ТЕХНОЛОГИЯ ИНТЕГРАЦИИ РАЗНОРОДНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ*

Жижимов О. Л., Федотов А. М., Шокин Ю. И.
Институт вычислительных технологий СО РАН
E-mail: 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

Аннотация: Доклад посвящен описанию технологии интеграции разнородных информационных ресурсов на основе *технологической платформы массовой интеграции распределенных источников гетерогенных данных.* Данная технология поддерживает 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 интеграции информационных ресурсов.

Технологическая платформа создана на основе программного комплекса с условным названием ZooSPACE, разработанного в ИВТ СО РАН.

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, необходима разработка специальных технологий, при помощи которых может быть реализована работа с информацией [1].

Стремительное развитие глобальных информационных и вычислительных сетей ведет к изменению фундаментальных парадигм обработки данных, которые можно охарактеризовать как переход к поддержке и развитию распределенных информационных ресурсов [2, 3]. Технологии использования распределенных информационных ресурсов получают все больший приоритет в информационном обществе. При этом наблюдаются переход к исключительно распределенной схеме создания, поддержания, хранения ресурсов и, в то же время — стремление к виртуальному единству посредством предоставления свободного доступа к любым ресурсам сети через ограниченное число «точек доступа». Постулируется принцип формирования математически однородного поля компьютерной информации, которое способно стать универсальным и машинонезависимым носителем данных.

На самом деле идея создания универсальной системы доступа к информационным ресурсам, распределенным в мировом пространстве, далеко не нова. По всей видимости, впервые ее четко осознал бель-

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

гийский социолог Поль Отле в конце XIX века, предложив совершенно новый метод, названным им *«Документацией»*:

«Цели Документации состоят в том, чтобы суметь предложить документированные ответы на запросы по любому предмету в любой области знания: 1) универсальные по содержанию; 2) точные и истинные; 3) полные; 4) оперативные; 5) отражающие последние данные; 6) доступные; 7) заранее собранные и готовые к передаче; 8) предоставленные как можно большему числу людей» (см. [4], с. 190).

«. . . человеческое знание позволит создать оборудование, действующее на расстоянии, в котором соединятся радио, рентгеновские лучи, кинематограф и микроскопическая фотография. Все предметы Вселенной, все предметы, созданные Человеком, будут регистрироваться на расстоянии с момента их создания. Тем самым будет создан движущийся образ мира — его память, его подлинная копия. Любой человек сможет прочесть отрывок, спроецированный на его личный экран» (см. [4] с. 16).

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 [3, 5, 6, 7, 8], 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 [9].

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 [10]. 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.

1. 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 [11]. 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 [12].

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) [13]. 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.

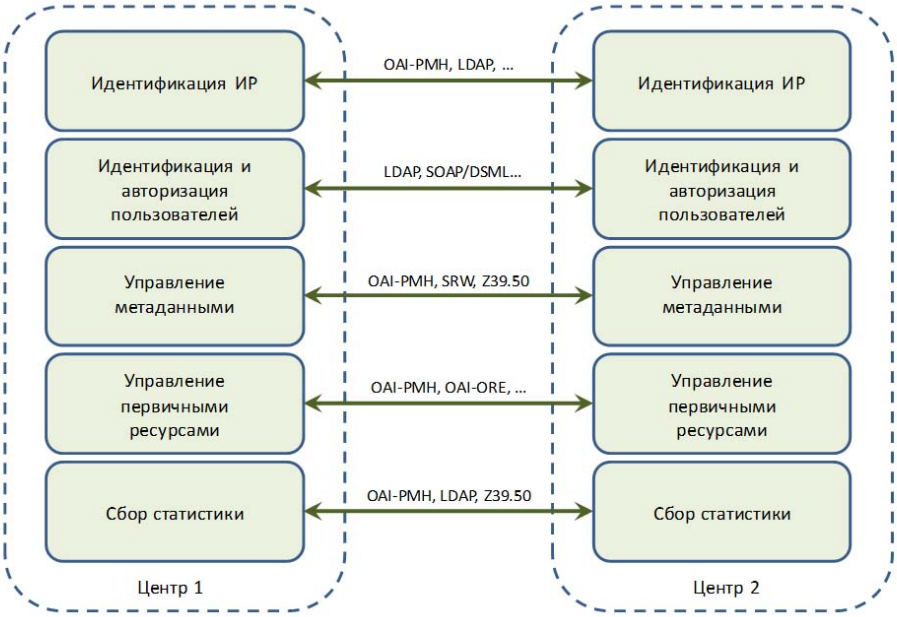

Рис. 1: Сетевое взаимодействие подсистем

Исходя из вышеперечисленных особенностей, платформа интеграции должна содержать следующие функциональные компоненты [12]:

