных в терминах промежуточной схемы, в запросы на языке источника в терминах его схемы.

Во многих системах стремятся описывать адаптеры и посредники декларативными средствами. Например, в системе TSIMMIS [TSIMMIS], разработанной для решения вопроса интеграции данных из различных источников, разработан логический язык запросов MSL над OEM моделью, который используется как язык для описания адаптеров и посредников с целью их последующей генерации. При этом вопрос об интерфейсе доступа к источнику данных скрывается в «действиях» правил преобразования входных структур в выходные.

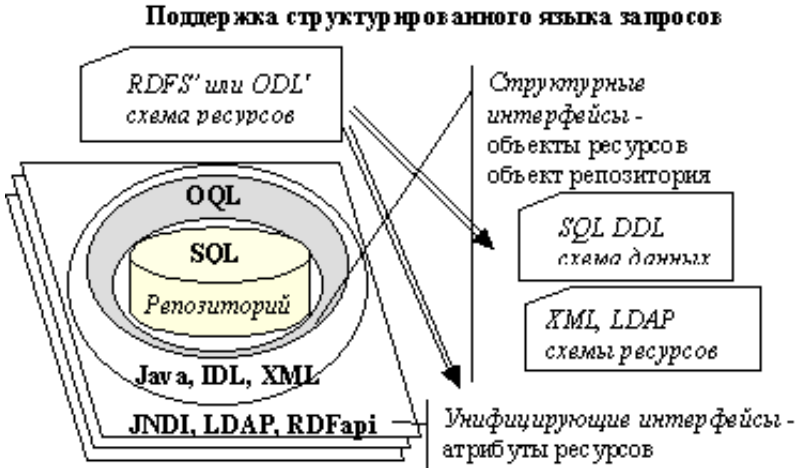
В отличие от интеграции гетерогенных баз данных, обладающих структурированными языками запросов, в нашем случае задача состоит в

поддержке операций атрибутного распределенного поиска, что несколько упрощает задачу. Мы не можем полагаться на поддержку структурированного языка запросов, поэтому в качестве интерфейса «адаптера источника данных» (интерфейса репозитория) используем открытые программные интерфейсы, например, одним из вариантов интерфейса является подмножество операций JNDI [JNDI]. Эти интерфейсы унифицируют («унифицирующие интерфейсы») доступ к данным репозиториев, используя программные возможности «локального» репозитория, для более высокоуровневых сервисов системы, обеспечивающих поддержку, например, репликации и обмена данными, индексирования и поиска информации.

Задачей унифицирующего интерфейса является предоставление высокоуровневым сервисам минимально необходимого набора операций с данными, в терминах атрибутированных ресурсов, соответствующих формальному описанию схемы репозитория. Сервисы более высокого уровня, используя описание схемы репозитория, могут работать с любым репозиторием, реализующим унифицирующий интерфейс.



Хотя мы не можем полагаться на поддержку структурированного языка запросов, но для случаев имеющих оный мы ориентируемся на автоматизированную поддержку унифицирующего интерфейса репозитория в таких случаях на более низких уровнях стека сервисов. Поддержка состоит генерации схем данных, программных интерфейсов, их реализаций по расширенному описанию схемы ресурсов репозитория. Сама схема ресурсов может описываться средствами расширенных диалектов языков RDFS или ODL (RDFS' и ODL'). Дополнительные указания, определяющие свойства операций репозитория, виды формирования схемы БД для хранения ресурсов, полнотекстовых индексов и т.п. приводятся в RDF-документах. Структурированным языком запросов является язык OQL, который для реляционных БД транслируется в SQL92 на основе соответствующего описания отображения, формируемого автоматически для генерируемых схем данных или описываемого вручную для унаследованных систем.



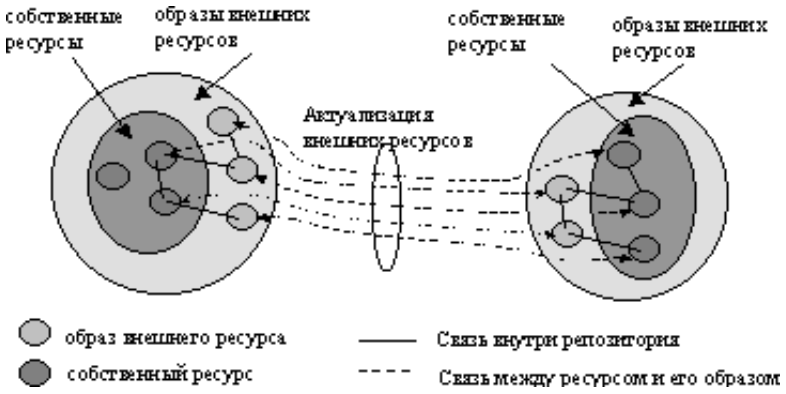
Для интеграции локальной системы в распределенную среду необходимо обеспечить только реализацию этого «унифицирующего» интерфейса. Набор операций этого интерфейса, минимально необходимый для работы высокоуровневых сервисов, включает:

1. добавление ресурсов;
2. изменение значений атрибутов по отдельности;
3. выборка всего ресурса или части его данных (указанного набора атрибутов);
4. выборка совокупности ресурсов, удовлетворяющих условиям, выраженным на некотором формальном языке запросов.

2.3.2. Связи между ресурсами разных репозиториев, дубликаты

Введение сервиса уникальной глобальной идентификации позволяет хранить информацию и осуществлять навигацию по связям между ресурсами не только в пределах одного репозитория, но и в рамках всей системы, обеспечивает возможность косвенного поиска, в том числе и по связям между ресурсами в разных репозиториях. Для того чтобы с учетом новых возможностей свести задачу поиска к суммированию ответов узлов, т.е. обеспечить применимость выбранных технологий распределенного поиска, инфраструктура обмена ИСИР предусматривает возможность поддержки следующих условий целостности.

