ПРАКТИКА ПОСТРОЕНИЯ И ЭКСПЛУАТАЦИИ РАСПРЕДЕЛЕННЫХ ИНФОРМАЦИОННЫХ СИСТЕМ НА ОСНОВЕ ПРОТОКОЛА Z39.50

Жижимов О.Л., Мазов Н.А., Скибин С.В.
Объединенный Институт геологии, геофизики и минералогии Сибирского Отделения РАН, пр. Акад. Коптюга 3, 630090 Новосибирск, Россия
E-mails: zhizhim@uiggm.nsc.ru, mazov@uiggm.nsc.ru, skibin@uiggm.nsc.ru

EXPERIENCE IN CONSTRUCTION AND OPERATION OF THE DISTRIBUTED INFORMATION SYSTEMS ON THE BASIS OF THE Z39.50 PROTOCOL

Oleg Zhizhimov, Nikolay Mazov, Sergey Skibin
United Institute of Geology, Geophysics and Mineralogy, Siberian Branch of the Russian Academy of Sciences, Novosibirsk, Russia
E-mails: zhizhim@uiggm.nsc.ru, mazov@uiggm.nsc.ru, skibin@uiggm.nsc.ru

Questions concerned with construction and operation of the distributed information systems on the basis of ANSI/NISO Z39.50 Information Retrieval Protocol are discussed in the paper.

The paper is based on authors' practice in developing ZooPARK server. Architecture of distributed information systems, questions of reliability of such systems, minimization of search time and administration are examined. Problems with developing of distributed information systems are also described.

Неуклонное накопление человечеством знаний и обработка их на компьютерах неизменно порождает ряд существенных проблем, связанных со стандартизацией правил работы с данными. В настоящее время существует множество различных стандартов, рекомендаций, соглашений, необходимых для полноценной работы с огромным количеством накопленной информации, доступной в рамках сети Интернет, разработано множество систем для эффективного поиска информации.

Известные поисковые системы (Google, AltaVista, InfoSeek, Yahoo и др.) являются многоцелевыми, производят поиск по огромным массивам информации. Так, например, система Google индексирует порядка 2 млрд. HTML-страниц. Для обеспечения быстрого поиска по такому объему информации применяются простые вычислительные методы, огромные вычислительные ресурсы (распределенные системы кластеров до 1000 компьютеров в каждом). При этом широко применяются распределенное хранение и обработка данных. Однако универсальные поисковые системы не в

состоянии использовать более сложные алгоритмы поиска, позволяющие учитывать тематику, производить поиск по различным поисковым атрибутам.

Наряду с многоцелевыми поисковыми системами существуют специализированные, или тематические, поисковые системы. Примером таких систем могут быть различные информационно-поисковые системы для обработки библиографической информации, кадровые системы, системы для обработки коллекций (музейных, биологических и др.), системы для ведения и использования различных словарей, тезаурусов, классификационных схем, системы обработки геоинформационной информации и др. Такие системы очень важны в повседневной жизни ученых и специалистов во многом потому, что часто бывает необходима достоверная и проверенная информация, получение которой не всегда могут обеспечить многоцелевые поисковые системы.

Помимо административных и правовых ограничений в получении информации, для ученых и специалистов существуют и чисто технические трудности. Частично решить эти проблемы могут специализированные информационные системы, которые намного лучше приспособлены для конкретной тематики и могут производить более структурированный, пертинентный и вычислительноемкий поиск информации. В таких системах возможно использование полуавтоматизированных схем работы с информацией, в отличие от многоцелевых систем, в которых используются в основном автоматизированные схемы. Профильный подход позволяет вести рецензирование, коррекцию информации, что может существенно повысить качество поиска и в идеале исключить большую часть нерелевантной информации, что не всегда удается при выдаче результатов поиска многоцелевыми поисковыми системами.

