Технологическая платформа массовой интеграции разнородных данных*

Жижимов О. Л., Федотов А. М., Шокин Ю. И.
Институт вычислительных технологий СО РАН
zhizhim@mail.ru, fedotov@sbras.ru, shokin@ict.nsc.ru

Послушайте, ребята,
Что вам расскажет дед.
Земля наша богата,
Порядка в ней лишь нет.

А. К. Толстой

Доклад посвящен описанию технологической платформы массовой интеграции распределенных разнородных источников данных, поддерживающей создание и функционирование широкомасштабных информационных инфраструктур на основе подхода виртуальной интеграции данных. Платформа массовой интеграции позволит создавать глобальные инфраструктуры из десятков и сотен гетерогенных баз данных и предназначена для решения стратегических задач в области автоматизации различных форм распределенной деятельности по управлению, контролю, планированию на уровне крупных предприятий, отраслей, корпораций и государственных институтов. В основе технологической платформы находится программный комплекс с условным названием ZooSPACE, разрабатываемый в ИВТ СО РАН.

Ключевые слова: распределенные информационные системы, интеграция гетерогенных данных, управление доступом к информационным ресурсам, Z39.50, LDAP, SRW/SRU.

1. Введение

Одним из основных результатов созидательной, социальной и интеллектуальной человеческой деятельности является создание и накопление информационных ресурсов с целью их дальнейшего использования и недопущения утраты опыта предыдущих поколений. Не будет преувеличением сказать, что уровень развития технологий накопления информации и эффективности использования накопленной ранее информации на протяжении всей истории человечества значительно влиял на уровень развития производительных сил. Утеря информации приводила к отбрасыванию цивилизации на века назад. Однако, чтобы эффективно пользоваться накопленной ранее информацией, необходимы специальные инструменты и специальные технологии, при помощи которых могут быть реализованы специальные приемы работы с информацией [1].

*Работа выполнена при частичной поддержке РФФИ: проекты 12-07-00472, 11-07-00561, президентской программы «Ведущие научные школы РФ» и интеграционных проектов СО РАН.

Важнейшей задачей, связанной с технологией работы с информацией, является исследование способов интеграции распределенных источников данных и создание научного задела в области распределенных информационных систем и баз данных в целях разработки инструментальной платформы (далее — платформа массовой интеграции), поддерживающей создание и функционирование широкомасштабных информационных инфраструктур на основе подхода виртуальной интеграции баз данных. Платформа массовой интеграции позволит создавать глобальные инфраструктуры из десятков и сотен гетерогенных баз данных и предназначена для решения стратегических задач в области автоматизации различных форм распределенной деятельности. Более узкой целью работы является разработка принципов и программных средств виртуальной интеграции распределенных источников данных на основе международных стандартов и рекомендаций для создания масштабных информационных инфраструктур, предназначенных для виртуализации доступа к данным различных СУБД с использованием единых правил и политик.

Под интеграцией информационных ресурсов понимается их объединение с целью использования (с помощью удобных и унифицированных пользовательских интерфейсов) разнородной информации с сохранением ее свойств, особенностей представления и пользовательских возможностей манипулирования с ней. При этом объединение ресурсов не обязательно должно осуществляться физически, оно может быть виртуальным, главное — оно должно обеспечивать пользователю восприятие доступной информации как единого информационного пространства. В частности, такие системы обеспечивают работу с гетерогенными наборами и базами данных или системами баз данных, обеспечивая пользователю эффективность информационных поисков независимо от особенностей конкретных систем хранения ресурсов, к которым осуществляется доступ.

