Технологическая платформа интеграции разнородных распределенных данных ZooSPACE[1]

ZooSPACE[2] – technological platform for the integration of heterogeneous distributed data

Технологічна платформа інтеграції різнорідних розподілених даних ZooSPACE[3]

О. Л. Жижимов, В. С. Никульцев, Е. В. Никульцева,
А. М. Федотов, Ю. И. Шокин
Институт вычислительных технологий СО РАН,
Новосибирск, Россия

Oleg Zhizhimov, V. S. Nikultsev, E. V.Nikultseva,
Anatoly Fedotov and Yury Shokin
Institute of Computational Technologies of the Siberian Branch
of the Russian Academy of Sciences,
Novosibirsk, Russia

О. Л. Жижимов, В. С. Нікульцев, Е. В. Нікульцева,
А. М. Федотов, Ю. І. Шокін
Інститут обчислювальних технологій СВ РАН,
Новосибірськ, Росія

В докладе рассматривается технологическая платформа массовой интеграции распределённых гетерогенных источников данных. В основе технологической платформы находится программный комплекс с условным названием ZooSPACE, разрабатываемый в ИВТ СО РАН. Комплекс ZooSPACE строится на основе нескольких слабосвязанных распределенных подсистем, обеспечивающих конфигурирование (ZooSPACE-L), доступ к ресурсам (ZooSPACE-Z), пользовательские и административные WEB-интерфейсы (ZooSPACE-W), сбор статистики (ZooSPACE-S) и мониторинг (ZooSPACE-M) всей системы. Обсуждается архитектура и состав каждой из подсистем.
Ключевые слова: распределенные информационные системы, интеграция гетерогенных данных, управление доступом к информационным ресурсам, Z39.50, LDAP, SRW/SRU.

The authors examine the technological platform for the mass integration of distributed heterogeneous data sources. Its core technological platform is a software system, code-named ZooSPACE, developed at ICT SB RAS. ZooSPACE complex is built on the basis of a loosely coupled distributed subsystems that provide configuration (ZooSPACE-L), access to resources (ZooSPACE-Z), user and administrative WEB-interfaces (ZooSPACE-W), collection of statistics (ZooSPACE-S) and monitoring (ZooSPACE-M) of the system. We discuss the architecture and composition of each of the subsystems.
Keywords: distributed information systems, the integration of heterogeneous data, access control to information resources, Z39.50, LDAP, SRW / SRU.

В доповіді розглядається технологічна платформа масової інтеграції розподілених гетерогенних джерел даних. В основі технологічної платформи лежить програмний комплекс з умовною назвою ZooSPACE, розроблений в ІОТ СВ РАН. Комплекс ZooSPACE базується на основі декількох слабкопов'язаних розподілених підсистем, що забезпечують конфігурування (ZooSPACE-L), доступ до ресурсів (ZooSPACE-Z), користувацькі та адміністративні WEB-інтерфейси (ZooSPACE-W), збір статистики (ZooSPACE-S) и моніторинг (ZooSPACE-M) всієї системи. Обговорюється архітектура і склад кожної із підсистем.
Ключові слова: розподілені інформаційні системи, інтеграція гетерогенних даних, управління доступом до інформаційних ресурсів, Z39.50, LDAP, SRW/SRU.

[1] Работа выполняется при финансовой поддержке Министерства образования и науки Российской Федерации (грант № «07.514.11.4130»), а также при частичной поддержке Интеграционных проектов СО РАН.
[2] The project is being fulfilled with the financial support from the Ministry of Education and Science of the Russian Federation (grant № «07.514.11.4130»), and with partial support from RAS Siberian Division Integration Projects.
[3] Робота виконується за фінансової підтримки Міністерства освіти і науки Російської Федерації (грант № «07.514.11.4130»), а також за часткової підтримки Інтеграційних проектів СО РАН.

В докладе описываются основные компоненты технологической платформы интеграции разнородных распределенных данных ZooSPACE, разработанной в результате выполнения Государственного контракта Министерства образования и науки РФ в рамках ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2013 годы» по теме «Разработка принципов и программных средств виртуальной интеграции распределённых источников данных на основе международных стандартов для создания масштабных информационных инфраструктур». Глобальной целью работы является разработка инструментальной платформы (далее – платформа массовой интеграции), поддерживающей создание и функционирование широкомасштабных информационных инфраструктур на основе подхода виртуальной интеграции баз данных. Более узкой целью работы является разработка принципов и программных средств виртуальной интеграции распределённых 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 [1-4], 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 [5].