Для получения наиболее полной информации, соответствующей конкретному запросу, повышения качества поиска специализированными информационными системами желательно их объединение в одну распределенную информационную систему (РИС). Обычно термин «распределенность» трактуется довольно широко. Здесь и далее мы будем подразумевать под РИС совокупность (объединение) гетерогенных (различных по структуре и форме построения) информационных систем, способных предоставлять схожую по структуре информацию. Ниже приведен список требований, которым должна удовлетворять РИС [1]:

- возможность работы с распределенными данными – информационная система должна допускать возможность работы с данными, расположенными на разных физических серверах, различных аппаратно-программных платформах и хранящихся в разнообразных внутренних форматах, в контексте одного клиентского сетевого соединения;

- логическая группировка данных – система должна позволять обрабатывать все запросы на логических группах баз данных, полностью скрывая тем самым физическое расположение последних;
- абстрактная модель данных – информационная система должна строиться на основе абстрактной схемы данных, на которую должны быть отображены конкретные базы данных. Это позволяет объединять данные из разнородных систем в одной логической группе;
- абстрактная система запросов – система должна оперировать не конкретным синтаксисом запросов, а его логической сутью на основе абстрактных атрибутов;
- метаинформация – система должна предоставлять полную информацию о себе и обо всех своих ресурсах;
- разграничение доступа – система должна быть способна предоставлять различные уровни привилегий для пользователей по доступу к информации;
- учет и контроль – система должна уметь собирать статистические данные по запросам пользователей и вести их бюджеты;
- открытость – система должна допускать расширение и быть основана на открытых стандартах и протоколах;
- связь с другими системами – возможность интегрировать свои ресурсы с ресурсами других информационных систем;
- демократичность в общении – система должна предоставлять как простые и понятные для неподготовленного пользователя интерфейсы, так и профессиональные интерфейсы для доступа к информации;
- связь с WWW – система должна иметь шлюз для доступа к ней из WWW.
- Объединение разнородных ресурсов в одну РИС предполагает, что члены объединения должны решить следующие вопросы [1, 4, 5, 7, 9]:
- договоренности о протоколах (Data Retrieval Protocol);
- договоренности о схемах данных (Meta Data);
- договоренности о наборах поисковых атрибутов (наборы, минимальные обязательные элементы) (Search Attributes);
- договоренности о форматах представления информации (Presentation Formats);
- договоренности о способах объединения результатов поиска (Data Fusion);
- определения схем маршрутизации (Query Routing);
- определения зон ответственности (Responsibility Division);
- обеспечения избыточности при выходе из строя какого-либо из узлов системы и как следствие (Resource Redundancy), договоренности об обмене данными;

- координации по наполнению данных, т. е. определение схем ведения заимствований описаний при дублировании схожей информации, постоянное координированное администрирование подсистем.

Очевидно, что вышеперечисленным требованиям РИС не могут удовлетворять системы, которые базируются на коммерческих СУБД со специфичными базами данных, поскольку каждая из таких систем, как правило, имеет собственные протоколы обмена информацией и схемы данных, несовместимые с другими родственными системами. Также этим требованиям не могут удовлетворять и обычные HTML-ресурсы, в силу отсутствия жестких стандартов на метаданные и способов построения конкретных систем.

Существуют различные решения унификации доступа к реляционным СУБД, основанные как на унификации API, так и на функционировании промежуточных подсистем и серверов (ODBC, JDBC, XML, XQL, MS ADO, CORBA, Applications Servers), а также на использовании разрозненных стандартов, таких как Marc, HTML, ISO-2709 и др. Однако универсальное решение проблемы построения профилированной РИС возможно только при совместной стандартизации сетевого обмена и схем данных.

Именно для этих целей в мировом информационно-библиотечном сообществе эффективно применяется протокол Z39.50 [1, 2], регламентирующий сетевой обмен и абстрактные схемы данных.

Протокол Z39.50 (ISO-23950) является удобной основой для создания профилированных РИС. Главной отличительной особенностью Z39.50 является стандартизация метаданных, схем данных, без чего невозможно построение РИС из разнородных источников информации. Причем Z39.50 является мировым стандартом, которого придерживаются многие информационные организации и объединения во всем мире, например, такие как Библиотека конгресса США, Ассоциация Российских библиотек АРБИКОН (в области электронных библиотек) и др. Использование протокола Z39.50 информационными сообществами как в России, так и за рубежом, показывает жизнеспособность РИС, построенных на основе этого протокола [8].