Исходя из общей и частной целей, с учетом анализа литературных источников и многолетней практики авторов в области создания программных комплексов для организации доступа к гетерогенным информационным ресурсам и базам данных [2, 3, 4, 5, 6], наиболее оптимальной архитектурой платформы массовой интеграции баз данных представляется архитектура слабосвязанных самодостаточных узлов некой распределенной информационной системы. Здесь и ниже эта система будет идентифицироваться под кодовым названием ZooSPACE. Этимология этого названия основана на двух элементах. Элемент «SPACE» подчеркивает распределенность системы, которая создает некое пространство, в котором могут функционировать информационные узлы и сервисы, обеспечивая самосогласованный доступ к информационным ресурсам и базам данных. Элемент «Zoo» подчеркивает некоторую преемственность предлагаемых решений по отношению к разработанным коллективом исполнителей ранее программных комплексов в области обеспечения унифицированного доступа к гетерогенным базам данных. В первую очередь имеется в виду программный комплекс ZooPARK, разные версии которого успешно эксплуатируются в России и в ближнем зарубежье на протяжении последних 13 лет [7].

Следует заметить, что проблема интеграции данных, как реальной, так и виртуальной, находящихся под управлением различных СУБД, изучается в мире уже давно. В этом направлении разработаны и успешно реализованы многие модели и технологии. Еще в 80-х годах прошлого века был разработан и документирован стандарт ANSI Z39.50 (Information Retrieval (Z39.50): Application Service Definition and Protocol Specification), последняя ревизия которого вышла в 2003 году [8]. Позднее ANSI стандарт был утвержден как стандарт ISO-23950. Спецификации этого стандарта включают

описание механизмов, структур и процедур доступа к базам данным безотносительно к их физической и логической реализации. Позднее идеология Z39.50 была перенесена на идеологию WEB-сервисов и архитектуру SOA. Это привело к созданию протокола SOAP/SRW и SRU, которые упрощали разработку конечных приложений, т.к. использовали технологии HTTP/XML (вместо ASN.1/BER), сохраняя общие принципы Z39.50 по абстрагированию от структур конечных СУБД и предоставляли универсальный способ доступа к данным для поиска и извлечения информации. Именно эти технологии сегодня используются во всем мире для интеграции данных из различных СУБД при построении действительно универсальных систем. На сегодняшний день в мире не существует технологии отличной от технологии Z39.50 и SRW/SRU, которые бы, с одной стороны, обладали требуемым потенциалом для интеграции данных различных СУБД, и, с другой стороны, обладали бы серьезной базой промышленной эксплуатации реальных информационных систем.

2. Платформа массовой интеграции

Для решения сформулированных проблем необходимо создание развитой инфраструктуры для представления и обмена метаданными (данными о ресурсах), без которой невозможно формирование единого информационного пространства [9]. Это можно рассматривать как первый шаг к интеграции и интероперабельности информационных систем. Под интероперабельностью любой информационной системы, в том числе и электронной библиотеки, понимается степень ее способности взаимодействовать с другими информационными системами, в том числе и с человеком. Но если при взаимодействии с человеком (как с информационной системой) основная нагрузка на обеспечение взаимопонимания ложится на человека, который в состоянии обработать даже очень плохо организованную информацию, то для обеспечения эффективного взаимодействия между собственно информационными системами требуются специальные технологические методы и общие соглашения [10].

В основе интеграции распределенных информационных систем лежит интеграция метаданных, которая основана на стандартах на формат для представления метаданных, одновременно с унификацией нормативно-справочной информации (профиля информационных систем) [11]. Под интеграцией данных с точки зрения пользователя следует понимать:

- возможность свободно группировать любые имеющиеся разнородные данные по любому признаку в произвольные реальные и/или виртуальные коллекции;
- возможность организовывать по всем массивам данных прозрачный для конечного потребителя сквозной поиск информации.

Реализация механизмов интеграции данных немыслима без их стандартизации — данные одного типа должны описываться и предоставляться единым образом в соответствии с нормативными документами. В частности, в стандартизованном виде должны предоставляться следующие типы информационных ресурсов:

- географические информационные ресурсы (картографические материалы, спутниковые снимки, данные полевых наблюдений и т. п.), а также соответствующие базы метаданных;
- фактографические базы данных и метаданных;

– библиографические базы данных и электронные каталоги;
– полнотекстовые базы данных и электронные библиотеки;
– авторитетные базы данных (описывающие субъекты информационного взаимодействия: персоны, организации и т. п.);
– другие ресурсы (аудио- и видеозаписи, электронные презентации и др.), снабженные стандартизованными метаданными.

