Возможности технологий ИСИР в поддержке Единого Научного Информационного Пространства РАН

Бездушный А. А. Бездушный А. Н. Нестеренко А. К. Серебряков В. А. Сысоев Т. М.
МФТИ ВЦ РАН ВЦ РАН ВЦ РАН ВЦ РАН
alix@7ka.mipt.ru bezdushn@ccas.ru alexn@ccas.ru serebr@ccas.ru tim@7ka.mipt.ru

Аннотация

Работа посвящена вопросам обеспечения интеграции информационных ресурсов РАН, информационной поддержки научных исследований в рамках Единого Научного Информационного Пространства (ЕНИП) РАН. Рассматриваются потребности, цели и задачи организации ЕНИП РАН, как среды взаимосвязанных распределённых гетерогенных систем. Обсуждается возможности применения технологии ИСИР для обеспечения решения этих задач, которые могут использоваться как для формирования инфраструктурных элементов среды, так и для создания адаптеров существующих систем и реализации новых с широкой гаммой вариации соответствующих служб. Кратко рассматривает архитектурные принципы последних разработок ИСИР, применение Semantic Web-технологий и многоуровневое деление архитектуры, инструментарий разработчика ИСИР и разработка информационных систем и Web-порталов на базе технологий ИСИР.

1 Проблемы и задачи ЕНИП РАН

В мире имеется достаточно большое количество информационных систем для работы с научными данными, наукоемкой информацией. Каждое научное учреждение имеет данные о публикациях сотрудников, о проводившихся или ведущихся научных исследованиях и проектах, располагает теми или иными наукоемкими данными. Многие из учреждений имеют собственные информационные системы для наукоемкой информации, которые в каком-то виде ее хранят.

Однако интересы пользователей не исчерпываются какой-то одной информационной системой, пусть даже замечательной. Как правило, интересы ученых шире поставленной задачи, находятся на стыке научных областей. Практически не возможно в рамках одной организации собрать информацию, которая удовлетворила бы запросы всех сотрудников. Даже, если бы в какой-то момент это удалось, то в силу огромного динамизма научных исследований невозможно было бы обеспечить какую-то приемлемую полноту, актуальность данных, для которых системы служит.

Попытки объединить данные научных учреждений в одной централизованной системе на достаточно высоком уровне не могут привести к положительному результату. Это можно увидеть на примере системы ERGO [13] и финского проекта создания национальной университетской системы[14]. Препятствиями на этом пути являются как объемы информации, так и сложность обеспечения полноты, актуальности данных, невозможность сведения данных всего разнообразия научных областей к единой структуре.

Российская Академия Наук объединяет большое число научно-исследовательских учреждений и коллективов, вовлеченных во все многообразие видов научной деятельности. Учреждения обладают уникальными научными информационными ресурсами. Однако информационные ресурсы остаются существенно разрознены, недостаточно систематизированы и структурированы, слабо применяются соглашения по стандартизации электронного представления и взаимодействия информационных ресурсов, соответствующие средства, призванные поддержать интеграцию информационных ресурсов, повышение точности поиска и т.п. [17]

В этой связи инициатива по организации Единого Научного Информационного Пространства РАН (ЕНИП РАН) призвана помочь научным коллективам сделать ряд шагов в направлении интеграции разнородных научных информационных и программных ресурсов отдельных научных учреждений, предоставлении пользователям более эффективных средства интеграции и поиска информации, научной коммуникации, сотрудничества и совместной работы. Под единым пространством понимается ни формирование централизованной системы, ни навязывание всем одних и тех же решений, а стремление последовательностью практических шагов, совместными усилиями научных коллективов РАН

- сформулировать взаимосогласованный набор соглашений, правил и открытых стандартов;
- приготовить совокупность макетов и типовых решений для реализации адаптеров прикладных

Труды 6ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL2004, Пущино, Россия, 2004.

систем, инфраструктурных служб, поддерживающих разные уровни интероперабельности распределенных гетерогенных данных и приложений;

- создать ряд информационных систем общего назначения, следующих этим соглашениям, использующих эти реализации, допускающих модульную организацию, наращивание функциональных возможностей;
- применить эти результаты для решения соответствующих задач учреждений РАН.

Все нацелено на то, чтобы помочь учреждениям РАН в решении общих информационных задач, в достижении требуемой интеграции с другими учреждениями РАН.

В общем случае можно сказать, что информационные системы учреждений РАН отличают огромные объемы и низкая структурированность данных, распределенный характер, неоднородность, независимость и разные условия сопровождения, управления и политики доступа к информационным источникам и сервисам. В таких случаях выделяю и стараются решить проблемы общего вида, среди которых следующие:

Техническая интероперабельность. Для обеспечения взаимодействия между разнородными информационными источниками необходимо поддерживать согласованные интерфейсы, протоколы и механизмы доступа к информационным ресурсам.

Синтаксическая интероперабельность. Данные, доступные из информационных источников, как правило, отличаются синтаксической неоднородностью, разнообразием моделей данных и форм представления данных. Следовательно, необходимо выработать и согласовать унифицирующий подход приведения данных к наиболее распространённым моделям данных и форматам.

