Технологии репликации данных и распределенного поиска в ИСИР РАН

А.Н.Бездушный <besdushn@ccas.ru>, Д. А. Ковалев <dk@programmer.net>,
В.А.Серебряков <serebr@ccas.ru>

Аннотация

В рамках проекта ИСИР РАН проводились исследования современных открытых протоколов, технологий, архитектур и систем для организации распределенной среды цифровых библиотек. На основе проведенного анализа были приняты решения относительно способов реализации распределенной гетерогенной среды информационной системы ИСИР РАН, изложению которых посвящена эта статья. Статья представляет соответствующую часть системы, используемые технологии - способы распределения и репликации данных, средства распределенного поиска и индексации и т.п. Данная работа выполняется в рамках проекта 99-07-90139 РФФИ “Интегрированная система информационных ресурсов РАН”.

Введение

Бурное развитие компьютерных, коммуникационных и особенно Интернет - технологий предоставляет возможность объединить информационные ресурсы в концептуально единую информационную среду. Организации РАН обладают значительными и постоянно возрастающими научными информационными ресурсами, но в большинстве своем они слабо систематизированы, существенно разрознены, как логически, так и физически. Многие представления научной информации преимущественно статические, плохо структурированы, не имеют средств каталогизации и поиска. Имеющиеся системы используют различные модели данных и схемы структурирования сходной информации, обладают разнообразными интерфейсами, ставя проблему организации единообразного информационного пространства из неоднородного набора составляющих систем. Эти факторы послужили предпосылками начала работ по проекту ИСИР РАН [BJKS, ABGKMS]. В рамках этого проекта ведется разработка программных средств для реализации распределенной гетерогенной информационной системы [BKS].

1. Анализ

Основная задача информационной системы - предоставить возможно более развитые средства обнаружения необходимой информации. В ходе разработки концепции ИСИР было проанализировано несколько основных подходов к организации информационно-поисковых систем.

Поисковые роботы. Этот класс систем использует развитие Интернет-технологий, решая задачу наибольшего охвата информации. Однако подобный объем информации может обрабатываться только автоматически, а отсутствие фиксированной структуры информации ограничивает возможности автоматической обработки, оставляя единственный критерий поиска - вхождение того или иного слова (фразы) в искомые документы. Поисковая информация, извлекаемая из документов (полнотекстовый индекс), концентрируется на поисковом сервере, выдающем в качестве ответа местонахождение и способ доступа к документу в виде URL. Основной недостаток такого подхода – это большое количество ресурсов, удовлетворяющих запросу. Большая часть формально удовлетворяющих запросу документов на деле не интересуют пользователя. Для улучшения качества поиска применяются несколько техник, автоматизированных и предполагающих участие экспертов, например, ранжирование найденных документов по близости вхождений искомых слов. Используется также гипертекстовая структура документа , в виде “индекса URL-цитируемости” документов (например, система Google PageRank [GOOGLE] и подобные). Имея наиболее широкий охват при фиксированных возможностях поиска, подход поисковых машин перестает масштабироваться при попытке улучшить возможности поискового сервиса. Две важные техники (зеркалирование и кэширование данных), позволяющие улучшить доступность и скорость работы

с Web-доступной информацией, нацелены на дублирование удаленной информации на серверах, более приближенных к обслуживаемой группе пользователей в смысле коммуникативных возможностей.