Партнерские репозитории (имеющие большое количество связей между своими ресурсами) могут настроить инфраструктуру обмена на поддержку дополнительного ограничения целостности по определенным (наиболее важным) видам связи: если на узле X содержится ресурс A, связанный с ресурсом B на узле Y, то на X должна поддерживаться копия поисковой информации (поисковых индексов) B (Bx), а на Y - копия поисковой информации A (Ay). Допускается отложенное распространение изменений на копии.



Отметим, что речь идет о поддержке ограниченного числа связей, в рамках небольших подмножеств тесно сотрудничающих репозиториев, кроме того, распространяются копии только непосредственно связанных ресурсов. Это ограничивает возможности поиска заданием условий только одного уровня косвенности, однако уменьшает количество требуемых пересылок и уровень дублирования информации до приемлемого уровня.

2.4. Сервис метарепозитория

Задачей этого сервиса является хранение и предоставление информации о формальном описании схемы репозитория, о функциональных возможностях репозитория, настройках и параметрах отдельных сервисов.

На всех этапах поддержки распределенной среды, начиная с унифицирующего интерфейса репозитория, сервисы ИСИР активно используют формальные описания схем репозиториев, разнообразные настройки и т.п. В основном эта метаинформация используется локально, однако компонентам типа преобразователей схем необходима возможность работать с описаниями удаленных репозиториев. Поэтому формальные описания публикуются в Интернет в виде XML-документов, и регистрируются в службе именования. Это дает возможность по URI получить URL описания и загрузить его для обработки. В дальнейшем по мере развития формальных методов сопоставления схем сервис метаописаний будет хранить и манипулировать этой информацией как с ресурсами с фиксированной схемой.

2.5. Сервис администрирования

Этот сервис предоставляет средства администрирования службами репозитория, к которым в первую очередь относятся сервисы, обеспечивающие работу репозитория в распределенной среде. На основе этого же сервиса может осуществляться администрирование данных локальной системы, если она такие возможности предоставит.

2.6. Сервис персональной инф**ормации**

Задачей этого сервиса является сопровождение персональной информации пользователей, за сопровождение информации о которых несет ответственность рассматриваемый репозиторий. К персональной информации относятся параметры

аутентификации, предпочтения пользователей, данные, связываемые с этими пользователями сервисами распределенной среды.

2.7. Транспортный сервис

ИСИР реализует настраиваемый “сервис обмена”, поддерживающий необходимые модели обмена сообщениями и реализующий требуемые виды обмена. Архитектура сервиса позволяет использовать разные транспортные протоколы, службы. Должна быть обеспечена поддержка протокола CIP [CIP] и Java интерфейса JMS [JMS].

Для представления передаваемых данных используется модель RDF, в которой модель атрибутированных ресурсов имеет прямое отражение - описание ресурса состоит из набора RDF-предложений вида «ресурс X имеет атрибут Y со значением Z» или «ресурс X связан связью Y с ресурсом Z».

Схема данных - набор допустимых классов ресурсов и т.п. - формально выражается на языке RDF-schema, дополненном набором дополнительных ограничений в рамках стандартного синтаксиса. При этом описанные выше обобщенные операции с унифицирующим интерфейсом репозитория уточняются следующим образом:

1. Имеется приложение (RDF-загрузчик), принимающее на стандартный вход RDF-модель в виде документа RDF/XML, проверяющее соответствие модели схеме репозитория, выраженной на RDF-schema, и загружающее описанные в RDF-модели ресурсы в репозиторий. Определен стандартный интерфейс к приложению, позволяющий с помощью ключей командной строки задать режим интерпретации модели - полное описание содержимого репозитория или список изменений к текущему состоянию.
2. Второе приложение - RDF-генератор - принимает с командной строки набор условий на искомые ресурсы, а также список атрибутов ресурсов, подлежащих выгрузке, и выдает RDF-модель, содержащую описания всех запрошенных атрибутов соответствующих ресурсов репозитория в виде документа RDF/XML, соответствующего схеме репозитория, выраженной на RDF-schema.

Далее будет подробно рассмотрено применение описанного интерфейса для обеспечения работы индексного и репликационного сервисов, и будет показано, как такая организация сводит действия по интеграции некоторого репозитория в ИСИР по уровню 1 и/или 2 к реализации двух вышеописанных приложений, созданию RDFschema-описания схемы данных репозитория и настройке вышестоящих сервисов ИСИР.

Описанный выше механизм специфицирует реализацию унифицирующего интерфейса на достаточно высоком уровне. С одной стороны, это расширяет класс систем, поддерживаемых инфраструктурой интеграции ИСИР, с другой - оставляет достаточно большую часть действий на реализацию для каждого конкретного репозитория (включая разбор, интерпретацию и генерацию RDF-документов и т.д.). Введя более строгие спецификации на требуемые операции, можно

выделить достаточно широкий подкласс систем, для которых можно еще более сузить необходимые действия. А именно, в ИСИР предусматривается реализация вышеуказанных RDF- загрузчиков и генераторов для репозиториев, поддерживающих интерфейс JNDI. Этот интерфейс поддерживает все необходимые манипуляции для случая иерархически организованного множества планарных ресурсов (т.е. ресурсов, у которых атрибуты имеют только простые значения), и при этом полностью параметризуется именами элементов схемы. Это позволяет реализовать «универсальные» загрузчик и генератор RDF для JNDI-репозитория, которые используют RDFschema-описание не только для проверки корректности входного документа, но и для параметризации JNDI-вызовов, и тем самым подходящих для целого класса JNDI-совместимых репозиториев.