Сбор метаданных. В сложившейся ситуации, когда сведения о ресурсах в большом объеме представлены в виде слабоструктурированного текста, когда поисковые системы осуществляют полнотекстовый поиск нужных данных по запросам в свободной форме, пользователь получает огромное количество «шумовой» информации, среди которой очень трудно выбрать действительно полезные знания. Учитывая это обстоятельство, для представления сведений о ресурсах стали использовать структурное представление, выделять понятие метаданных, описывающих содержимое ресурса в виде набора именованных значений, в том числе связей с другими ресурсами. Метаданные используются для автоматизированного анализа содержимого ресурса, построения поисковых индексов и позволяют обеспечить достаточно высокую точность и эффективность поиска разнотипной информации. Центральной компонентой в обслуживании слабоструктурированных и унаследованных коллекций информации является процесс "сбора" метаданных, в ходе которого из коллекций в соответствии с требованиями синтаксической интероперабельности извлекаются и структурируются метаданные, формируется индексная информация для обеспечения локального поиска, маршрутизации распределенных запросов, ранжирования результатов запросов.

Семантическая интероперабельность. Метаданные могут относиться к различным предметным областям, в рамках одной иметь разные выражение и интерпретацию. Создание и согласование стандартных прикладных профилей метаданных и онтологий упростит интеграцию разнообразных систем, позволит автоматизировать обмен метаданными, их обработку и преобразование, повысить точность и эффективность поиска. Глубина структуризации метаданных о ресурсах должна определяется задачами конкретной системы. В узкопрофессиональных системах она является высокой с тем, чтобы поддержать соответствующие процессы, возможность проведения специальных исследований. Тем не менее, для общих задач интеграции информационных ресурсов высокая степень структуризации не требуется и усложняет процесс. Необходимо выработать подход к наращиванию степени структуризации метаданных, который позволил бы специализировать общие схемы метаданных под потребности конечных приложений. Разработать набор элементов метаданных для общей научной информации и некоторые профили метаданных конкретных научных областей, согласуя их с научным сообществом и международными стандартами. Обеспечить выделение и согласование стандартных классификаторов ресурсов и тезаурусов.

Поддержка глобальной идентификации ресурсов. Использование глобально уникальных идентификаторов дает возможность установления взаимосвязей между ресурсы разных репозиториев (под репозиторием мы понимаем интероперабельный информационный источник, в указанном выше смысле) распределённой среды, объединять связанные данные отдельных репозиториев в виртуально-единые ресурсы. Это предоставит пользователям возможность производить навигацию среди ресурсов всей информационной системы, выполнять косвенный поиск, в том числе и по связям между ресурсами в разных репозиториях, упрощает задачу объединения результатов поисковых запросов разных репозиториев.

Совместный «поиск» - маршрутизация запросов и объединение ответов. Для понижения нагрузки на сеть и повышения эффективности распределенные запросы должна выполняться не во всем множестве репозиториев, а только в соответствующем запросу подмножестве. Этот процесс, называют маршрутизацией запросов, при принятии решения использует “предварительные

знания” - информацию, распространяемую в среде именно с целью обоснованной рассылки поисковых запросов, формируемую на основе локальных индексов. Процесс объединения ответов репозиториев, к которым был направлен запрос, в единый ответ системы должен обеспечивать как устранение вторичных вхождений описаний одного и того же ресурса (дублирования описаний), которые с большой вероятностью могут появиться из разных частей распределенной среды, так и обеспечение совместного ранжирования результатов, поступающих от этих частей.

Балансировка нагрузки. Для снижения нагрузки на телекоммуникационные и вычислительные ресурсы при обработке запросов, при доступе к часто используемой информации применяются механизмы балансировки нагрузки. Балансировка нагрузки предполагает репликацию метаданных с маломощных серверов на более мощные. В этом случае происходит концентрация поисковой информации на ограниченном числе мощных серверов, участвующих в ответе на поисковые запросы. В рамках обмена и репликации данных встают проблемы обеспечения связывания и интеграции ресурсов независимо сопровождаемых источников информации, выявления дубликатов.

Распределенная авторизация доступа и принцип единой аутентификации. Различные информационные источники, составляющие распределённую среду, имеют различные механизмы контроля доступа к информации. Средства контроля доступа должны быть также предоставлены и интегрированным средой, должен быть указан общий подход к безопасности систем. Для того чтобы избавить пользователя от необходимости регистрироваться в каждом информационном источнике, должен быть поддержан принцип единого входа.

Возможным способом эффективной интеграции данных в контексте потребностей Российской Академии Наук. Организации, обладающие крупными цифровыми коллекциями, например, библиотеки, смогут с помощью решений ЕНИП предоставить доступ к своим информационным ресурсам в рамках. Научные учреждения, не обладающие подобными ресурсами, могут обеспечить возможность индексирования своих данных, экспорт данных и агрегацию их в более крупные хранилища. Такие хранилища могут предоставляться крупными учреждениями, научными центрами и отделениями РАН. Таким образом, можно поддержать агрегацию научных данных в соответствии с организационной структурой РАН (рис. 1). Находящиеся на верхнем уровне информационные хранилища отделений, в основном поддерживая профильные коллекции информации и услуг, могут участвовать в маршрутизации запросов объединенной среды, доступ к которой открывается через «единое окно», портал научного пространства РАН.