Распределенные системы. Этот класс систем ([HARVEST], [DESIRE]) использует принцип распределения ответственности за информацию между многими организациями, входящими в систему. Каждая из них поддерживает собственную коллекцию информации в виде документов, при этом имеет возможность ввести более или менее строгие правила оформления этих документов, позволяющие организовать автоматизированную выборку дополнительной информации для поиска. В основном, это библиографическая информация. Экстрагированная информация хранится отдельно на поисковом сервере, в виде набора атрибутов (например, в формате SOIF [SOIF/TIO]), описывающих документ, и указателя на его местонахождение (URL). Добавив к этому полнотекстовые индексы и информацию о гипертекстовой структуре коллекции, можно получить систему с более богатыми поисковыми возможностями - в поисковые критерии добавить условия на атрибуты (назовем этот вид поиска *атрибутным поиском*). Аналогичные принципы организации распределенного поиска развивались и в области справочных систем и интернет-служб, где информация также представлялась в виде набора атрибутов, описывающих объект [WHOIS++, LDAP]. Общий принцип совместного атрибутного поиска в таких распределенных системах сводится к тому, что информация о каждом документе независима от других, и ответ совокупной системы есть сумма ответов входящих в нее частей.

Балансировка нагрузки предполагает репликацию поисковой информации, часто - и самих данных, с маломощных серверов на более мощные. Ответственность за изменение и манипуляции с информацией при этом остается на исходном сервере, а внешнее ПО обеспечивает актуальность копии на мощном сервере, который отвечает на поисковые запросы как по своим, так и по реплицированным данным. Происходит концентрация поисковой информации на ограниченном числе мощных серверов, участвующих в ответе на поисковые запросы. Необходимо заметить, что специфика информационных систем, как правило, не требует немедленного распространения изменений, и поддержки каких-либо распределенных транзакций.

Маршрутизация запросов. Другая важная технология обеспечения эффективного распределенного атрибутного поиска - маршрутизация запросов. Основная задача этой технологии состоит в автоматическом выборе из множества серверов (потенциальных участников обработки запроса) подмножества, содержащего основную часть искомых документов. Выбор делается на основании относительно компактной предварительной информации о содержимом каждого из серверов. Техника маршрутизации, когда перенаправление запроса, сбор и предобработка ответов производятся на стороне “сервера” – поисковой системы, имеет также модификации. Протокол работы с сервером может предусматривать выдачу в ответ на поисковый запрос не (или не только) искомой информации, но и “предложений” обратиться на другой сервер с этим или модифицированным запросом (“referrals”), чтобы получить дополнения к ответу. Таким образом, задача опроса рекомендованых серверов и сбора консолидированного ответа перекладывается на клиента ([LDAP] v3).

Специализированные системы. В это класс систем, существенный с точки зрения задач ИСИР попадают различные информационные, справочные, экспертные и другие специализированные системы, эксплуатируемые и вновь разрабатываемые в организациях РАН (например, [JKM]), которые содержат массу представляющей интерес информации в структурированном виде. Важнейшими являются разнообразные библиотечные и справочные системы, хранящие наукоемкую информацию - информацию о публикациях, конференциях, проектах, сотрудниках организаций, связях, совместных программах и т.п. Важнейшим фактором улучшения структуры информации является введение разнообразных классов информационных единиц и понятия связей между ними - кроме документов, можно выделить персоны, организации, мероприятия и т.п. Это позволяет во многих случаях четко специфицировать в запросе и получить непосредственно интересующую информацию, а не набор документов,

возможно содержащих ее в том или ином виде. Наличие информации о связях между информационными единицами, а также их глобальная уникальная идентификация позволяет получить дополнительные возможности, такие как различение схожих ресурсов (однофамильцев и т.п.), косвенные поисковые критерии - возможность искать информацию непосредственно об интересующих объектах, имея сведения только о связанных с ними объектах.

Однако строгие требования к структуре информации ограничивают темпы роста таких коллекций, т.к. требуют огромного согласованного труда экспертов. Тем более важную роль играют технологии, обеспечивающие распределенный совместный поиск по этим системам. При этом задача, по сравнению с описанными выше системами, усложняется целым рядом дополнительных проблем. Объединяемые системы используют различные СУБД. Системы достаточно сильно разбросаны географически, и средства связи между системами не всегда гарантируют определенный уровень связи. Эти обстоятельства не позволяют использовать средства СУБД для организации распределенного поиска.