Спецификации JNDI стандартизуют не только клиентский JNDI API, но и средства адаптации для него новых протоколов и систем (JNDI SPI), тем самым серьезно облегчая задачу поддержки унифицирующего интерфейса. Кроме того, достаточно много информационных систем, в первую очередь, поддерживающих протокол LDAP, уже имеют JNDI-адаптеры.

Аналогично JNDI, в дальнейшем планируется реализация «универсальных» загрузчиков и генераторов для других популярных протоколов и систем, например Z39.50, SDLIP.

Интересно также отметить, что эти и подобные популярные протоколы, сводимые на уровне репозиториев к единому интерфейсу, возникают и на конечном пользовательском уровне ИСИР, в виде шлюзов для этих протоколов к сервису распределенного поиска. Таким образом, инфраструктура ИСИР имеет возможность как использовать данные, доступные по различным популярным протоколам, так и предоставлять собственные данные по ним.

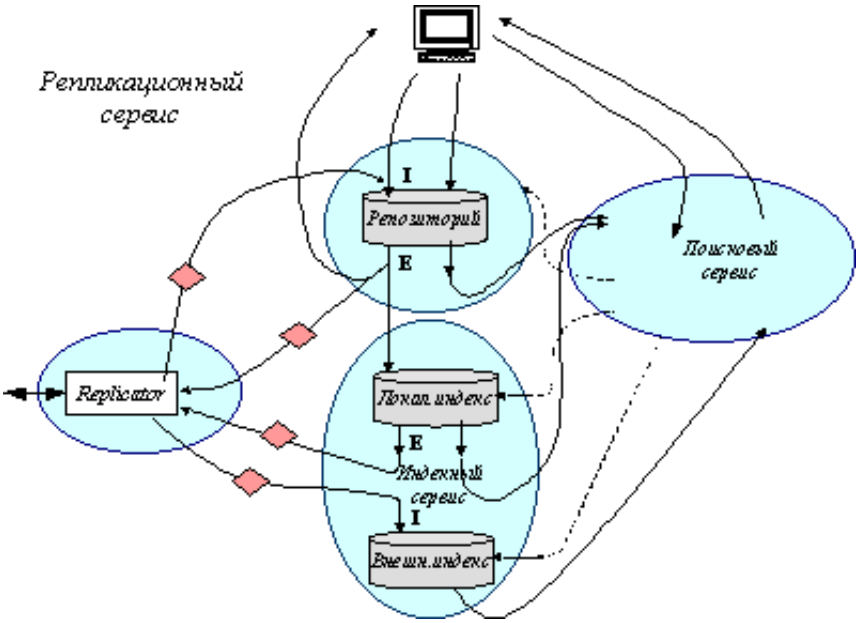
2.8. Сервис репликации

Сервисы распределенной системы предполагает автоматизированный обмен информацией между репозиториями, происходящий на постоянной основе и позволяющий минимизировать взаимодействие при ответе на пользовательский запрос. Виды обмена включают:

- обмен данными между отдельными репозиториями через унифицирующие интерфейсы репозиториев;
- репликация и обновление реплик локальных поисковых индексов для обеспечения балансировки нагрузки при выполнении операций поиска;
- концентрация на поисковых серверах предварительной информации о содержимом репозиториев (описателей коллекций) для маршрутизации запросов.

Анализ этих видов обмена приводит к выводу, что все они укладываются в общую модель обмена сообщениями, широко используемую в задачах интеграции и распределенных коммуникациях. Целый класс приложений среднего слоя (Message-OrientedMiddleware) успешно использует модель обмена сообщениями для интеграции и обеспечения распределенной работы разного рода приложений.

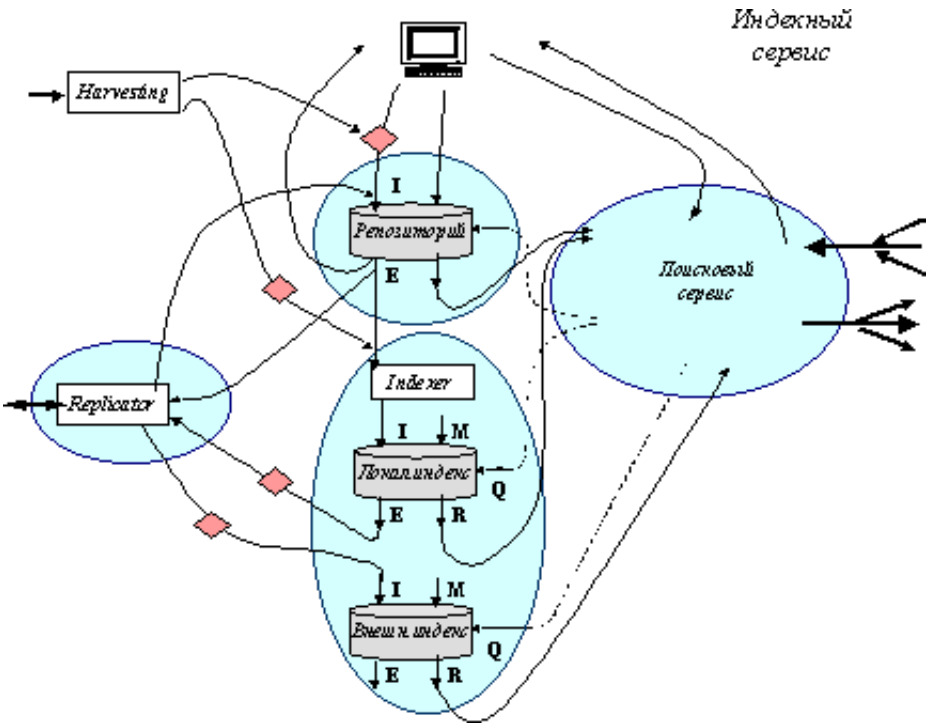
Две основные модели обмена – PTP (point-to-point) и PS (publisher-subscriber) предоставляют гибкие средства конфигурирования обмена.