Как уже говорилось ранее, применимость протокола Z39.50 не ограничивается библиотеками, и существуют, например, проекты по созданию единых информационных систем с использованием Z39.50 в рамках Новосибирского научного центра, в других объединениях различных регионов России, общероссийские проекты. При этом объединяться могут тезаурусы, классификационные схемы, полнотекстовые документы, музейные данные, геоинформационные описания и многие другие виды информации [9–11].

Авторы настоящего доклада являются разработчиками серверного программного обеспечения Z39.50 – «ZooPARK» [3], успешно функционирующего сегодня в РИС г. Новосибирска («Региональная библиотечная

система»), СО РАН («Интегрированная РИС СО РАН»), РИБС г. Москвы («Корпоративная сеть публичных библиотек Москвы»), РИС LibWeb («Распределенный каталог LibWeb») и в других. По различным данным программное обеспечение ZooPARK составляет от 60 до 70% серверного программного обеспечения Z39.50 в России и входит в десятку самых популярных серверов Z39.50 в мире.

Модульный сервер Z39.50 ZooPARK допускает работу с данными различных СУБД и удовлетворяет следующим требованиям:

- поддержка протокола Z39.50-1995 (v.3);
- работа с различными СУБД;
- переносимость на различные аппаратные платформы.

Внутренняя структура сервера такова, что доступ к конкретным базам нных, будь то CDS/ISIS или MySQL, осуществляется через специальные динамически подгружаемые модули – провайдеры данных. При помощи специального провайдера удаленного доступа (Z-Remote) происходит взаимодействие с другими серверами Z39.50. Функциональная схема сервера ZooPARK представлена на рис. 1.

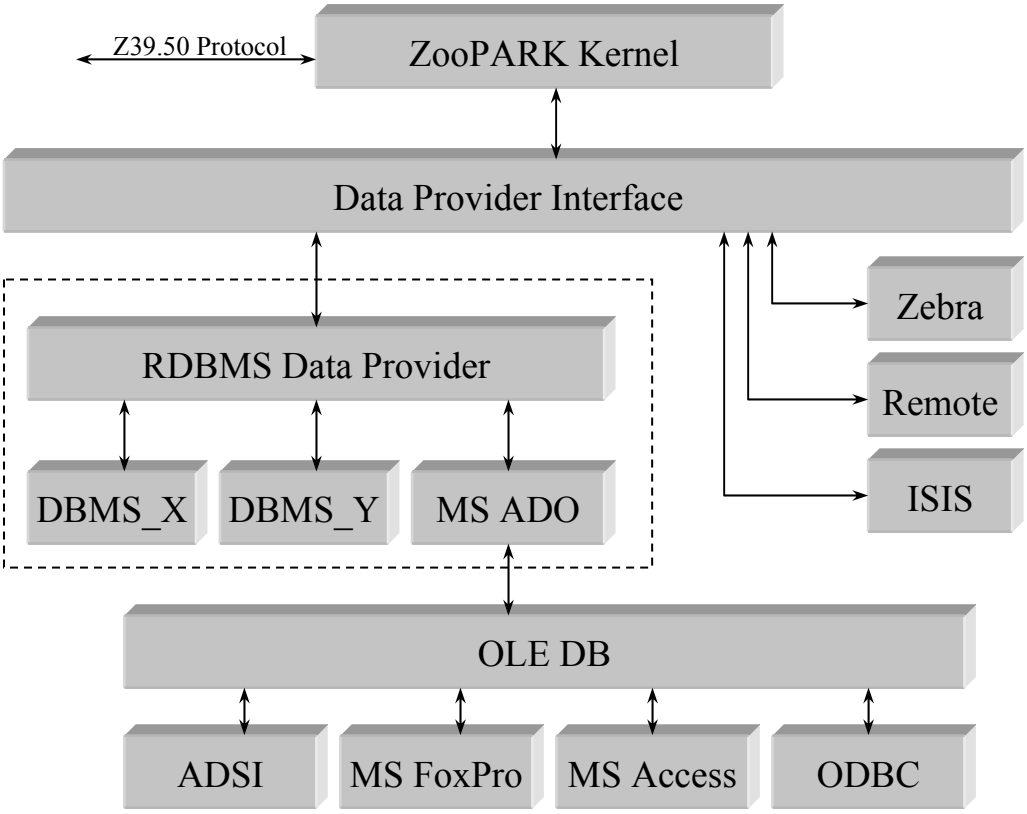

Рис 1. Функциональная схема доступа к данным в сервере ZooPARK

Важными аспектами такого подхода являются расширяемость и независимость тела сервера от типа базы данных. Каждый конкретный про-

вайдер может взаимодействовать с конкретным типом БД, причем он и только он становится жестко привязанным к конкретной СУБД, все остальные модули сервера при этом не зависят от источника данных.