2. Архитектура, сервисы распределенной среды ИСИР

2.1. Модель данных, определения

Как следует из приведенного анализа, ценность специализированных систем с точки зрения задач ИСИР состоит, прежде всего, в наличии хорошо структурированной информации. Для того чтобы разумно использовать эту структуру информации в системе типа ИСИР, необходима модель данных, максимально близкая к уровню ER-моделирования.

Используемая распределенной средой ИСИР модель основывается на предложениях Kahn и Wilensky [KW] и состоит в следующем. . Каждый информационный *ресурс* хранится некоторым *репозиторием* как глобально и уникально именуемый [HDL] набор структурированных данных (*сведений о* ресурсе, его свойств, атрибутов, связей) и, возможно, *содержания*, например, одного или более форматов представления каталогизируемого ресурса. Эти структурированные данные, описывающие ресурс, а в связи с этим называемые *метаданными*, используются для организации поиска и каталогизации ресурсов. Под этими терминами мы понимаем следующее.

Информационный ресурс – это единица информации, представляющая собой уникально именуемый набор данных, структурированных в виде именованных атрибутов. Ресурсы могут находиться в различных отношениях с другими ресурсами. Любой информационный ресурс принадлежит к некоторому классу, определяющему, исходя из его назначения, набор возможных атрибутов и отношений. Пример информационного ресурса – объект класса “публикация” с атрибутами “название”, “аннотация”, “ключевые слова”, “ISBN”, “текст”, находящийся в отношении “авторства” с объектом класса “персона”. Атрибут “текст”, в свою очередь, может структурировать свое содержимое по разделам и т.д. Каждая из участвующих в формировании распределенной среды систем рассматривается (вне зависимости от реального назначения, которое может сильно отличаться от задач поиска информации) как хранилище информационных ресурсов (репозиторий).

Под репозиторием понимается средство хранения информации, предоставляющее некоторый четко специфицированный способ (интерфейс - *схему данных*, *модель операций*) для управления ею, включающий в той или иной мере способы доступа, выборки и манипулирования информационными ресурсами. Для физического хранения информации в репозитории может использоваться один или несколько *узлов*.

Узел - физическая единица, обеспечивающая единообразное хранение всех или части ресурсов репозитория.

Коллекция – это совокупность информационных единиц (информационных ресурсов, ИР), объединенных общими свойствами (например, общей принадлежностью, или тематической направленностью).

Такой достаточно высокий уровень абстракции позволяет концептуально объединить информацию из разных систем, во всем разнообразии используемых ими моделей более низкого уровня, описывая их в единой терминологии атрибутированных ресурсов и связей между ними. В то же время он достаточен для рассматриваемых нами задач поиска (см. ниже). Различие - как структурное, так и семантическое - "аналогичных" объектов разных репозиториев выявляется в ходе сопоставления *схем данных репозитория*.

Схема данных репозитория - набор формальных опеределений, фиксирующих состав допустимых для репозитория классов ресурсов, состав и строение атрибутов ресурсов этих классов, допустимые виды связей между ресурсами и разные ограничения на данные ресурсов.

Важно заметить, что обсуждавшиеся ранее классы систем с точки зрения задач ИСИР также представляют собой репозитории или наборы репозиториев. Разница состоит только в отсутствии отношений между ресурсами в схеме этих репозиториев (или наличии только связей, обусловленных изначальным гипертекстовым представлением), и использовании только планарной структуры атрибутов.

2.2. Задача интеграции

Обозначенная во введении общая задача ИСИР состоит в организации единого информационного пространства РАН. Сюда входят задачи по выборке и структуризации метаданных из различных электронных представлений, а также средства их ввода в структурированном в виде. Эти задачи вкратце упомянутые выше (поисковые машины, Harvest/Essence и т.д.) очень важны, в рамках проекта ИСИР проводится отдельная работа по поддержке интерфейсов с соответствующими системами. Второй класс задач, решению которых посвящена данная статья, состоит в предоставлении средств интеграции информации разнообразных информационных систем (репозиториев), тем или иным способом накопивших структурированную информацию.