2.9. Сервис актуализации

Задачей этого сервиса отслеживание актуализации информации, за сопровождение которой несет ответственность рассматриваемый репозиторий. Отслеживание может заключаться в информировании администраторов репозитория о нарушении условий актуальности информации.

2.10. Индексный сервис, описатели коллекций



2.10.1. Локальный поисковый сервис

Задача локального поискового сервиса – основываясь на предоставленном минимуме операций унифицирующего интерфейса репозитория, предоставить гарантированные возможности по поиску информации в терминах атрибутированных объектов.

Предполагается, что внешний интерфейс к локальному поисковому сервису ИСИР будет включать несколько шлюзов для популярных поисковых протоколов (Z39.50,SDLIP, LDAP etc.), Web-интерфейс, поисковый язык на основе ODMG OQL.

Запрос во внутреннем представлении – это некоторое дерево, во внутренних вершинах которого находятся логические связки “и, или, не” или имена отношений, определенных в схеме репозитория, в листьях – элементы, задающие условия на значения атрибутов искомых ресурсов, представляемые в виде троек «имя атрибута из схемы, операция, значение». Если некоторой внутренней вершине соответствует имя отношения, то в ее поддереве не должно быть других таких же вершин. Это поддерево интерпретируется как множество условий на значения атрибутов ресурсов, связанных с искомым соответствующим отношением.

Такое внутреннее представление запроса дает возможность осуществлять прямой атрибутный поиск и задавать косвенные условия на атрибуты непосредственно связанные с искомыми ресурсами.

2.10.2. Структура локальных поисковых индексов

Важно заметить, что схема ресурсов, в терминах которой задается поисковый запрос, может специфицировать сложноструктурированные атрибуты, состоящие из нескольких полей и подструктур. Например – сложный атрибут «адрес», с полями «индекс», «город», «телефон» и т.д. Однако, условия могут задаваться только на значения атомарных атрибутов или атомарных составляющих сложных атрибутов (в примере с адресом - «адрес.город=Москва» и т.п.). Это означает, что поисковая информация, экстрагируемая поисковым сервисом, имеет планарную структуру – сложных атрибутов нет.

На основании анализа различных поисковых систем, реализующих ту или иную часть необходимой функциональности, в ИСИР реализуется следующее решение в отношении к структуре индексов.

В формальное описание схемы репозитория вводятся описатели индексов для всех полей сложных атрибутов, допустимых в поисковых запросах. Описание включает тип индекса, который нужно построить, и набор настроек, зависящий от типа. Утилита-индексатор использует это расширенное RDFschema-описание, выделяя необходимые данные из выходных документов RDF-генератора, взаимодействующего с репозиторием через унифицирующий интерфейс.

Выделяются несколько видов индексов:

- **Атрибутные** - индексируются значения атрибута целиком. Допустимые операции включают “=, >, <, *включает*”. Указывается тип данных, задающий необходимую интерпретацию значений, представленных в символьном виде.
- **Полнотекстовые** и **ключевые слова** - значения атрибутов разбиваются на отдельные термы (ключевые слова), которые индексируются по отдельности. Применимые операции - “*содержит слово, часть слова*”.. . Эти виды различаются по способам выделения слов, основ слов, сопоставления весов.
- **Связи** - поддерживаются только двусторонние связи, которые выражаются в схеме двумя способами. Первый предполагает наличие в обоих классах ресурсов атрибутов, содержащих URI противоположных ресурсов. (В этом случае для каждого из таких атрибутов имеется описание, указывающее имя связи, которую он реализует, и ссылку на описание двойственного ему атрибута. Второй способ требует наличия в полученных данных отдельного RDF-ресурса, производного от специального класса ISIRRelation и содержащего всю необходимую информацию.

2.10.3. Реляционные структуры для работы с поисковыми индексами

Реализация локального поискового сервиса ИСИР основывается на РСУБД. В качестве РСУБД может использоваться широкий набор ПО разных производителей, т.к. набор необходимых требований не выходит за рамки подмножества SQL-92.

Принципиальная реляционная схема для работы с поисковыми индексами ИСИР включает:

1. Таблицу описаний индексов *isir_index_info*, в которой отражены имена индексов, их тип, и имена соответствующих таблиц с данными.
2. Таблицу указателей ресурсов, *isir_indexed_resources*, содержащую URI всех проиндексированных ресурсов.
3. Для каждого атрибутного индекса – таблицу- словарь , содержащую все уникальные значения индексируемого атрибута, и таблицу, регистрирующую все вхождения этих значений.
4. Для каждого индекса по ключевым словам - аналогичные таблицы, содержащие собственный словарь значений и перечисление их вхождений.
5. Для полнотекстовых индексов - единый словарь термов для всех атрибутов, для которых задано построение полнотекстового индекса. Таблица, хранящая перечисление вхождений термов, содержит, кроме пары идентификаторов (терма и ресурса), указатель позиции терма в тексте значения. Указатель позиции используется при поиске фраз, для оценки близости вхождений термов.
6. Реестр связей, содержащий наименования связей, и таблицу, фиксирующую наличие отношений между ресурсами.