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. Количество узлов в ZooSPACE не нормируется и может быть любым. Система ZooSPACE может состоять из одного единственного узла.

Такой 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:

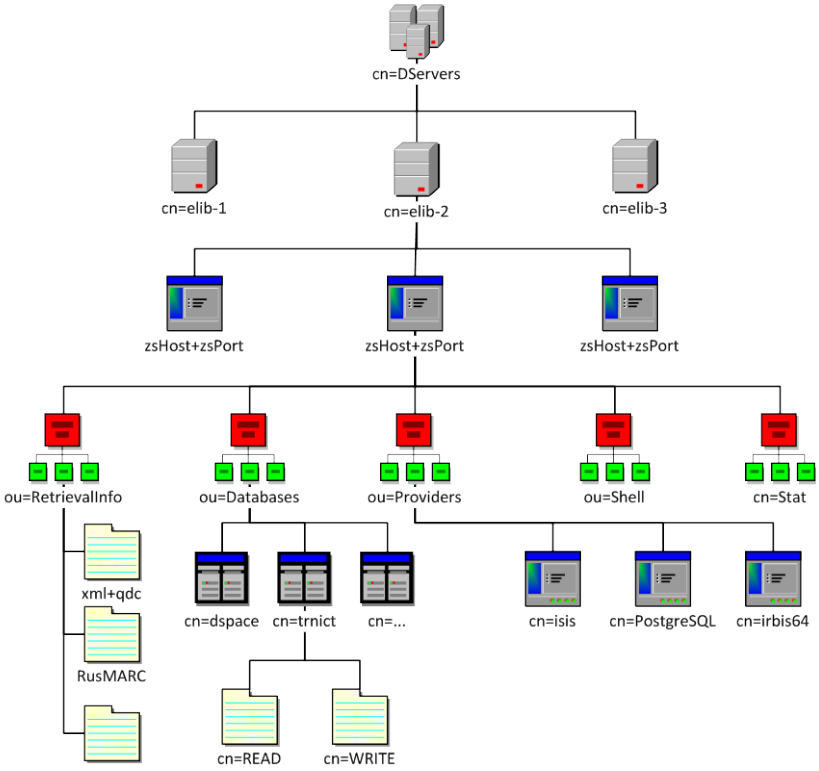

Рис. 1 Структура узла описания хостов в каталоге ZooSPACE

ZooSPACE-L обеспечивает функционирование справочной и административной подсистемы ZooSPACE. Подсистема интегрирует совокупность LDAP серверов узлов, функционирующих в соответствии с единой для всех политикой и хранящих в виде единой иерархической базы данных (системный каталог ZooSPACE, далее каталог ZooSPACE) всю конфигурационную и административную информацию ZooSPACE (см. Рис.1). Все LDAP серверы подсистемы ZooSPACE-L связаны правилом двусторонней репликации каталога ZooSPACE по сетевому протоколу LDAP(S). Количество LDAP серверов в ZooSPACE-L не нормировано. Общую функциональность может обеспечить один единственный сервер LDAP. Количество LDAP серверов ZooSPACE-L может не совпадать с количеством узлов ZooSPACE в соответствии с Рис. 2.