Под интеграцией мы понимаем следующее. Распределенная система ИСИР РАН ориентируется на объединение организаций, каждая из которых поддерживает коллекцию ресурсов, представляющих общий интерес. 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 ИСИР, 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 предоставлять средства обслуживания и машрутизации поисковых запросов и их результатов, предоставление 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.

2.3. Набор сервисов ИСИР

Поставленная задача интеграции решается с использованием сервисной архитектуры, базирующейся на опыте таких реализаций, как Desire, NCSTRL и т.д.

Репозиторий. 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, предоставляет различные возможности по работе с данными, и т.д.

Унифицирующий интерфейс репозитория. 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 программ-оболочек (wrapper) и программ-посредников (mediator) [Wie92]. Во многих системах стремятся описывать эти программы декларативными средствами. Например, в системе TSIMMIS[Gar97, TSIMMIS] разработан логический язык запросов MSL над OEM моделью, который используется как язык для описания оберток и посредников с целью их последующей генерации. В своей работе мы выделили процедурную прослойку доступа к источнику данных - “унифицирующий” интерфейс доступа к репозиторию. Задача этой прослойки – используя возможности ПО “локального” репозитория, предоставить сервисам более высокого уровня(обмена и интерации) минимально необходимый набор операций с данными, в терминах атрибутированных информационных ресурсов, ограниченных соответствующим образом описанной схемой данных репозитория. Сервисы более высокого уровня параметризуются формальным описанием схемы данного репозитория, и могут работать с любым репозиторием, реализующим унифицирующий интерфейс. Набор операций, минимально необходимый для работы сервисов интеграции, включает:

1) **добавление** ресурсов, по предоставлении минимально необходимого набора атрибутов;
2) **изменение значений** каждого атрибута в отдельности по указании уникального (для репозитория) идентификатора ресурса и новых значений атрибута;
3) **выборка** всего ресурса или его части его данных (указанного набора атрибутов) по уникальному идентификатору;
4) **выборка** совокупности ресурсов, удовлетворяющих условиям, выраженным на некотором формальном языке.

Для представления передаваемых данных используется модель RDF, в которой модель атрибутированных ресурсов имеет прямое отражение. Схема данных - набор допустимых классов ресурсов и т.п. - формально выражается на языке RDF-schema, дополненном набором дополнительных ограничений в рамках стандартного синтаксиса. Операции с унифицирующим интерфейсом репозитория обеспечиваются RDF- загрузчиком и генератором. Первый принимает RDF/XML документ, проверяет его соответствие модели RDF-schema, и загружает описанные ресурсы в репозиторий. Второй, используя набор условий на искомые ресурсы, список атрибутов ресурсов, подлежащих выгрузке, выдает RDF/XML документ соответствующий RDF-schema. Предусматривается реализация вышеуказанных RDF- загрузчиков и генераторов для репозиториев, поддерживающих интерфейс JNDI[JNDI]. Этот интерфейс поддерживает все необходимые манипуляции для случая иерархически организованного множества планарных ресурсов, и при этом полностью параметризуется именами элементов схемы. Это позволяет реализовать “универсальные” загрузчик и генератор RDF для JNDI-репозиторий, которые использует 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. Достаточно много информационных систем, в первую очередь, поддерживающих протокол LDAP[LDAP], уже имеют JNDI-адаптеры.

Аналогично JNDI, в дальнейшем планируется реализация “универсальных” загрузчиков и генераторов для таких популярных протоколов, как Z39.50[Z39.50], SDLIP[SDLIP]. Эти 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.