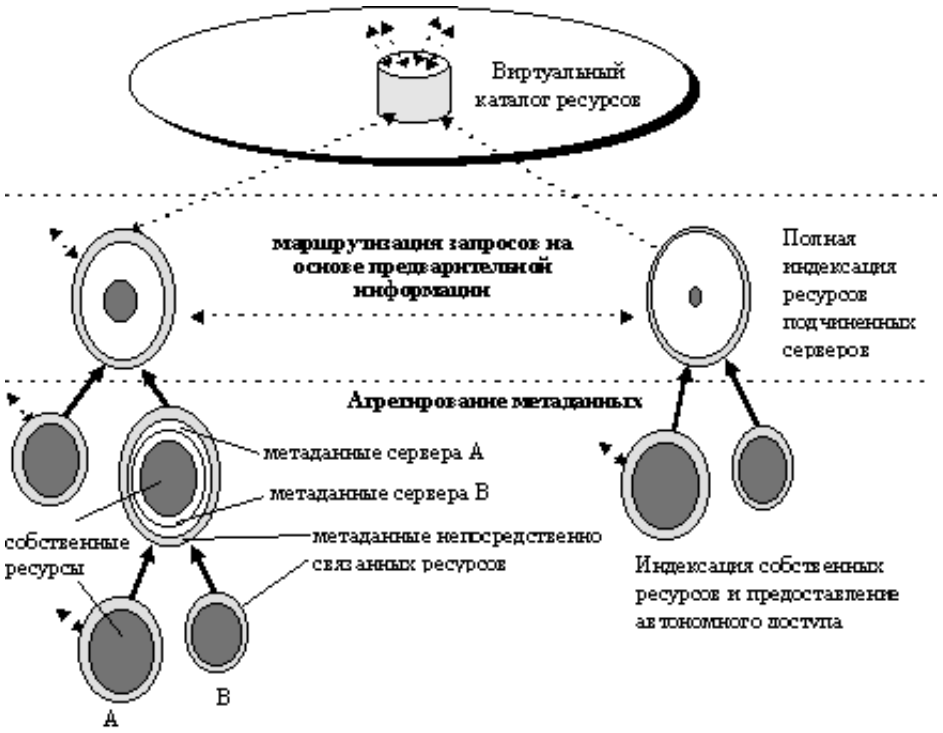
Алгоритмы поиска с использованием этих структур сводятся к задаче генерации SQL-запросов по внутреннему представлению поискового запроса.

2.11. Сервис распределения предварительной информации

Основная технология, адаптированная ИСИР для достижения нужной эффективности распределенного поиска, состоит в концентрации поисковой информации на подмножестве узлов системы, обладающих большими вычислительными мощностями и хорошо связанных друг с другом. При этом, источником обновления этой информации, ответственным за ее актуальность и полноту, остается исходный репозиторий. Инфраструктура ИСИР просто настраивается на поддержку актуальных копий поисковых индексов, сформированных локальным поисковым сервисом репозитория, в “вышестоящем” “поисковом” репозитории, концентрирующем поисковую информацию. Копии размещаются в структурах локального поиска, наряду с информацией о его собственных ресурсах - содержимое таблиц ресурсов и словарей индексов смешиваются, а таблицы вхождений объединяются. “Свои” и “чужие” ресурсы различаются по признаку “is_local”. Для всех ресурсов поддерживается временная метка последнего обновления. При обработке поисковых запросов локальный поисковый сервис “вышестоящего” репозитория отвечает на запросы как о собственных данных, так и о данных удаленных репозиториев. Обычно он только формирует объединенный ответ, а описание свойств ресурса, его содержание предоставляется пользователям исходным репозиторием.

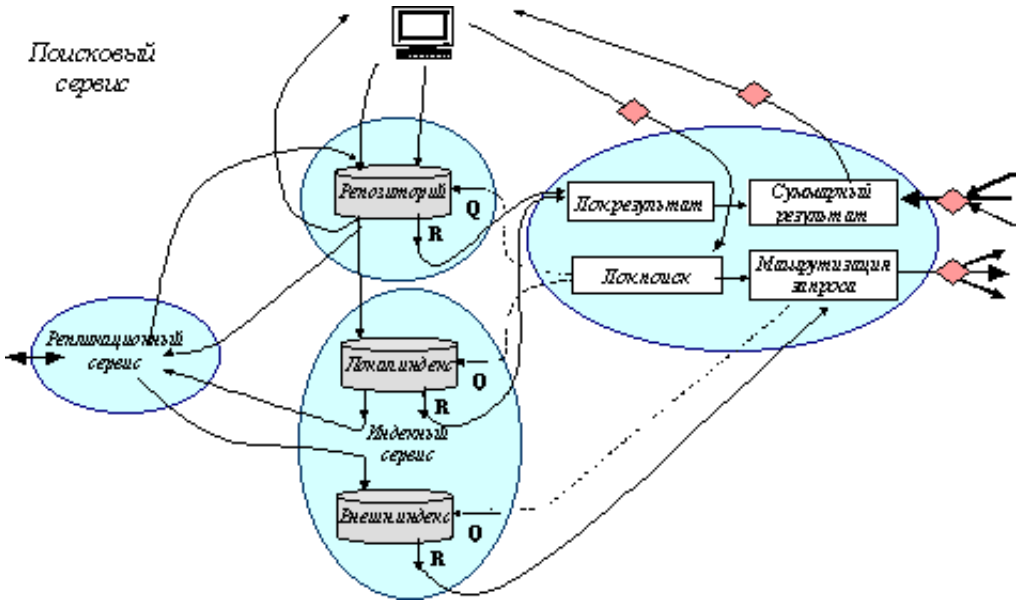
Таким образом, концентрация поисковой информации подразумевает образование множества иерархических конфигураций, в каждой из которых поисковые индексы с подчиненных узлов реплицируются на родительский узел, и далее, возможно, до корня. Для поисковых индексов разрабатывается соответствующая RDF-схема, по-существу специфицирующая возможности унифицирующего интерфейса репозитория и позволяющая осуществлять обмен индексами с помощью сервиса обмена данными.