Несмотря на то, что в настоящее время в мире существует довольно большое число разработчиков продуктов с использованием протокола Z39.50, в России по данным проекта RUSLAN на 1 июня 2002 г. существовало всего 4-6 разработок, а реальные системы для широкого применения в информационно-библиотечном сообществе поставляют только 2-3 разработчика (http://www.ruslan.ru:8001/rus/z3950/stat/stat.shtml). Это позволяет намного легче осуществить различные договоренности в рамках использования Z39.50 с учетом Российских особенностей (например, о различных кириллических кодировках). Доля иностранных производителей на российском рынке невелика. Таким образом, системы на основе протокола Z39.50 являются удобной платформой для построения открытых распределенных информационных систем. А в тех случаях, когда в какой-то области не существует стандартизованных схем данных, наборов поисковых атрибутов, разработчики серверов Z39.50 на территории России и специалисты из конкретных областей могут довольно легко выработать отраслевые, региональные стандарты, которых затем будут придерживаться новые члены различных информационных сообществ России. Это возможно в рамках протокола Z39.50 (ISO-23950), который позволяет вносить расширения локальными группами – подкомитетами.

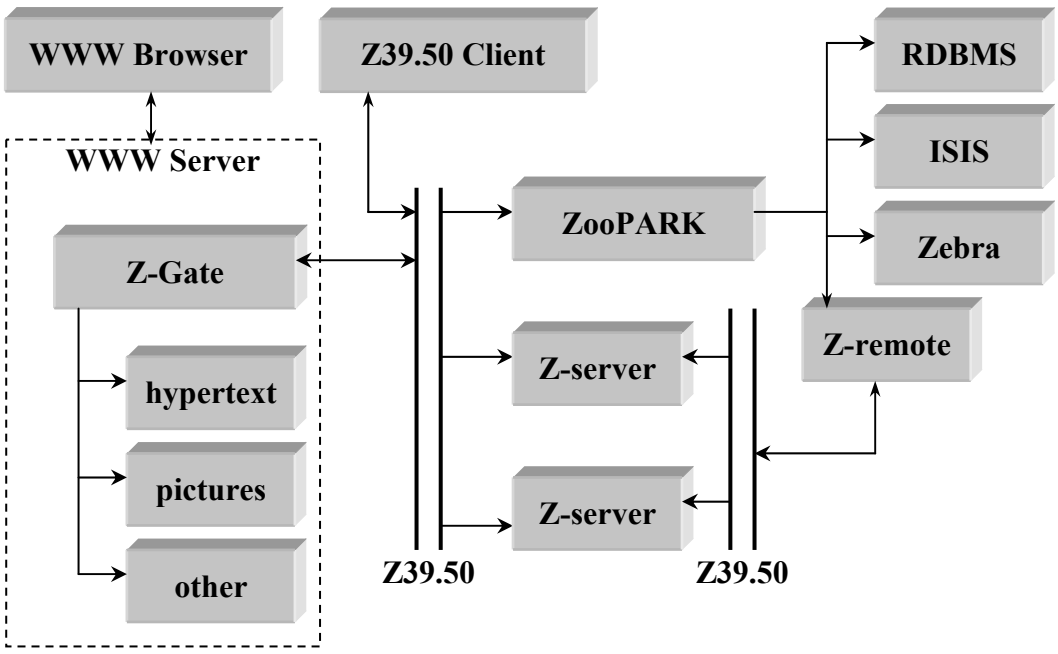

Рис.2 Обобщенная схема РИС с использованием серверов Z39.50

Рассмотрим наиболее часто используемые варианты архитектуры РИС. Обобщенная схема РИС представлена на рис. 2. Она состоит из множества географически распределенных серверов Z39.50, каждый из которых поддерживает определенный набор баз данных. Для доступа к таким РИС применяются клиенты Z39.50, которыми могут быть как специализированные программы, так и WWW-шлюзы.