Локальный поисковый сервис. Из-за разнообразия программных средств, моделей данных, неравнозначности предоставлемых поисковых возможностей и т.д., не представляется возможным непосредственно воспользоваться собственными поисковыми службами репозиториев для обеспечения операций распределенного поиска. Поэтому мы выделили и предоставляем локальный поисковый сервис, использующий унифицирующий интерфейс репозитория для выборки всей или указанной части информации репозитоория. Этот сервис обеспечивает такой уровнь взаимодействия с репозиториями, на котором имеются гарантированные возможности по поиску и выборке информации, ограниченные в каждом случае собственной схемой данных репозитория. Поисковые возможности сервиса включают:

- возможность атрибутного поиска, т.е. возможность задавать условия на значения атрибутов искомых ресурсов
- возможность полнотекстового поиска по текстовым атрибутам,
- возможность поиска с одним уровнем косвенности, т.е. возможность задавать атрибутные условия на ресурсы, непосредственно связанные с искомыми.
- возможность комбинировать условия, используя связки “и”, “или”, “не”.

Сервис может извлекать из репозитория и возвращать клиенту значения только тех атрибутов, которые могут использоваться в операциях поиска (поисковая информация), для текстовых атрибутов используется их компактное представление в виде полнотекстовых индексов (*локальные поисковые индексы*).

На основании анализа различных поисковых систем, реализующих ту или иную часть необходимой функциональности, в ИСИР реализуется следующее решение. 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. Программа-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), *полнотекстовые* и *ключевые слова* (значения атрибутов разбиваются на отдельные термы), *связи* (поддерживаются только двусторонние связи).

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

Сервис именования. Важно заметить, что для организации распределенного функционирования необходима глобальная система идентификации ресурсов во всех репозиториях, независимая от схем идентификации в каждом из репозиториев. Проблема глобальной идентификации ресурсов решается в ИСИР в предположении, что в задачи каждого из интегрируемых репозиториев входит поддержка идентификаторов собственных ресурсов, уникальных в пределах репозитория. ИСИР использует внешнюю службу именования (Handle System [HDL]) для назначения ресурсам глобальных идентификаторов - URI. Служба именования используется сервисами ИСИР 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.

Сервис распределенного поиска. 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.

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, концентрирующий поисковую информацию. Копии размещаются последним в структурах локального поиска, наряду с информацией о его собственных ресурсах - содержимое таблиц ресурсов и словарей индексов смешиваются, а таблицы вхождений объединяются.

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-QR]. 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.

Сервис обмена данными/репликации данных. Сервисы распределенной системы предполагает автоматизированный обмен информацией между репозиториями, происходящий на постоянной основе и позволяющей 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

Эти виды обмена укладываются в общую модель обмена сообщениями, широко используемую в задачах интеграции и распределенных коммуникациях. Две основные модели обмена – PTP (point-to-point) и PS (publisher-subscriber) предоставляют гибкие средства конфигурирования обмена. ИСИР реализует настраиваемый “сервис обмена”, поддерживающий обе модели и реализующий требуемые виды обмена. Архитектура сервиса позволяет использовать ему разные транспортные протоколы, службы. Имеется поддержка протокола CIP[CIP] и Java интерфейса JMS[JMS].

С точки зрения потоков данных в рамках обмена для поддержки распределенного поиска, распределенная система представляет собой некоторый граф, в котором можно выделить (взаимопересекающиеся) опорные иерархии. Один опорный “лес” деревьев - это структура потоков реплик, используемых для балансировки нагрузки. Эта структура может частично совпадать с административной структурой подчинения организаций, может быть основана исключительно на договоренности между администраторами репозиториев, исходя из мощностей их оборудования. Структура может быть выбрана с учетом перераспределения реплик в тематические коллекции, для улучшения качества и уменьшения стоимости маршрутизации. Достаточно хорошо совпадает с этим и направление обмена индексами (описателями коллекций). Индексы и поисковая информация накапливаются на мощных серверах, одновременно перераспределяясь в тематические коллекции. Далее, сокращенные описания тематических

коллекций передаются и накапливаются на нескольких “точках входа” в систему, осуществляющих первый шаг маршрутизации.