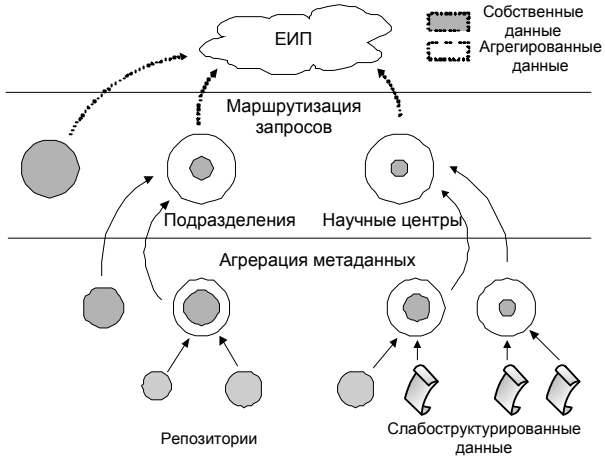

Рисунок 1. Агрегацию данных в соответствии с организационной структурой РАН.

Научные учреждения заинтересованы в получении и предоставлении доступа к данным о научных достижениях, научной деятельности сотрудников и организаций. Эта информация представляет интерес для конечных пользователей системы. Она позволит сотрудникам получить информацию о смежных работах в других коллективах. Организации могут предоставлять свои данные в слабоструктурированном виде, например как текстовые или XML-документы. Формат таких документов, как правило, отражает аспекты конкретной предметной области и сложившихся потребностей организации. Кроме того, многие организации поддерживают собственные сайты, заполненные статическими страницами, и необходимо предоставить им простой инструментарий, который позволил бы как управлять сайтами через web-интерфейс, так и агрегировать их информацию в объединенное научное пространство.

Для интеграции таких данных в систему необходимо выделять в них структуру, приводить к агрегатным моделям и схемам данных и связывать между собой и с имеющимися в среде данными. К примеру, организация может предоставлять информацию о научных публикациях, отчётах, о результатах научной деятельности за некоторый промежуток времени. Желательно, чтобы информация была автоматизированным образом проанализирована и структурирована (рис. 2) с выделением авторов публикации, издательства, и других атрибутов – названия, ISBN и пр. При интеграции данных может выясниться, что в хранилище уже имеется некоторая информация об авторах публикации, например другие статьи, или информация об участии в конференциях и пр., желательно, чтобы было обеспечено автоматизированное сопоставление и связывание соответствующих ресурсов.

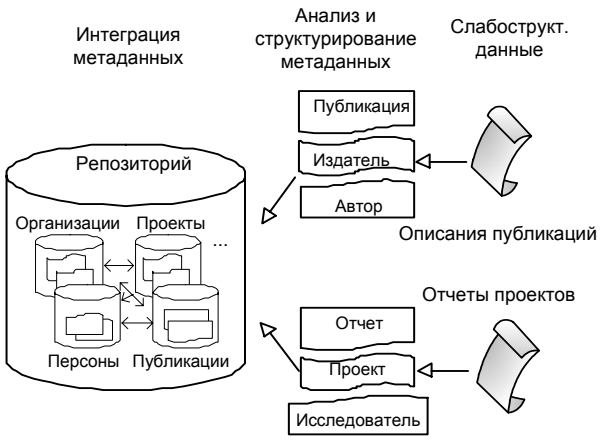

Рисунок 2. Анализ и структуризация слабоструктурированных данных.

Для решения вышеуказанных задач необходимо сформировать и стандартизовать тематические профили метаданных[10], их модели данных, определить средства записи схем данных, правила их специализации и согласования, выработать типовые интерфейсы взаимодействия [15]. Необходимо предоставить учреждениям РАН типовые программные и инструментальные средства [16], которые упростят или обеспечат участие учреждений РАН в едином информационном пространстве, структуризацию данных и их интеграцию в единую среду. Помимо прочего, стоят задачи создания сети взаимосвязанных инфраструктурных или агрегатных узлов, скорее всего, соответствующих структурным и тематическим объединениям РАН (отделения, научные центры РАН), предоставления доступа к среде в целом, ее соответствующим (объединениям) фрагментам через «единое окно».

2 Технологии ИСИР - архитектура и средства разработки

Рассмотренные выше видение и требования к формированию ЕНИП РАН в определенной степени могут быть поддержаны решениями на основе кросс-платформенной технологии ИСИР[1-7]. Технологии ИСИР позволяют поддержать многие элементы жизненного цикла распределенного гетерогенного информационного пространства. Помимо средств интеграции систем, программными компонентами ИСИР предоставляется также ряд средств для построения сложных информационных корпоративных порталов, шлюзов к информационным системам. Архитектура ИСИР базируется на открытых технологиях и стандартах [11, 12]. Следование стандартам мы считаем ключевой позицией в решении задач интеграции данных и приложений.