Наиболее часто применяются следующие схемы РИС:

1. Сам сервер Z39.50 обращается напрямую к различным источникам данных по фирменным протоколам, либо с использованием промежуточных технологий, преобразует полученную информацию к требуемому виду в соответствии с протоколом Z39.50.

2. Клиент Z39.50 (либо WWW-шлюз) обращается к ряду серверов независимо, сам производит обработку полученных данных.

3. Существуют выделенные серверы Z39.50 (в частности, сервер «ZooPARK» на схеме), которые производят распараллеливание запроса, осуществляют маршрутизацию, обработку полученной информации и возвращают клиенту кумулятивный результат.

Первый вариант является неким элементарным звеном РИС, но может и самостоятельно выполнять многие функции, «распределенность» при этом подразумевается в распределенном хранении данных. Так, один сервер ZooPARK может отображать данные из различных типов СУБД, в том числе разнесенным административно и географически. Основным требованием к такой схеме является хорошая пропускная способность каналов связи, так как некоторые СУБД могут создавать большой служебный трафик передачи данных.

Второй вариант подразумевает большие вычислительные затраты программ-клиентов и продуманную логику обработки данных из набора разрозненных серверов. Этот вариант может быть оправдан при построении сводного портала на Web какого-либо конгломерата источников информации. При такой схеме появляется возможность выбора подмножества серверов, в т.ч. использование произвольного отдельного сервера, возможно, не включенного в локальное объединение РИС. Важным моментом при этой схеме является возможность установить принадлежность полученных данных источнику информации. Но существует и ряд проблем, решение которых весьма затруднено, а именно: проблемы с настройкой избыточности, сортировкой полученных данных, маршрутизацией запросов. Решение этих проблем требует наличия высококлассных специалистов при конкретном WWW-шлюзе и создает определенные трудности в поддержке и настройке системы.

В третьем варианте клиенты (WWW-шлюзы) становятся более простыми. Основная мощь вычислений ложится на серверы Z39.50. При такой схеме РИС вся настройка работы производится в рамках серверов системы. Программы-клиенты же просто получают требуемую им информацию, не

зная о распределенности компонентов системы. Однако в этом случае пропадает возможность гибкой настройки конечным пользователем наборов серверов, параметров поиска и пр. К тому же необходимы более тесные контакты между членами РИС. Так, например, необходимо вести согласованное администрирование настроек Z-серверов. Тем не менее, на наш взгляд, такая схема является наиболее предпочтительной. В настоящее время активно ведутся работы по увеличению гибкости данной схемы в рамках сервера ZooPARK. В частности, ведутся следующие работы:

- оценка необходимости применения и построения схем маршрутизации запросов;
- создание избыточных массивов информации (Resource Redundancy);
- упрощение и централизация администрирования сервера ZooPARK;
- повышение простоты, надежности и масштабируемости решений;
- расширение набора поддерживаемых СУБД;
- повышение эффективности различных форматных преобразований данных по протоколу Z39.50.

Каналы связи между серверами имеют различную пропускную способность, а серверы в рамках единой РИС также обычно обладают различными вычислительными мощностями.

Так, в библиотечных корпоративных проектах, как правило, существуют головные организации – центральные библиотеки, которые могут себе позволить иметь лучшие серверы и более высококвалифицированный технический персонал, нежели малые библиотеки, входящие в сообщество.

Исходя из вышеперечисленного, а также ряда других факторов возникает вопрос о маршрутизации запросов [5], составлении сетей серверов со взвешенными связями, оценке необходимой избыточности. Это позволит при выходе из строя какого-либо сегмента РИС сохранить полную функциональность системы. Избыточность требует построения схем обмена информацией. Пока такой процесс не может быть полностью автоматизирован, поскольку члены РИС обладают своими уникальными наработками вне протокола Z39.50 (СУБД, схемы данных и др.) и часто не хотят их менять. С другой стороны, обмен информацией в рамках ограниченных пропускных возможностей сетей при больших объемах данных не позволяет в автоматизированном режиме дублировать информацию по сетям, используя серверы Z39.50.