- систему идентификации информационных ресурсов;
- систему идентификации, аутентификации и авторизации пользователей;
- систему контроля доступа к данным и метаданным;
- 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 [14], Z39.50 [7], SRW/SRU [5], LDAP [15] и др. (см. рис. 1).

Реализация сервисов SRW/SRU даст существенно новое качество информационной системы — возможность включения ее ресурсов в глобальные поисковые системы на более высоком уровне, нежели уровень внешней индексации статических WEB страниц другими системами. Другие возможные типы поиска связаны с поиском по заданным шаблонам и с поиском с привлечением онтологии. Поиск с привлечением онтологии является более интеллектуальным типом поиска. Для его реализации требуется дополнительная информации информация о предметной области, включающая определения терминов, сущностей и связей. Следует отметить, что представление этой дополнительной информации должно соответствовать глобальным договоренностям международным стандартом, иначе, поиск с привлечением словарей, тезаурусов и онтологии всегда будет ограничен текущей системой, а интероперабельность не будет реализована.

Отметим, что основу разработки технологии составляют, прежде всего, стандарты и международные рекомендации, формирующие профиль системы, под которым понимается набор из одного или нескольких базовых нормативно-технических документов (стандартов и спецификаций), ориентированных на решение определенной задачи (реализацию заданной функции либо группы функций приложения или среды) с указанием, при необходимости, выбранных классов, подмножеств, опций базовых стандартов, которые являются необходимыми для выполнения конкретной функции.

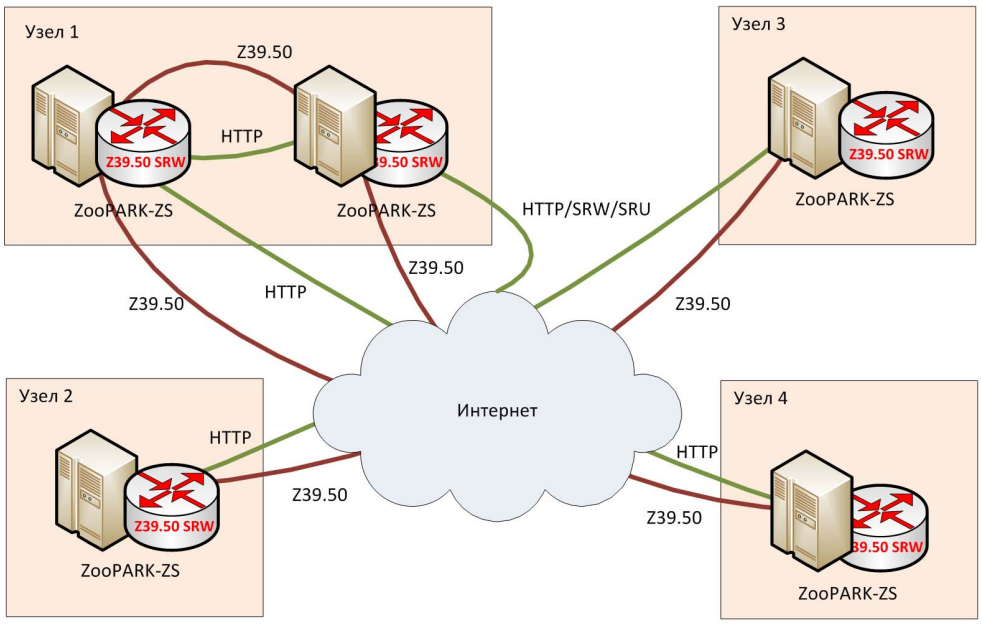

Рис. 2: Инфраструктура узлов ZooSPACE

2. 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.