Анализ показывает, что до определенного уровня эта иерархия может совпадать с иерархией подчинения соответствующих организаций - владельцев репозиториев. Например, узел центральной библиотеки может концентрировать поисковую информацию филиалов. Это стимулирует использование согласованных схем ресурсов. Однако, на верхних уровнях на первый план выступают критерии не административного, а технического характера - необходимо сконцентрировать поисковую информацию на наиболее мощных серверах, а также осуществить перераспределение ресурсов в тематические коллекции (см. далее), чтобы обеспечить хорошее качество маршрутизации запроса.



2.12. Сервис распределенного поиска, сервис маршрутизации запросов

В задачи этого сервиса входит формирование результатов поисковых запросов к распределенной системе на основе данных входящих в нее репозиториев. Для организации эффективного распределенного поиска сервис использует технологии балансировки нагрузки и маршрутизации запросов на основе предварительной информации. Предварительная информация, используемая сервисом для маршрутизации запросов, формируется из локальных поисковых индексов каждого репозитория. Важные моменты использования этих технологий для обеспечения более богатых поисковых возможностей описаны н иже в разделе «Связи между ресурсами разных репозиториев».

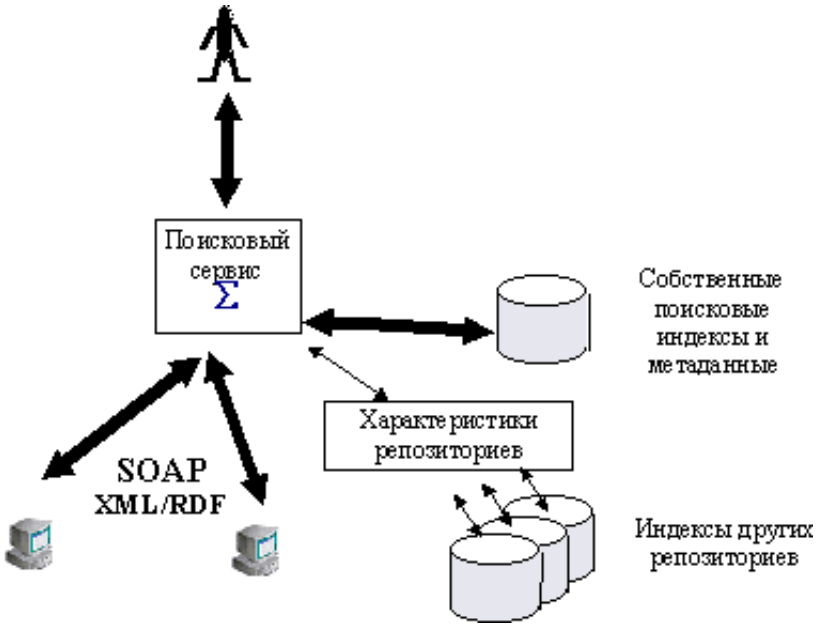


2.12.1. Маршрутизация запросов

Классическая задача маршрутизации запросов состоит в сужении множества узлов - участников обработки запроса на основе предварительной информации об их содержимом. Эта информация (описатели коллекций) анализируется на соответствие пришедшему запросу, некоторое количество наименее “перспективных” серверов отбрасывается, сокращая тем самым накладные расходы на формирование ответа за счет его возможной неполноты. Обзор различных подходов к формированию описателей, методов оценки релевантности, а также экспериментальные данные приведены в [RCDL2000-1].

В принципе, этот процесс может повторяться на каждом из получивших запрос узлов, в отношении “подчиненных” им узлов, тем самым оправдывая термин “маршрутизация” - запрос маршрутизуется по иерархии узлов, происходит отсечение ветвей.

Необходимо заметить, что в ИСИР автоматическое отсечение узлов считается неприемлемым, т.к. может привести к невозможности найти некоторый ресурс в принципе (такая возможность следует из состава используемых описателей коллекций - см. далее). Поэтому предварительная информация используется для упорядочения узлов в порядке убывания оценки их релевантности запросу, с тем чтобы пользователь мог самостоятельно принять решение об участии того или иного узла в ответе.



2.12.2. Модель шага маршрутизации

Действия поискового сервиса ИСИР на одном шаге маршрутизации описываются следующей последовательностью действий (некоторые варианты оптимизации этого процесса будут приведены ниже):

1. получить формальные (числовые) оценки соответствия Relevancy(q,Ci) каждой из коллекций-кандидатов Ci запросу q, на основе описателей

коллекций (см. ниже)
2. составить числовую оценку «ценности» каждой коллекции Value(q,Ci) как взвешенную суперпозицию вышеупомянутой оценки соответствия и обратной оценки стоимости запроса к коллекции Expense(Ci). Последние могут присваиваться коллекциям администраторами системы – пропорционально вычислительным и коммуникативным мощностям соответствующих узлов).
Value(q,Ci)=a1*Relevancy(q,Ci)+a2/Expense(Ci)
3. упорядочить коллекции в порядке убывания Value(q,Ci)
4. разослать запрос выбранным узлам, суммировать ответ

В качестве меры соответствия коллекции Ci запросу q – Relevancy(q,Ci) –используется оценка количества документов этой коллекции, удовлетворяющих ему.