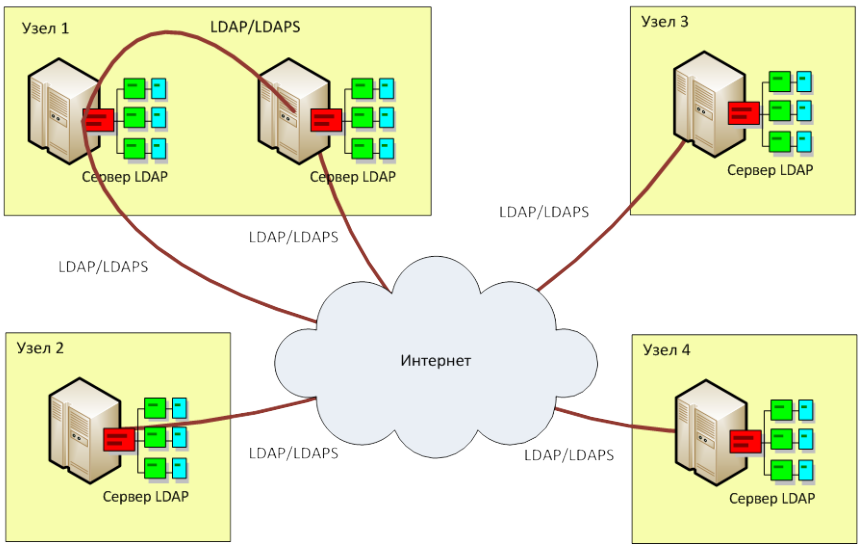

Рис.2 Инфраструктура подсистемы ZooSPACE-L

ZooSPACE-Z обеспечивает функционирование подсистемы доступа к базам данных системы ZooSPACE. Она интегрирует совокупность Z39.50 и SRW/SRU серверов узлов, функционирующих в соответствии с единой для всех политикой (см. Рис.3). В качестве базовых серверов этой подсистемы используется сервер ZooPARK-ZS – модифицированный сервер ZooPARK v.6.1, дополненный необходимой функциональностью в части взаимодействия с подсистемой ZooSPACE-L. Количество серверов ZooPARK-ZS в ZooSPACE-Z не нормировано. Общую функциональность может обеспечить один единственный сервер ZooPARK-ZS. Каждый сервер ZooPARK-ZS в ZooSPACE-Z взаимодействует с подсистемой ZooSPACE-L по протоколу LDAP/LDAPS для получения конфигурационной и административной информации из каталога ZooSPACE. Аутентификация и авторизация всех пользователей ZooSPACE-Z также происходит в подсистеме ZooSPACE-L. Каждый сервер ZooPARK-ZS в ZooSPACE-Z предоставляет интерфейсы доступа к данным по протоколам Z39.50 и SRW/SRU в соответствии со спецификациями этих протоколов и обеспечивает взаимодействие с серверами СУБД, которые по отношение к подсистеме ZooSPACE-Z являются внешними, но могут использовать политику аутентификации и авторизации своих пользователей в подсистеме ZooSPACE-L. Одной из обязательных функцией серверов ZooPARK-ZS является возможность переадресовывать запросы на доступ к данным на другие серверы ZooPARK-ZS подсистемы ZooSPACE-Z, а также на серверы Z39.50 и SRW/SRU, не входящие в ZooSPACE-Z, по соответствующим протоколам.