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;
- 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).

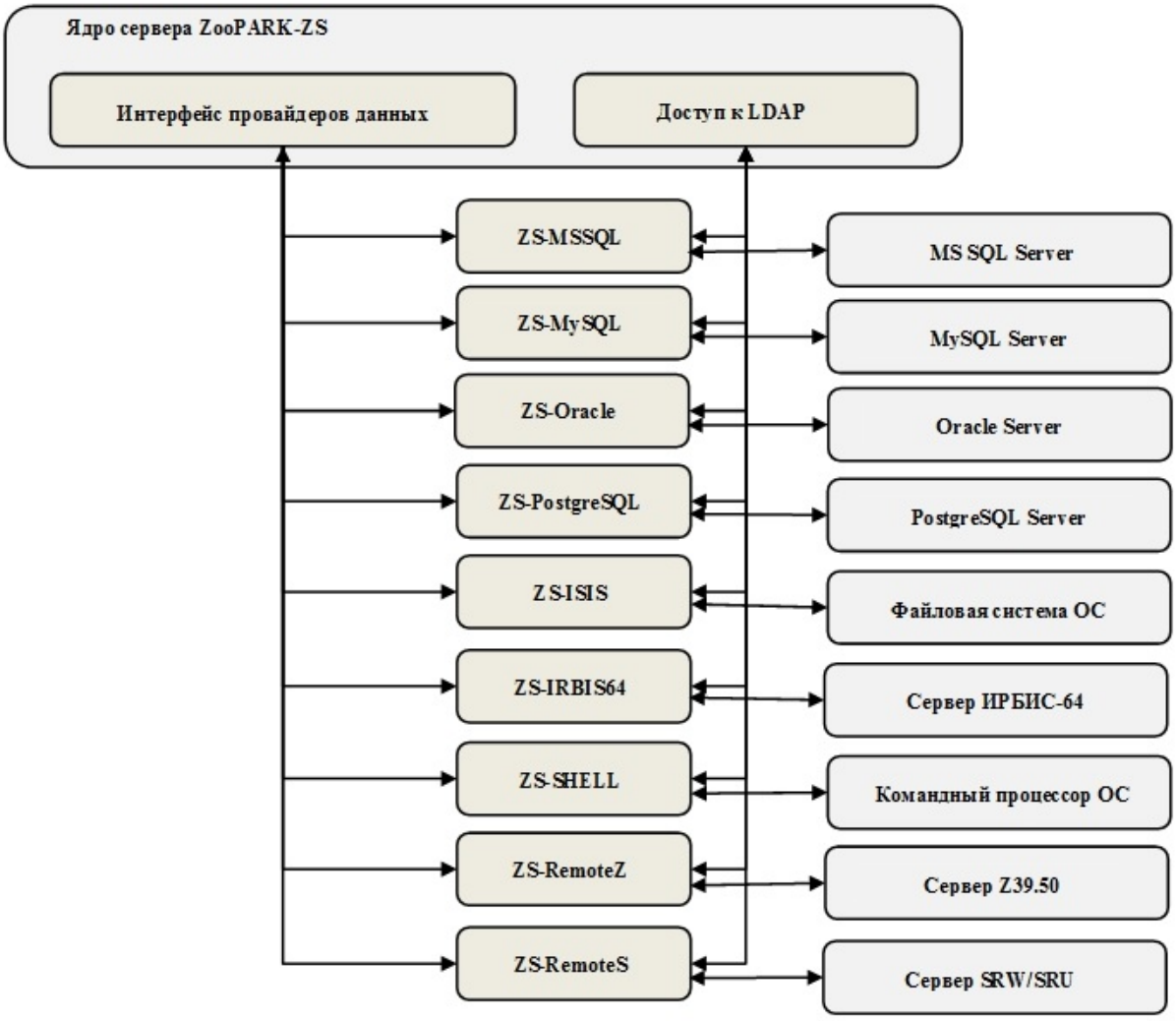

Рис. 3: Доступ к данным сервера ZooPARK-ZS

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 спецификациями;
- 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 (см. рис. 4):