Для интеграции информации, соответствующей различным моделям данных, необходима каноническая модель данных, которая была бы наиболее удобна для решения рассматриваемых вопросов. Прежде всего, желательно, чтобы эта модель данных соответствовала «объектной» парадигме, поскольку это более абстрактная, существенно более богатая семантически и более естественная форма представления информации, например, чем реляционная модель данных, подходящая для задач хранения, поиска и получения информации. Во-вторых, желательно следовать последним Web-стандартам, в первую очередь требованиям XML-технологий - представление данных в форме XML весьма удобно для обмена информацией в Интернет, и обеспечивает требуемый уровень синтаксической интероперабельности. В-третьих, модель данных должна допускать существование распределенной информации.

Как следствие этих требований, мы выбрали технологии Semantic Web [2,3,9] как базис архитектуры. Этот W3C проект продолжает линию эволюции Web – от гипертекста к структурированным XML-документам, и далее к эффективной машинной обработке данных и интеграции разбросанной в Web информации. Resource Description Framework (RDF), модель данных Semantic Web, представляет собой обобщение ER-модели данных для их применения в Web. RDF модель данных хорошо согласуется с требованиями концептуального проектирования. Для записи RDF-данных W3C-спецификация предлагает XML-синтаксис (RDF/XML). Это XML-представление «объектных» данных используется нами для всех задач, связанных с обменом и представлением информации. Язык RDF Schema позволяет определять структурированные словари метаданных. Применение стандартных словарей свойств информационных ресурсов, предоставляемых, в частности, Dublin Core Metadata Initiative (DCMI), Publishing Requirements for Industry Standard Metadata (PRISM) и пр., гарантирует высокую степень семантической интероперабельности, и облегчает интеграцию данных. RDFS служит основой для более богатых языков описания схем, в частности Web Ontology Language (OWL), который используется в архитектуре ИСИР для определения схем данных и онтологий предметных областей.

Архитектура интегрированной распределённой среды, формируемой на базе решений ИСИР, следует традиционному многоуровневому строению, принятому в федеративных базах данных и системах на базе механизма посредников. В подходе цифровых библиотек такому строению следует, например, архитектура MIA[11]. В нашем случае специализация архитектуры связана с применением технологий Semantic Web (рис. 3). Нижний уровень соответствует разнородным провайдерам данных, информацию которых необходимо интегрировать в среду. Уровень адаптеров обеспечивает единообразный доступ к данным репозиториев, отображение модели данных провайдера в каноническую модель данных репозиториев, в нашем случае модель данных

Semantic Web. Каждый источник погружается в стандартную оболочку, предоставляемую ИСИР, либо должен самостоятельно поддерживает согласованные интерфейсы, протоколы, модель данных и язык запросов. Провайдер данных и оболочка могут быть распределены по сети. На данный момент имеются адаптеры для RDBMS, ODBMS и RDF источников информации, ведется разработка поддержки прозрачного взаимодействия с LDAP каталогами. Для спецификации канонической (интегрирующей) схемы данных репозитория, объединяющей «схемы экспорта» отдельных источников информации, мы используем язык RDFS схем, а точнее подмножество Web Ontology Language (OWL).

Специализированные компоненты-посредники, надстраиваемые над адаптерами, позволяют репозиториям участвовать в обмене и репликации данных, маршрутизации запросов между узлами. Эти компоненты позволяют прозрачно объединить данные информационных источников в интегрированную информационную среду, которая соответствует «интегрирующей» схеме данных. Интегрированные данные могут представляться различным клиентам, в частности имеется возможность построения web-порталов к информационной системе, а также поддержка протоколов SOAP, Z39.50 и SDLIP.

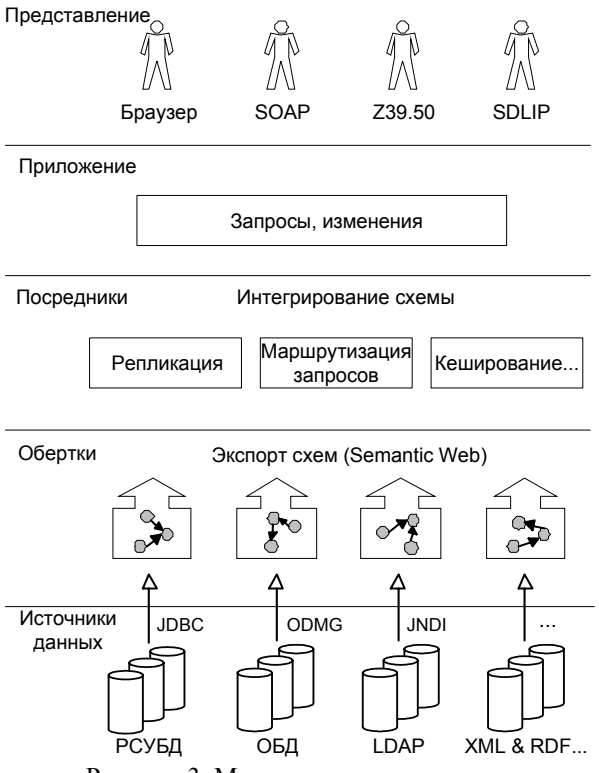

Рисунок 3. Многоуровневая архитектура.