Отдельные группы организаций могут обеспечивать более плотное взаимодействие между своими репозиториями, настраивая поддержку связей, а также предоставляя поисковые интерфейсы – точки входа для своих “подсистем”.

Сервис метаописаний. 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 ([RDFS, DAML, OIL]), и регистрируются в службе именования.

2.4. Отображение схем

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 данной постановке задачи невозможно. С другой стороны, поисковые запросы пользователей к распределенной системе формулируется для некоторой вполне определенной, обычно, широко распространенной схемы. 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 архитектура ИСИР предусмотрены следующие возможности преобразования данных:

- На этапе создания унифицирующего интерфейса репозитория. Оболочка может отобразить структуры данных локального репозитория в схему ресурсов, максимально приближенную к одной из “стандартных”.
- На этапе обмена с помощью компонент-преобразователей.
- На этапе создания локальных поисковых индексов и описателей коллекций. Схема репозитория может видоизменяться, например, за счет слияния нескольких атрибутов в один индекс, переименования и т.п.
- На этапе обработки запроса. При перенаправлении запроса некоторому репозиторию, программы-посредники могут переформулировать исходный запрос в запрос, обращенный непосредственно к репозиторию.

Возможности преобразования существенно различны на разных этапах, и решают различные задачи. Первые два предоставляют наиболее широкие возможности преобразования, и предполагают полное семантическое преобразование, позволяющее прямой обмен данными между репозиториями. Преобразования этого класса совершаются компонентами-медиаторами над данными в RDF-представлении, на основании формальных описаний отображения схем (см. [DAML+OIL]). Вторые две возможности ориентированы на сохранение семантики поиска, и допускают потерю структуры.

2.5. Связи между ресурсами разных репозиториев, дубликаты

Введение сервиса уникальной глобальной идентификации позволяет хранить информацию, 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 определенных условий целостности. 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.

3. Заключение

В статье описан подход к разработке гетерогенной распределенной информационной системы ИСИР РАН. Система состоит из нескольких взаимосвязанных сервисов, поддерживающих единую среду. Реализация системы ведется на основе открытых стандартов, с использованием языка Java и его технологий для достижения кросс-платформенной переносимости. Внимательно изучается опыт схожих открытых проектов ([HARVEST,DESIRE,IMESH,ISAAC] и др.), с целью обеспечения интероперабельности. К сожалению, ограничения на объем не позволяют изложить технические решения ИСИР более подробно. Значительно больший по объему вариант статьи можно найти на сайте проекта.

4. Литература

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

[BJKS] Бездушный А.Н., Жижченко А.Б., Кулагин М.В., Серебряков В.А., Интегрированная система информационных ресурсов РАН и технология разработки цифровых библиотек, Программирование, том. 26, N. 4, 2000, pp. 177–185.

[BKS] А.Н. Бездушный, Д. А. Ковалев, А.А. Филиппова, Использование протокола LDAP для поддержки распределенности гетерогенных информационных систем, Электронные библиотеки: перспективные методы и технологии, электронные коллекции, 2-я вероссийская конференция, Протвино, 2000 г.

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

[DAML+OIL] The DAML language is being developed as an extension to XML and the Resource Description Framework. The latest release of the language (DAML+OIL) provides a rich set of constructs with which to create ontologies and to markup information so that it is machine readable and understandable. http://www.daml.org

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

[DESIRE] The DESIRE Project (http://www.desire.org). Development of a European Service for Information on Research and Education

[Gar97] H. Garcia-Molina, Y. Papakonstantinou, D. Quass, A. Rajaraman, Y. Sagiv, J. Ullman, and J. Widom. The TSIMMIS project: Integration of heterogenous information sources, March 1997.

[GINF] GINF (http://www-diglib.stanford.edu/diglib/ginf/). General Interoperability Framework based heavily on RDF.