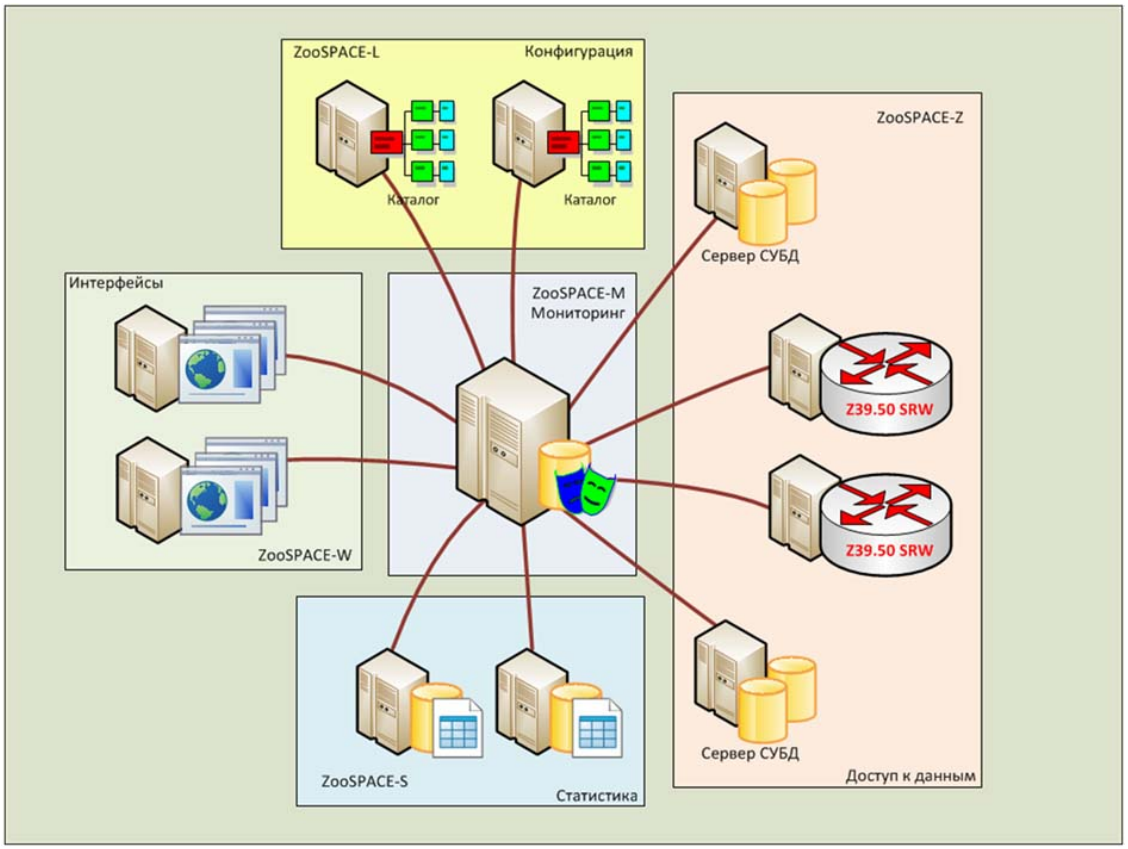

Рис. 4: Основные подсистемы ZooSPACE

– 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.

Литература

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

[2] *Жижимов О. Л., Федотов А. М., Чубаров Л. Б., Шокин Ю. И.* Технология создания распределённых информационно-вычислительных ресурсов СО РАН // Тр. Первой международной конференции САИТ–2005. — 12–16 сентября 2005 г., Переславль-Залесский. «Системный анализ и информационные технологии». — Т. 2, Москва. — С. 161–165.

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

[4] *Отле П.* Библиотека, библиография, документация: Избранные труды пионера информатики / Пер. с англ. и фр. М.: ФАИР-ПРЕСС, Пашков дом, 2004.

[5] *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.

[6] *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.

[7] *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.

[8] *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.

[9] *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.

[10] 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.

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

[12] *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.

[13] *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.

[14] *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.

[15] *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.