Решения ИСИР содержат стандартные средства, обеспечивающие функционирование всех уровней архитектуры. Стандартная реализация уровня адаптеров базируется на так называемом ядре ИСИР, обеспечивающем отображение модели данных провайдера в модель данных языка Java. Ядро следует стандартизуемому подходу к прозрачному хранению Java-объектов в различного рода СУБД.

Фреймворк (средства поддержки и разработки) ИСИР в определенной степени автоматизирует настройку ядра на источник. Рассмотрим случай провайдера данных в виде реляционной базы данных, как наиболее распространённый на практике. При разработке нового репозитория с новой БД необходимо лишь представить схему данных, которые планируется поставлять в интегрированную среду, - это RDFS-схема данных или OWL-онтология предметной области. Инструментарий фреймворка предоставляет настраиваемую генерацию по OWL схеме данных DDL скрипта реляционной БД, спецификаций объектно-реляционного отображения и хранимых Java-классов, используемых ядром ИСИР для представления реляционных данных в объектном виде. Системные сервисы ИСИР параметризуются OWL-описанием схемы данных для автоматической настройки на предметную область конкретного приложения.

В случае, когда необходимо интегрировать в информационную среду уже существующую систему, вдобавок к описанию онтологии предметной области на OWL, необходимо специфицировать объектно-реляционное отображение, позволяющее отобразить структуру онтологии на структуру базы данных, используемой в организации. Помимо возможностей ручной настройки, имеются прототипы средств обратной инженерии баз данных, которые предоставляют организациям возможность быстрого перехода от унаследованной базы данных к интегрированной информационной системе ИСИР.

Как было сказано, основной задачей ядра ИСИР является формирование «объектной базы данных» из исходного источника данных, для поиска и управления объектами в ней используются подмножества объектного языка запросов OQL и интерфейса объектных баз данных ODMG, . Помимо поддержки хранения объектов (хранимых persistent объектов), ядро ИСИР включает дополнительные базовые сервисы, необходимые в реальных системах. Прежде всего, это служба безопасности, обеспечивающая прозрачную проверку прав доступа к данным при их чтении, модификации и пр. Служба поддерживает возможность назначения персональных прав доступа к защищённым объектам, каскадную проверку доступа для зависимых подобъектов. Другая полезная служба ядра – это обеспечение учета изменений и доступа к объектам. Предусмотрены средства категоризации объектов иерархическими классификаторам и тезаурусам, для которых обеспечивается поддержка эффективного исполнения разнообразных запросов к рекурсивным структурам, возможно, с горизонтальными связями, например, в случае реляционных БД большинство

обращений реализуется одним SQL запросом. В ядро встраивается механизм поддержки версий объектов.

Прикладные и системные программные компоненты ИСИР (рис. 4) работают с данными либо в их объектном представлении, используя ODMG и OQL, либо в Semantic Web + XML-представлении, используя RDF/XML и XPath. XML-представление объектных данных позволяет нам применить к ним всю мощь XML-технологий и интегрировать в архитектуру различные разработки, связанные с XML[12]: XSLT и XForms, OWL/XSD Validation, SOAP и Web-сервисы...

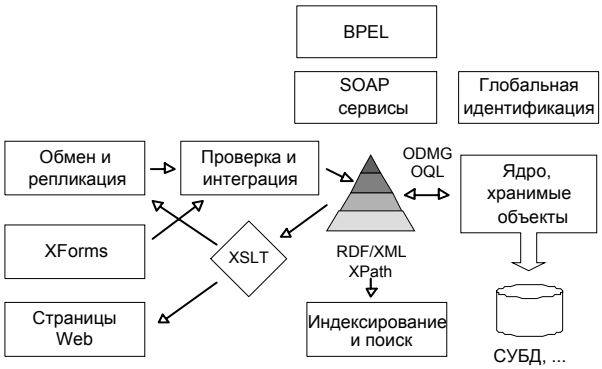

Рисунок 4. Взаимосвязь применяемых технологий.

Функционирование распределённой среды обеспечивается следующими компонентами [5, 18]:

Сервисы обмена и репликации в распределённой среде базируются на RDF/XML-представлении данных и принципах Semantic Web. В отличие от стандартных средств репликации, применяемых для реляционных баз данных, здесь взаимодействующие репозитории могут иметь различную схему данных, и при обмене между ними производится преобразование от одной схемы к другой на основе «отображения» OWL-онтологий, либо более сложного XSLT-преобразования «на лету». Средства интеллектуального импорта данных, используемые при обмене, поддерживают гибкие возможности по настройке алгоритма импорта, модификации имеющихся данных. Производится связывание и интеграция поступающей информации с имеющейся на основе частичного совпадения атрибутики объектов, названия, возможно обнаружение опечаток в данных. При интерактивном импорте возможно взаимодействие с пользователем в случае возникновения конфликтных ситуаций.

Сервис глобальной идентификации, базирующийся на CNRI Handle System, обеспечивает поддержку глобальной идентификации ресурсов репозитория. С его помощью можно найти репозиторий-владелец ресурса, владельцев его копий и реплик, получить доступ к ресурсу через Web и пр.