Исходя из вышеперечисленных особенностей, платформа интеграции должна содержать следующие функциональные компоненты [10]:

– систему идентификации информационных ресурсов;
– систему идентификации, аутентификации и авторизации пользователей;
– систему управления метаданными;
– систему управления информационными ресурсами, в том числе полнотекстовыми;
– систему сбора статистики;
– систему мониторингу доступности сервисов и ресурсов.

Реализация этих подсистем должна основываться на открытых спецификациях, связанных с международными стандартами. В распределенной среде должны быть задействованы механизмы синхронизации данных, например, на основе репликаций. При этом в качестве протоколов сетевого взаимодействия должны выступать стандартные протоколы, например, OAI [12], Z39.50 [4], SRW/SRU [2], LDAP [13] и др. (см. рис. 1).

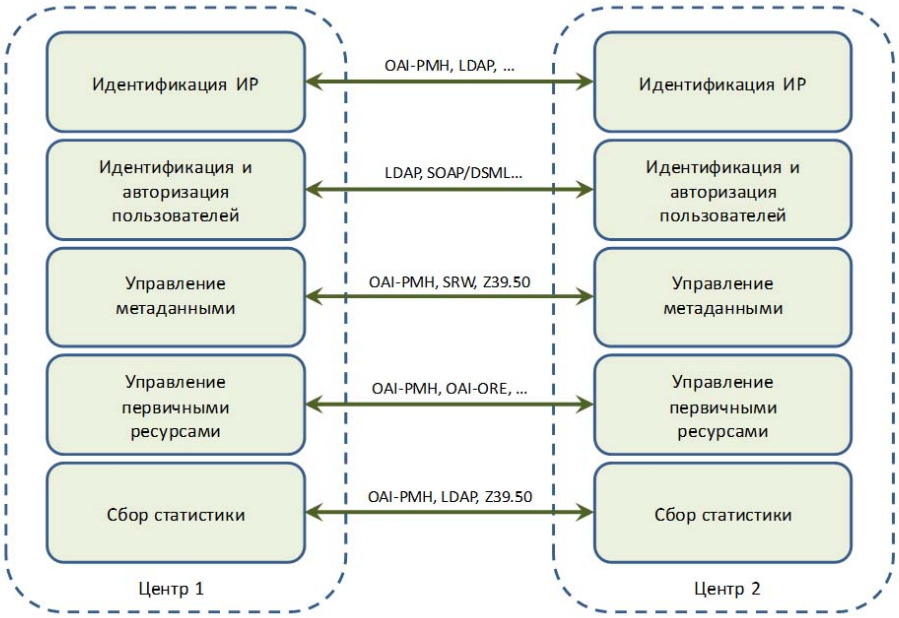

Рис. 1. Сетевое взаимодействие подсистем

3. Общая инфраструктура ZooSPACE

Инфраструктура ZooSPACE реализуется на произвольном количестве слабосвязанных самодостаточных узлов, функционирующих в соответствии с единой политикой. Взаимодействие узлов между собой осуществляется посредством сетевых протоколов прикладного уровня на основе транспортного протокола TCP/IP в соответствии со схемой (см. рис. 2). Количество узлов в ZooSPACE не нормируется и может быть любым. Система ZooSPACE может состоять из одного единственного узла.