В настоящее время авторами изучается эта проблема. Предполагается, что в ближайшее время решение будет найдено.

Но это лишь одно из направлений будущей деятельности, а пока в РИС возникают неотложные проблемы, которые необходимо решать в ближайшее время. К таким проблемам можно отнести следующие:

- некорректное администрирование, настройка параметров Z-серверов и СУБД;
- неправильное заполнение схем данных;
- маленький процент грамотных специалистов в области Z39.50;
- несоответствие элементов данных в записях различных членов одной РИС между собой при схожести записей (например, данные по одинаковым книгам в электронных картотеках);
- отсутствие скоординированности по использованию наборов поисковых атрибутов.

Так, например, решению подобных проблем могут способствовать требования, предъявляемые обществом «Открытые информационные системы» к третьему этапу конкурса в области ЭБ, рекомендациями других библиотечных ассоциаций. Также помогают схемы заимствования записей из эталонных источников, таких как Всероссийская книжная палата, Российская национальная библиотека, Российская государственная библиотека; постепенное понимание проблем участниками проектов, их нарастающее стремление к сотрудничеству.

В заключение хочется выразить надежду, что в ближайшее время удастся преодолеть ряд административных и технических трудностей при построении РИС, характерных для «подросткового периода». Это, в свою очередь, позволит создать полнофункциональные профилированные РИС, которые помогут ученым и специалистам более эффективно сконцентрироваться на работе с информацией, не отвлекаясь на технический процесс поиска и отбраковки нерелевантных потоков данных, получаемых из многоцелевых поисковых систем.

Литература

[1] *Жижимов О.Л.* Введение в Z39.50. – Новосибирск: Изд-во НГОНБ, 3-е изд., доп. и перераб. 2002.

[2] ANSI/NISO Z39.50-1995. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. Z39.50 Maintenance Agency Official Text for Z39.50-1995, July 1995.

[3] *Жижимов О.Л., Мазов Н.А., Скибин С.В.* Текущее состояние программного обеспечения Z39.50 ОИГГМ СО РАН (ZooPARK). // Матер. 9 Междунар. конф. «Крым-2002»: Москва, Издательство ГПНТБ России, 2002. – Т. 2. – С. 542–544.

[4] Минимальные требования к Z39.50 службе для участников III этапа конкурса «Российские корпоративные библиотечные системы» // Проект RUSLAN (http://www.ruslan.ru:8001/rus/z3950/minreq3.html).

[5] *Некретьстянов И.* Маршрутизация запросов в системах распределенного поиска // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: 2 Всерос. науч. конф.: – Сб. докл. – Протвино, 2000. – С. 280–287.

[6] *Вольфенгаген В.Э., Калиниченко Л.А., Мендкович А.С. и др.* Информационные системы и научные телекоммуникации. (Проблематика и разработки по проектам РФФИ) // Вестник РФФИ. – 1998. – № 4. – С. 4–50.

[7] *Калиниченко Л.А.* Методы и средства интеграции неоднородных баз данных. – М.: Наука, 1983.

[8] *Жижимов О.Л., Мазов Н.А.* Состояние и перспективы использования протокола Z39.50 в информационном сообществе России // Информационное общество. – 2000. – № 2. – С. 39–43.

[9] *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 Всерос. науч. конф.: Сб. докл. – Протвино, 2000. – С. 102–106.

[10] *Мазов Н.А., Жижимов О.Л.* Унификация построения и организации доступа к тезаурусам и классификационным схемам в распределенных информационных системах по протоколу Z39.50 // Электронные библиотеки: перспективные методы и технологии, электронные коллекции: 2 Всерос. науч. конф.: Сб. докл. – Протвино, 2000. – С. 230–233.

[11] *Мазов Н.А., Жижимов О.Л.* Применение протокола Z39.50 в распределенной информационной системе Сибирского отделения РАН // Библиотечно-информационные ресурсы в науке, образовании, культуре и бизнесе: Материалы междунар. конф. – Самарканд, 1999. – С. 118–125.