Сервисы индексирования и поиска обеспечивают построение атрибутно-полнотекстового индекса и поиск по нему для хранимых объектов репозитория, для RDF/XML данных внешних информационных источников, а также текстовых, HTML и XML файлов. При построении индекса и поиске поддерживается нормализация словоформ для русского и английского языков.

SOAP web-сервисы репозитория позволяют унифицировать механизмы доступа к функциям репозитория - импорт/экспорт, манипуляция данными, идентификация, полнотекстовый поиск и пр. Помимо системных web-сервисов, каждый репозиторий может определять собственные специфичные web-сервисы, выполняющие специфичные для этого репозитория задачи, например выдающие отчёты или производящие некоторые вычисления. ИСИР содержит компоненты, максимально упрощающие написания таких web-сервисов, интегрированные со средствами разработки web-страниц xml-отчётов (iXSP).

BPEL электронные регламенты. Мы также разрабатываем средства, обеспечивающие дизайн и исполнение координированных рабочих процессов web-сервисов в рамках языка BPEL. Эти механизмы позволяют создавать составные Web-сервисы на базе имеющихся сервисов среды. Такие механизмы могут использоваться для спецификации регламентов взаимодействия репозиториев, которые позволяют автоматизировать сложные процессы интеграции данных, извлечения данных[16].

Как уже упоминалось, ИСИР-решения позволяют не только построить интегрированную информационную систему, но и предоставить к ней доступ конечным пользователям через Web-портал. Интегрирующий портал представляет собой «единое окно» в электронную среду, открывающее доступ ко всей собранной в ней информации. Помимо этого, заинтересованные организации могут поддерживать собственные корпоративные порталы, предоставляющие доступ к информационному подмножеству, поддерживаемому этой организацией. Собственные порталы организаций могут иметь более специализированную схему данных и расширенный набор сервисов для пользователя. Например, библиотечный портал может предоставлять читателю доступ к списку выданных ему книг, позволять забронировать печатное издание, получить доступ к некоторым электронным изданиям.

ИСИР-технологии специально оптимизированы для создания информационных порталов. Средства построения порталов плотно интегрированы с инфраструктурой ИСИР и ориентируются на применение новейших W3C XML-технологий.

Представление страниц сайта опирается на принципы XML/XSLT-публикации. Для описания web-страницы разработчик сначала указывает логику выборки RDF/XML-информации из

репозитория – для этого используется средство формирования xml-отчётов iXSP («XML серверные страницы»). Вторым шагом является написание или визуальный дизайн XSLT-преобразования, приводящего выбранную из репозитория RDF/XML-информацию к желаемому формату. Одно из преимуществ использования XSLT-публикации состоит в многообразии целевых форматов – это может быть не только XHTML, но и WML для мобильных устройств, PDF для печати и отчётов и пр.

Для построения форм ввода и редактирования данных используется новейший W3C-стандарт XForms, применяющий механизм форм к XML-данным (в нашем случае - RDF/XML). Выборка данных для изначального представления на форме производится с помощью iXSP, а обработка результатов формы интегрирована со средствами интеллектуального импорта данных фреймворка ИСИР.

ИСИР-решения содержат все стандартные базовые блоки, необходимые для быстрой разработки порталов – регистрация и персонализация, портлеты и агрегация контента, новости, форумы, рассылки, система ведения документации и управления контентом [6], административный интерфейс и пр. [7] Эти компоненты реализованы на основе упомянутых выше средств. ИСИР-решения апробированы в ряде применений, как научных[4,8], так и коммерческих.

3 Инфраструктурные службы распределённой среды

3.1 Обмен данными

Как видно из описанного выше, в распределённой среде нам необходима поддержка обмена как минимум двух типов информации между репозиториями: обмен (репликация) метаинформации и обмен индексами для осуществления совместного поиска.

Обмен информацией строится на базовых принципах, предложенных в спецификации протокола Common Indexing Protocol (CIP). Согласно данным предложениям, обмен рассматривается только между парой серверов, более сложные случаи сводятся к серии таких обменов. Для пары серверов определяют два различных отношения: “push” и “poll”. В первом случае (“push”) соединение инициируется сервером, который передаёт информацию другой стороне. Во втором случае (“poll”) инициирующая сторона запрашивает необходимые ей данные у другого сервера.

У инициирующей обмен стороны имеется ряд правил, которые описывают, когда осуществлять взаимодействие и какие данные передавать или запрашивать. Обмен начинается или в результате наступления заданного события (например, изменение определённого количества ресурсов со времени последней связи), или по заданному расписанию. Состав передаваемых объектов определяется запросом, который может накладывать ограничения на их тип, значение свойств, принадлежность заданному репозиторию, а так же выделять ресурсы, изменившиеся после указанного момента времени.

Наиболее простым по затратам методом подключения источника данных к распределённой среде является реализация поддержки отношения типа “push”. Для этого достаточно уметь выгружать данные в формате RDF/XML, и передавать их по протоколу SOAP на заданный адрес, например, один раз в день. Данный подход может быть использован для наиболее удалённых или маломощных узлов сети, а так же в том случае, когда нет возможности установить сервисы ИСИР или совместимые с ними.