[GOOGLE] Google Search Technology. http://www.google.com/intl/ru/technology/index.html

[Harvest] Harvest is an integrated set of tools to gather, extract, organize, search, cache, and replicate relevant information across the Internet. http://harvest.transarc.com/

[HDL] A general-purpose global name service enabling secure name resolution over the Internet. http://www.handle.net/

[ICE] The Information and Content Exchange protocol. http://www.w3.org/TR/NOTE-ice

[IMESH] The IMesh toolkit (http://www.imesh.org/toolkit). An architecture and toolkit for distributed subject gateways.

[Isaac] The Internet Scout Project, Isaac network. (http://scout.cs.wisc.edu/research/isaac/index.html). Uses LDAP and CIP for building distributed resource discovery service.

[JKM] Жижимов О.Л., Коджесян В.С., Мазов Н.А. Пример распределенной информационной системы на основе метаданных и международных стандартов, Электронные библиотеки: перспективные методы и технологии, электронные коллекции, 2-я вероссийская конференция, Протвино, 2000 г.

[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/

[KW] Robert Kahn, Robert Wilensky, A Framework for Distributed Digital Object Services, May 13, 1995, cnri.dlib/tn95-01, http://www.cnri.reston.va.us/home/cstr/arch/k-w.html

[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

[MARC] The Network Development and MARC standards office. http://www.loc.gov/marc/

[QR] J.Kirriemur et al. Cross-searching Subject Gateways. The Query Routing and Forward Knowledge Approach. http://www.dlib.org/dlib/january98/01kirriemuir.html
[RCDL2000-1] Маршрутизация запросов в системах распределенного поиска. И. Некрестьянов, СПбГУ. Материалы RCDL-2000 http://www.protvino.ru/dl2000/reports/pdf/066.pdf
[RDF] The Resource Description Framework (RDF) integrates a variety of web-based metadata activities including sitemaps**,** content ratings**,** stream channel definitions**,** search engine data collection (web crawling), digital library collections, and distributed authoring, using XML as an interchange syntax. The RDF specifications provide a lightweight ontology system to support the exchange of knowledge on the Web. http://www.w3.org/RDF
[RDFschema] This specification describes how to use RDF to describe RDF vocabularies. The specification also defines a basic vocabulary for this purpose, as well as an extensibility mechanism to anticipate future additions to RDF. http://www.w3.org/TR/rdf-schema
[SCI99] I. Kuralenok, V. Dobrynin, I. Nekrestyanov, M. Bessonov and A. Patel. Distributed search in topic-oriented document collections. In Proc. Of World Multiconference on Systemics, Cybernetics and Informatics (SCI'99), volume 4, August 1999.
[SDLIP] Simple Digital Library Interoperability Protocol. http://www-diglib.stanford.edu/~testbed/doc2/SDLIP
[SIGIR95] James P. Callan, Zhihong Lu, and W. Bruce Croft. Searching distributed collections with inference networks. In Proc. of the SIGIR'95.
[SOAP] Simple Object Access Protocol is a lightweight XML based protocol for exchange of information in a decentralized, distributed environment. http://www.w3.org/TR/SOAP
[SOIF/TIO] A Tagged Index Object for Use In Common Indexing Protocol. ftp://ftp.isi.edu/in-notes/rfc2655.txt
[TSIMMIS]www.db.stanford.edu/tsimmis
[UDDI] Universal Description, Discovery and Integration specification. http://www.uddi.org
[Wie92] G. Wiederhold. Mediators in the architecture of future information systems. In IEEE Computer 25:3, pp. 38-49.
[WSDL] Web Services Description Language. http://msdn.microsoft.com/xml/general/wsdl.asp
[XML] The Extensible Markup Language (XML) is the universal format for structured documents and data on the Web. http://www.w3.org/XML
[Z39.50] Z39.50 Resource Page. http://www.niso.org/z3950.html