2.12.3. Описатели коллекций

Рассмотрим подходы к формированию описателей, обеспечивающие приемлемую точность оценки при достаточной компактности. Техника маршрутизации запросов имеет смысл, когда объем предварительной информации о коллекции заметно меньше, чем объем данных коллекции. Применительно к ИСИР, описатель коллекции должен быть существенно компактнее ее поискового индекса. Поисковый индекс включает реестр проиндексированных ресурсов и для каждого индекса:

- словарь значений (количество записей равно количеству уникальных значений);
- перечень вхождений (количество записей для текстовых атрибутов - порядка количества слов во всех индексированных текстах).

Описатель коллекции в ИСИР, адаптированный из предложений WHOIS++ , где он именуется центроидом (centroids), для каждого поискового индекса представляет из себя:

- словарь значений, для каждого терма которого указано nR(t,Ai,Cj) - количество ресурсов, имеющих терм в значении соответствующего атрибута(ов).

Описатель коллекции легко формируется по поисковому индексу.

Подобный описатель на несколько порядков компактнее поискового индекса, более того - начиная с некоторого момента не зависит от объема коллекции, даже для текстовых атрибутов, т.к. имеет количество записей порядка размера словаря. Однако он не позволяет точно вычислить количество документов, удовлетворяющих запросу, заставляя прибегать к статистическим оценкам. Наиболее характерный пример - текстовые атрибуты, условия поиска на которые могут включать условия типа «содержит фразу». В ИСИР предполагается использовать (описанный в [RCDL2000-1] применительно к полнотекстовому поиску) простой статистический подход к оценке Relevancy(q,C), в

предположении, что термы распределены по ресурсам коллекции равномерно и независимо. В этом предположении вероятность вхождения терма в значении атрибута A некоторого ресурса равна отношению nR(t,A,Ci)/|Ci|, где |Ci| - общее количество ресурсов в коллекции. Вероятность совместного вхождения нескольких термов равна произведению соответствующих вероятностей.

2.12.4. Тематические коллекции

Инфраструктура интеграции предполагает перераспределение поисковой информации в процессе репликации (балансировки поисковой нагрузки) в тематические коллекции, т.е. коллекции, содержащие ресурсы по одной (или нескольким смежным) теме. Автоматизация этого процесса может быть достигнута за счет формирования индексов для атрибутов, ссылающихся на рубрики рубрикаторов и т.п. В этом случае администраторы системы получают возможность осуществить репликацию ресурсов, имеющих сходные значения разных рубрикаторов, в рамках одной тематической коллекции.

Кроме повышения эффективности маршрутизации, это дает возможность сильно сократить объем описаний на самом верхнем этапе маршрутизации за счет использования *сокращенных описаний*. Сокращенные описания строятся по полным, путем отбрасывания информации о термах {tj}, для каждого из которых количество содержащих его ресурсов коллекции CnR(tj,C) не превосходит заданного порога.

2.13. Сервис агрегирования результатов запросов

Сервис решает задачи агрегирования результатов поисковых запросов, их ранжирования.

Задача сервиса распределенного поиска состоит в формировании ответа распределенной системы на основании ответов отдельных ее частей – локальных поисковых сервисов, входящих в нее репозиториев. Подразумевается, что введенные ранее условия целостности по связям между ресурсами разных репозиториев выполнены, поэтому консолидированный ответ системы действительно можно получить, суммируя ответы отдельных репозиториев. Необходимо обеспечить приемлемое время ответа распределенной системы (сравнимое со скоростью ответа поисковых машин - порядка нескольких секунд – десятков секунд), в условиях существенно неоднородных вычислительных и коммуникативных возможностей репозиториев.

2.14. Сервис преобразования

Важным аспектом функционирования распределенной гетерогенной среды является обеспечение отображения схем различных репозиториев, позволяющих осуществлять преобразование запросов и данных. Введение унифицирующего интерфейса репозитория позволяет описывать данные единообразно, в терминах атрибутированных ресурсов, однако никак не фиксирует структуру и семантику соответствующих классов ресурсов, атрибутов и т.п. Создание единой согласованной схемы в такой постановке задачи невозможно из-за большого

количества объединяемых систем, различного подчинения соответствующих организаций, разных предпосылок при выборе той или иной модели данных и т.д. С другой стороны, поисковые запросы пользователей к распределенной системе формулируется для некоторой вполне определенной, обычно, широко распространенной схемы. Так, некоторые системы предоставляют возможность перед формулировкой запроса выбрать одну из поддерживаемых схем, основанных на популярных стандартам, например, [DC, MARC, Z39.50].

Для решения этих проблем в архитектуре ИСИР предусмотрены следующие возможности преобразования данных:

- На этапе создания (унифицирующий интерфейс репозитория). Оболочка реализуется таким образом, чтобы отобразить структуры данных локального репозитория в схему ресурсов, максимально приближенную к одной из «стандартных».
- На этапе обмена (после выборки данных через унифицирующий интерфейс, или перед их загрузкой, с помощью компонент-преобразователей).
- На этапе создания локальных поисковых индексов и описателей коллекций. На этом этапе схема репозитория может видоизменяться, например, за счет слияния нескольких атрибутов в один индекс, переименования и т.п. Аналогичные преобразования могут производиться при формировании описателей коллекции на основе поисковых индексов.
- На этапе обработки запроса. При перенаправлении запроса некоторому репозиторию, программы-посредники могут переформулировать исходный запрос в запрос, обращенный непосредственно к репозиторию.