Если репозиторий, передающий данные, должен поддерживать отношение “poll”, требования к нему возрастают: необходимо иметь веб-сервис обмена, принимающий запросы на репликацию, а так же уметь выбирать ресурсы согласно приведённым выше критериям. Существующие технологии ИСИР удовлетворяют этим требованиям.

В качестве транспортного протокола используется SOAP, собственно данные передаются как бинарные объекты в соответствии с соглашением “SOAP with Attachments”. В спецификации CIP для этих целей был предложен простой текстовый протокол, но разработчиками было принято решение использовать SOAP благодаря его популярности, а так же существующей поддержке ядром ИСИР.

3.2 Репликация ресурсов

Упомянутый выше сервис глобальной идентификации, помимо того, что гарантирует уникальность генерируемых идентификаторов, так же позволяет централизованно хранить дополнительную информацию, ассоциированную с идентификатором. В частности, вместе с идентификатором хранится информация о репозитории, которому принадлежит данный ресурс.

Метаданные ресурса могут быть изменены только в рамках репозитория, которому относится ресурс. Если информация была реплицирована на другие сервера, то там она рассматривается как информация “только для чтения”. В частности, это ограничение не позволяет редактировать одни и те же метаданные в двух разных местах.

Таким образом, из всех доступных копий метаданных в распределённой среде только одна считается актуальной, а остальные потенциально могут содержать устаревшую информацию. Для поддержания копий в актуальном состоянии используется процесс репликации. Репликация использует общий сервис обмена данными, в качестве передаваемой информации выступает RDF/XML представление ресурсов.

3.3 Совместный поиск

Совместный поиск позволяет пользователю искать нужную ему информацию среди всех распределённых по среде ресурсов. При этом неизбежны некоторые компромиссы, из-за которых поиск выполняется не настолько точно, как если бы запрос осуществлялся в рамках одного репозитория. Это связано с тем, что ресурсы хранятся децентрализовано и могут изменяться в произвольный момент времени, а так же с большими размерами сети и различными пропускными способностями каналов. В то же время поиск должен выполняться достаточно эффективно по времени, что не даёт возможности опрашивать все репозитории.

Для решения этой задачи используется методика маршрутизации запросов, основанная на обмене описателей коллекций и индексов.

Индексирование. Для хранимых ресурсов строится атрибутно-полнотекстовый индекс, с помощью которого можно искать как полнотекстовую информацию (например, в текстовых полях описаний), так и выбирать ресурсы по значению их атрибутов. Индекс хранится в реляционной СУБД, и обновляется по заданному расписанию. Кроме непосредственно поиска, индекс так же используется для формирования описателей коллекций, которые основаны на частотных характеристиках встречающихся в индексе термов.

Описатели коллекций предназначены для определения степени релевантности репозитория какому-либо поисковому запросу, и применяются при маршрутизации запросов.

Протокол поиска. Стандартизованный поисковый протокол необходим как для совместного поиска, так и для предоставления возможности поиска внешним системам. Для ИСИР был выбран протокол Simple Digital Library Interoperability Protocol (SDLIP), благодаря его ориентированности на XML, относительной простоте и большим возможностям (в частности, “поиск в найденном”). Протокол SDLIP может работать на основе различных транспортных протоколов. В стандарте есть примеры использования SDLIP поверх HTTP и CORBA. В ИСИР SDLIP используется поверх SOAP, и оформлен в виде веб-сервиса.

С учётом вышесказанного поиск осуществляется следующим образом. У репозитория, на который поступил запрос, имеется индекс своих и реплицированных на него ресурсов, на основании которого определяются соответствующие запросу ресурсы, находящиеся в данном репозитории. Помимо индекса, у репозитория могут быть описатели коллекций или индексы нескольких удалённых репозиториев. При наличии индекса можно сразу отобрать подходящие под запрос удалённые ресурсы, в случае описателей коллекций выбирается подмножество серверов, которые с наибольшей вероятностью имеют необходимую информацию. Далее, на выбранные серверы отправляется SOAP/SDLIP запрос, и ожидается поступление результатов поиска. Следует заметить, что удалённые серверы так же могут распространить этот запрос на другие репозитории в процессе обработки. После поступления ответов результаты агрегируются, удаляются дубликаты, и производится общее ранжирование результатов (на основе информации о релевантности, поступившей от каждого сервера).

Решения о том, в каких случаях использовать реплицирование ресурсов, в каких реплицирование индексов, а в каких – описателей коллекций, принимается проектировщиком среды.

Заключение

В работе рассмотрены вопросы интеграции информационных ресурсов РАН, информационной поддержки научных исследований в рамках Единого Научного Информационного Пространства РАН. Обсуждены возможности технологии ИСИР, ее применения для обеспечения интеграции информационных ресурсов РАН. Сопоставление требований ЕНИИР РАН и возможностей кросс-платформенной технологии ИСИР позволяет сделать вывод о том, что проблемы ЕНИР РАН в существенной степени могут быть поддержаны решениями ИСИР.