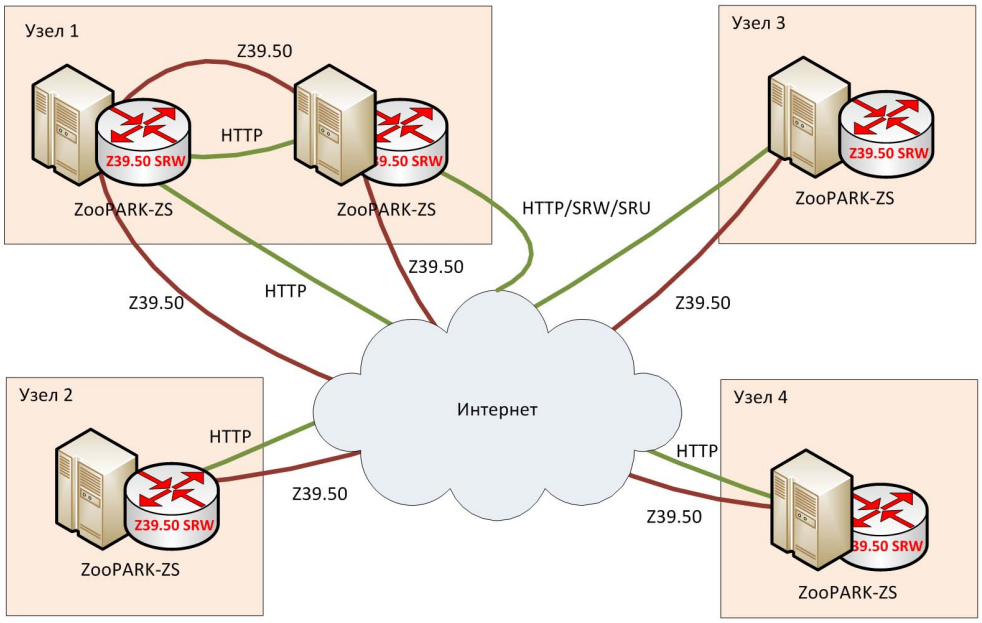

Рис. 2. Инфраструктура узлов ZooSPACE

Создаваемая платформа массовой интеграции предназначена для создания и поддержки функционирования масштабных, динамически формирующихся информационных инфраструктур из большого числа автономных баз данных. ZooSPACE должна обеспечивать функциональные характеристики:

- поддержку унифицированного по информационной инфраструктуре представления данных, которое позволяет выполнять поисковые запросы, не зависящие от физического расположения данных;
- предоставление прикладных программных интерфейсов для выполнения массовых поисковых запросов и управления информационной инфраструктурой;
- обработку массовых запросов к совокупности баз данных реляционного и иерархического типов;
- выбор поискового пространства запроса по метаданным, описывающим характеристики баз данных информационной инфраструктуры;
- синтаксический контроль запроса с соответствующей диагностикой до начала его выполнения;
- подключение/отключение баз данных и вычислительных ресурсов по инициативе их администраторов в процессе функционирования инфраструктуры;