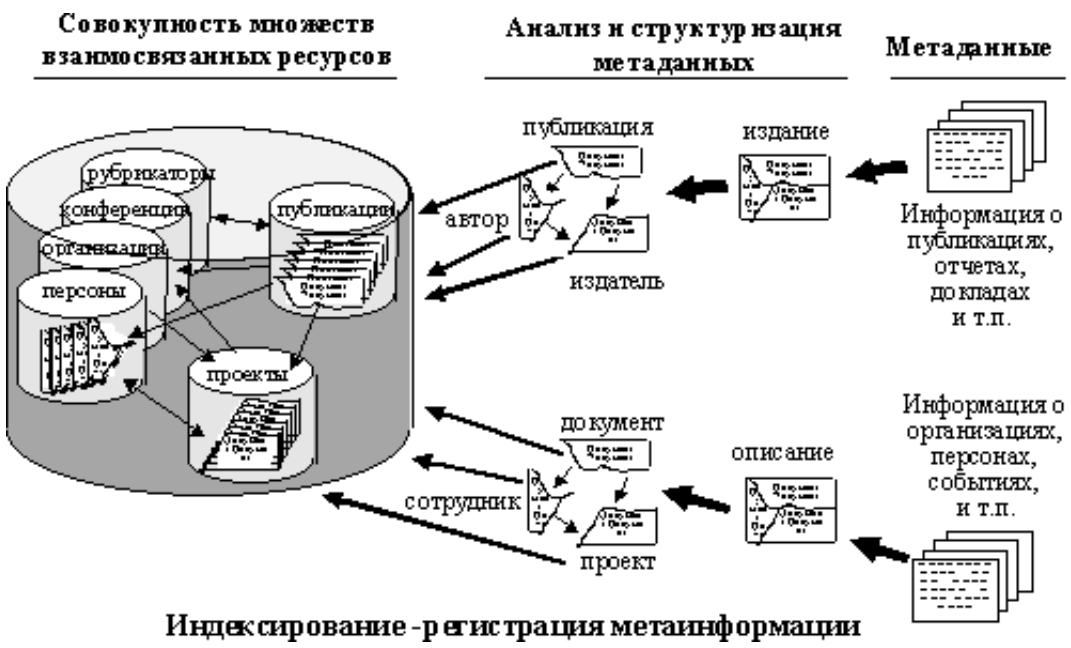
Заметим, что возможности преобразования существенно различны на разных этапах, и решают различные задачи. Первые два из описанных этапов предоставляют наиболее широкие возможности преобразования, и предполагают полное семантическое преобразование, позволяющее прямой обмен данными между репозиториями. Однако, такая степень преобразований достаточно трудоемка, требует подробных формальных описаний, и не всегда возможна на основании имеющегося в распоряжении набора данных.

Вторые две возможности (на этапах формирования локальных поисковых индексов и переформулировки запроса) ориентированы в первую очередь на сохранение семантики поиска, и допускают потерю структуры. Например, при поиске можно поставить соответствие между атрибутом «аннотация» в одной схеме, и атрибутами «краткое содержание» и «ключевые слова» в другой, совместив соответствующие индексы.

Преобразования первого класса совершаются компонентами-медиаторами над данными в RDF-представлении, на основании формальных описаний отображения схем.

2.15. Сервис сбора метаданных и индексной информации

Сервис решает задачи по извлечению и структуризации метаданных.



3. Литература

[URI] T. Berners-Lee, R. Fielding, L. Masinter. “Uniform Resource Identifiers (URI): Generic Syntax”, IETF RFC 2396, August 1998.

[HDL] The Handle System Home Page, http://www.handle.net/

[DC] The Dublin Core Metadata Initiative. http://purl.org/dc

[MARC] ISO TC46/SC4; http://lcweb.loc.gov/loc/standards/isotc46/, http://www.loc.gov/marc/

[Z39.50] ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Z39.50 Maintenance Agency Offical Text for Z39.50-1995, July 1995; http://lcweb.loc.gov/z3950/agency; http://www.niso.org/

[CIP] Common Indexing Protocol. http://www.rfc-editor.org/cgi-bin/rfcsearch.pl?searchwords=CIP &num=1500&format=ftp

[Wie92] G. Wiederhold. Mediators in the architecture of future information systems. In IEEE Computer 25:3, pp. 38-49.

[TSIMMIS] www.db.stanford.edu/tsimmis

[JMS] Java Messaging Service. http://java.sun.com/products/jms/

[JNDI] Java Naming and Directory Interface - unified interface to multiple naming and directory services. http://java.sun.com/products/jndi/

[LDAP] Lightweight Directory Access Protocol. The protocol is designed to provide access

to directories supporting the X.500 models. http://www.rfc-editor.org/cgi-bin/rfcsearch.pl?searchwords=LDAP&num=1500&format=ftp

[RCDL2000-1] Маршрутизация запросов в системах распределенного поиска. И.Некрестьянов, СПбГУ. RCDL-2000, http://www.protvino.ru/dl2000/reports/pdf/066.pdf

---

Об авторах

Бездушный Анатолий Николаевич - к.ф-м.н., с.н.с. ВЦ РАН. Сфера деятельности: компиляторы, базы данных, Интернет, Web, программирование, параллелизм, сети, информационно-поисковые технологии, интеграция данных .
Тел. (095)135-5471,135-5280,
e-mail: bezdushn@ccas.ru

Д. А. Ковалев - аспирант ВЦ РАН. Сфера деятельности: Интернет, Web, программирование,информационно-поисковые технологии, интеграция данных.
e-mail: dk@programmer.net

Серебряков Владимир Алексеевич - с.н.с., зав.отделом ВЦ РАН. Сфера деятельности: компиляторы, базы данных, параллелизм, программирование, сети, информационно-поисковые технологии.
Тел. (095)135-5471,135-5280,
e-mail: serebr@ccas.ru