Литература

[1] A.N. Bezdushnyi, A.B. Zhizhchenko, M.V. Kulagin, and V.A. Serebryakov, "Integrated Information Resource System of the Russian Academy of Sciences and a Technology for Developing Digital Libraries", Programming and Computer Software, Vol. 26, No. 4, 2000, pp. 177–185

[2] А. А. Бездушный, А.Н. Бездушный, А.К. Нестеренко, В.А. Серебряков, Т.М. Сысоев, "Архитектура RDFS-системы. Практика использования открытых стандартов и технологий Semantic Web в системе ИСИР", Пятая Всероссийская научная конференция: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL’2003, Санкт-Петербург, Россия, 2003. http://rcdl2003.spbu.ru/proceedings/J1.pdf

[3] А. А. Бездушный, А.Н. Бездушный, А.К. Нестеренко, В.А. Серебряков, Т.М. Сысоев, " RDFS как основа среды разработки цифровых библиотек и Web-порталов ", Российский научный электронный журнал “Цифровые библиотеки”, том. 6, выпуск 3, 2003. http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2003/part3/BBNSS

[4] А. А. Бездушный, А.К. Нестеренко, Т.М. Сысоев, А.Н. Бездушный, В.А. Серебряков, " Архитектура и технологии RDFS-среды

разработки цифровых библиотек и Web-порталов ", Российский научный электронный журнал “Цифровые библиотеки”, том. 6, выпуск 4, 2003. http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2003/part4/BNSBS

[5] А.Н. Бездушный, Д. А. Ковалёв, В.А. Серебряков, " Архитектура Сервисов Интегрированной Системы Информационных Ресурсов (ИСИР) ", Российский научный электронный журнал “Цифровые библиотеки”, том. 5, выпуск 1, 2002. http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2002/part1/BKS

[6] Т.М. Сысоев, А.К. Нестеренко, А. А. Бездушный, А.Н. Бездушный, В.А. Серебряков, "О реализации службы управления содержанием ", Российский научный электронный журнал “Цифровые библиотеки”, том. 6, выпуск 6, 2003. http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2003/part6/SNBBS

[7] А.К. Нестеренко, А. А. Бездушный, Т.М. Сысоев, А.Н. Бездушный, В.А. Серебряков, " Служба управления потоками работ по манипулированию ресурсами репозитория ", Российский научный электронный журнал “Цифровые библиотеки”, том. 6, выпуск 5, 2003. http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2003/part5/NBSBS

[8] А. А. Вежневец, А. Н. Бездушный, " Вопросы построения информационного портала поддержки использования результатов фундаментальных исследований ", Российский научный электронный журнал “Цифровые библиотеки”, том. 6, выпуск 6, 2003. http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2003/part6/VB

[9] W3C Semantic Web Activity. http://www.w3.org/2001/sw/

[10] Dublin Core Metadata Initiative. http://dublincore.org/

[11] MIA Development: Architecture and Functional Model, Tracy Gardner, UKOLN, University of Bath

[12] Apache Cocoon Project. http://cocoon.apache.org

[13] European Research Gateways Online http://www.cordis.lu/ergo

[14] Laitinen, Sauli; Sutela Pirjo & Tirronen, Kerttu, Development of Current Research Information Systems in Finland, proceeding of CRIS-2000

[15] А.Н. Бездушный, А.Б. Жижченко, Н.Е. Калёнов, М.В. Кулагин, В.А. Серебряков, Предложения по наборам метаданных для научных информационных ресурсов ЕНИП РАН, Шестая Всероссийская научная конференция: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL’2004, Пущино, Россия, 2004.

[16] Нестеренко А.К., Сысоев Т.М., Бездушный А.А., Бездушный А.Н., Серебряков В.А., Интеграция распределенных данных на основе технологий Semantic Web и рабочих процессов, Шестая Всероссийская научная конференция: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL’2004, Пущино, Россия, 2004.

[17] Соколов И.А., Босов А.Б., Бездушный А.Н. Об Информационном Web-портале Российской Академии Наук // Системы и средства информатики, Выпуск 13, ISSN 0869-6527. М: Наука, 2003, с. 139-155

[18] А. Н. Бездушный, Д. А. Ковалев, В. А. Серебряков. Технологии репликации данных и распределенного поиска в ИСИР РАН. // Сборник докладов Третьей Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции", Петрозаводск, 09.2001

Capabilities of ISIR technology in supporting of United Scientific Informational Space of RAS

Bezdushny A. A., Bezdushny A. N., Nesterenko A. K., Serebriakov V. A., Sysoev T. M.

This paper is devoted to questions of integration of informational resources of RAS and informational support of investigations in United Scientific Informational Space of RAS. We consider needs, goals and tasks of USIS RAS, as environment of interconnected distributed systems. The capabilities of applying ISIR technology for solving this task is discussed. This technology may be used to form infrastructural elements of environment, to create adapters for existing systems and to implement new system. We shortly consider architectural principles of last ISIR’s work outs, applying of Semantic Web technology, multilayer architecture, ISIR developer’s kit and development of informational systems and Web portals based on ISIR.