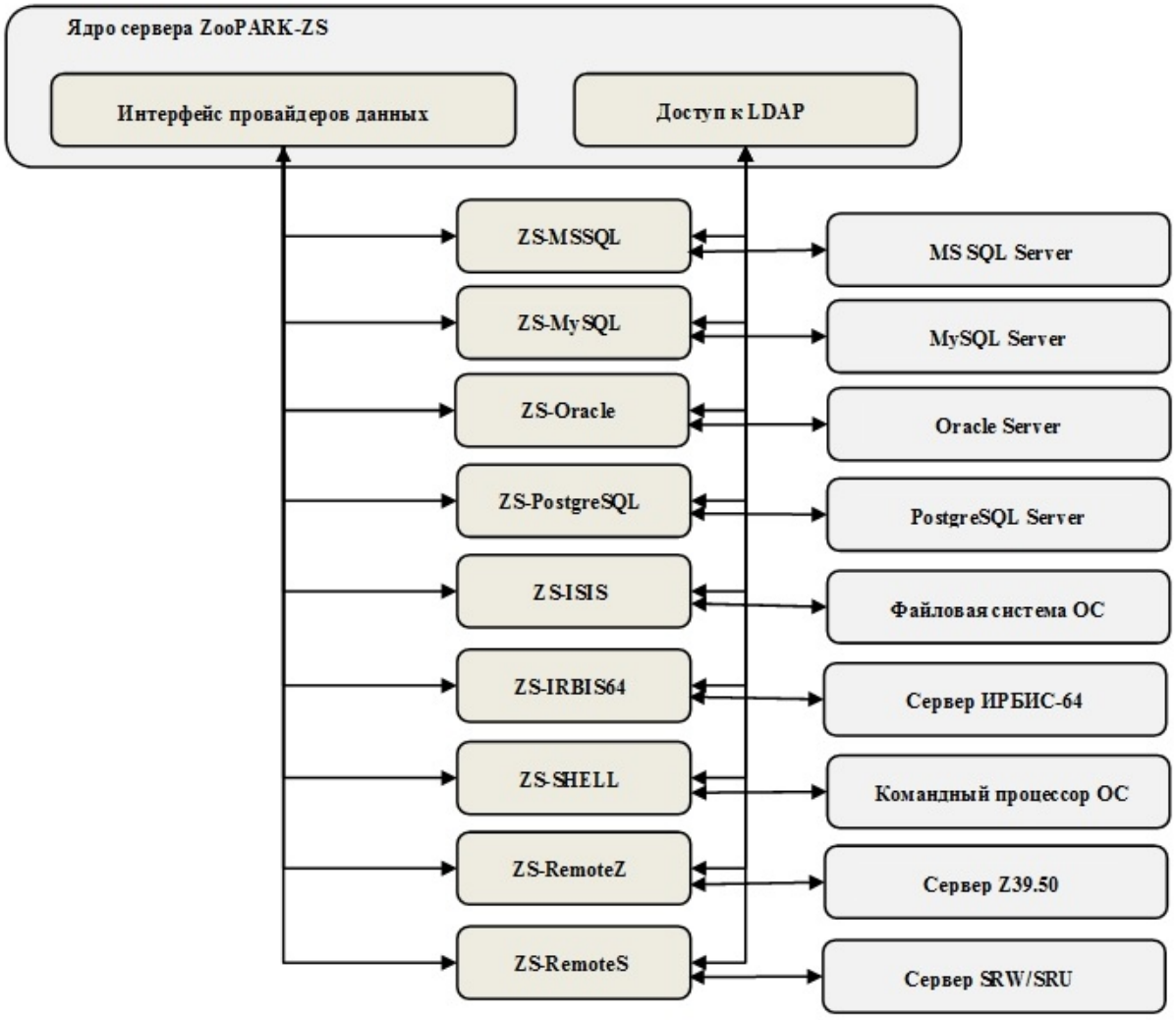

Рис. 3. Доступ к данным сервера ZooPARK-ZS

– защиту хранимых в информационной инфраструктуре данных от несанкционированного доступа.

Как и для базового сервера ZooPARK, доступ к базам данных для сервера ZooPARK-ZS реализован через единый для всех типов поддерживаемых СУБД интерфейс (интерфейс провайдера данных). При этом вся логика взаимодействия с конкретной СУБД локализована в специальном модуле – провайдере данных, который представляет собой динамически загружаемую библиотеку, причем загрузка этой библиотеки происходит на этапе выполнения по мере обращения. Такой режим загрузки модулей не требует перезагрузки сервера при изменении модулей. Для каждого типа СУБД создан свой провайдер данных.

Реализация такого подхода позволяет удовлетворить требованиям для провайдера данных:

– возможность взаимодействия с конечной СУБД в соответствии с ее спецификаци-

ями;

- обеспечение конвертации внешних запросов в синтаксис и семантику целевой СУБД;
- обеспечение конвертации извлекаемой из целевой СУБД информации во внешние структуры данных;
- реализация для различных ОС (Linux,Windows 2003/2008, Solaris x86, FreeBSD).

Схематично доступ к базам данных сервера ZooPARK-ZS представлен на диаграмме в соответствии (см. рис. 3).

Выбор инфраструктуры узлов позволяет обеспечить достаточно гибкую распределенную информационную систему и реализовать всю необходимую функциональность, которая обеспечивается подсистемами ZooSPACE. В качестве подсистем ZooSPACE выступают следующие:

- ZooSPACE-L — обеспечение функционирования справочной и административной подсистемы ZooSPACE.
- ZooSPACE-Z — обеспечение функционирования подсистемы доступа к базам данных системы ZooSPACE.
- ZooSPACE-M — обеспечение функционирования системы мониторинга всех компонент ZooSPACE.
- ZooSPACE-S — обеспечение функционирования подсистемы сбора статистики работы всех компонент ZooSPACE.
- ZooSPACE-W — обеспечение реализации пользовательских и административных WEB-интерфейсов доступа к другим подсистемам ZooSPACE.

В заключении отметим, что разрабатываемый в ИВТ СО РАН подход к построению распределенных информационных систем позволяет обеспечить возможность интеграции разнородных и разнотипных информационных ресурсов в единую информационную среду и унифицированного поиска благодаря использованию унифицированной модели работы с данными (в идеологии протокола Z39.50). Созданная система сервисов предоставляет широкому кругу потенциальных пользователей стандартизированный доступ к данным и алгоритмам их обработки. Такой подход позволяет обеспечить высокую степень информационной поддержки междисциплинарных научных исследований.

Список литературы

[1] *Шокин Ю. И., Федотов А. М., Барахнин В. Б.* Проблемы поиска информации. Новосибирск: Наука, 2010. 198 с.

[2] *Жижимов О. Л., Пестунов И. А., Федотов А. М.* Структура сервисов управления метаданными для разнородных информационных систем [Электронный ресурс] // Электронные библиотеки: российский научный электронный журнал. — 2012. — Москва: Институт развития информационного общества. — Т. 15. — № 6. — ISSN 1562-5419.

[3] *Жижимов О. Л., Амельченко С. А.* Информационная система проекта «Электронная Сибирь»: сервисы управления данными // Вестник ДВО РАН. — 2012. — № 2. — с. 123–128. — ISSN 0869-7698.

[4] *Жижимов О. Л., Мазов Н. А.* Принципы построения распределенных информационных систем на основе протокола Z39.50. — ОИГГМ СО РАН, Новосибирск: ИВТ СО РАН. — 2004. — ISBN 5-9554-0017-6. — 361 с.

[5] *Федотов А. М.* Методологии построения распределенных систем // Вычислительные технологии. — 2006. — Т. 11. — С. 3–17.

[6] *Шокин Ю. И., Федотов А. М., Жижимов О. Л.* Технология распределенных информационных систем // Материалы конференции «Современные информационные технологии для научных исследований». Магадан, 2008. — С. 18–21.

[7] *Жижимов О. Л., Мазов Н. А.* Серверный комплекс ZooPARK — итог 10-летней эксплуатации [Электронный ресурс] // XVI Международная конференция «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса» — Крым-2009 (Судак, Украина, 08.06 – 12.06.2009): Материалы конференции. — М.: ГПНТБ России, 2009. — ISBN 978-5-85638-132-9. — Гос. регистр. № 0320900806.

[8] ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification. - NISO Press, Bethesda, Maryland, U.S.A. — Nov 2002. — ISSN: 1041-5653. — ISBN: 1-880124-55-6.

[9] *Шокин Ю. И., Федотов А. М.* К вопросу о развитии информационной инфраструктуры СО РАН // Вычислительные технологии. — 2009. — т. 6, № 6. — с. 127–137.

[10] *Жижимов О. Л., Федотов А. М., Шокин Ю. И.* Технологическая платформа массовой интеграции гетерогенных данных // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2013. т. 11, вып. 1. с. 24–41.

[11] *Федотов А. М., Барахнин В. Б., Жижимов О. Л., Федотова О. А.* Технология создания корпоративных информационных систем учета трудов научных работников // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2011. т. 9, вып. 2. с. 31–41.

[12] *Жижимов О. Л., Федотов А. М., Федотова О. А.* Построение типовой модели информационной системы для работы с документами по научному наследию // Вестн. Новосиб. гос. ун-та. Серия: Информационные технологии. 2012. т. 10. — № 3. — С. 5–14.

[13] *Федотов А. М., Шокин Ю. И., Жижимов О. Л., Молородов Ю. И.* Служба директорий LDAP как единая информационная среда // Открытое и дистанционное образование. — 2007. — Томск. — № 4(28). — С. 31–41.