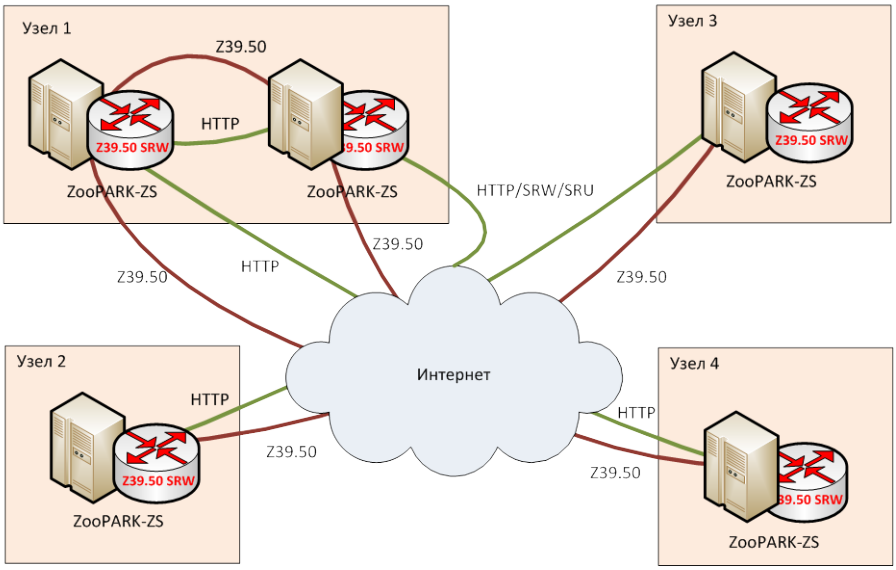

Рис. 3 Инфраструктура подсистемы ZooSPACE-Z

ZooSPACE-M обеспечивает функционирование системы мониторинга всех компонент ZooSPACE. В качестве платформы мониторинга ресурсов ZooSPACE выбран программный комплекс Nagios Core. Nagios Core это программное обеспечение с открытым исходным кодом (http://nagios.org/projects/nagioscore), предназначенное для контроля функционирования IT инфраструктуры и своевременного оповещения администраторов о проблемах, с оборудованием и сетевыми сервисами, которые возникли или могут возникнуть в процессе эксплуатации. Сконфигурированная в соответствии с требованиями ZooSPACE подсистема ZooSPACE-M (см. Рис.4) на основе Nagios Core и расширенная дополнительными модулями позволяет постоянно контролировать доступность серверов, сервисов и ресурсов ZooSPACE, уведомлять уполномоченных администраторов о возникновении критических ситуаций и модернизировать карту доступности информационных ресурсов ZooSPACE в LDAP каталоге. Результаты мониторинга и текущее состояние мониторируемых ресурсов отображаются через специальные административные WEB-интерфейсы подсистемы ZooSPACE-W.

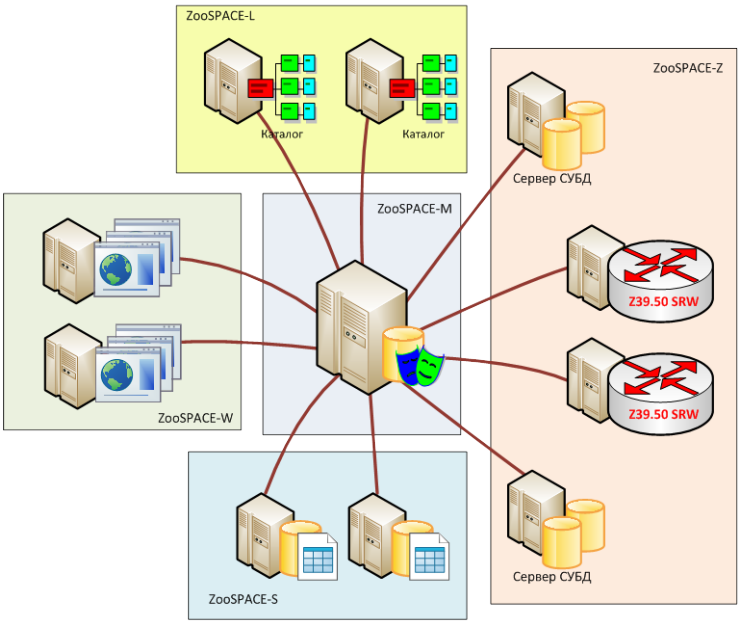

Рис. 4 Взаимодействие сервера подсистемы ZooSPACE-M
с другими серверами ZooSPACE

ZooSPACE-S обеспечивает функционирование подсистемы сбора статистики работы всех компонент ZooSPACE. Подсистема основана на специальных серверах ZooSTAT-ZS с дополнительными модулями сбора, обработки и утилизации log-файлов серверов ZooPARK-ZS [6]. Подсистема ZooSPACE-S является распределенной, управление и конфигурирование ее возможно через подсистему ZooSPACE-L.

ZooSPACE-W обеспечивает предоставление административных и пользовательских WEB интерфейсов для доступа к ресурсам ZooSPACE (см. Рис.5). Каждый сервер подсистемы ZooSPACE-W хранит одинаковый набор программного обеспечения, реализующий необходимую функциональность для формирования WEB интерфейсов и внутренней обработки данных. Клиент может обращаться к любому из серверов без потери функциональности. Наличие нескольких серверов в подсистеме ZooSPACE-W повышает уровень доступности серверов и минимизирует трафик между разными узлами.

Программное обеспечение WEB сервера состоит из нескольких блоков.

Блок Z реализует интерфейсы доступа к подсистеме ZooSPACE-Z. Этот блок обеспечивает поиск и представление данных из различных СУБД в соответствии с выбранным профилем.

Блок L реализует интерфейсы доступа к подсистеме ZooSPACE-L как набор административных интерфейсов доступа к каталогу ZooSPACE. Доступ только для администраторов. Блок содержит функции авторизации пользователей. Фактически этот блок реализует интерфейсы для просмотра и модернизации каталога ZooSPACE.

Блок S реализует интерфейсы доступа к подсистеме ZooSPACE-S. Возможны разные уровни доступа. Интерфейсы предназначены для просмотра статистической информации о работе системы ZooSPACE.

Блок M реализует интерфейсы доступа к подсистеме ZooSPACE-M. Возможны разные уровни доступа. Интерфейсы предназначены для просмотра результатов мониторинга различных компонент ZooSPACE.

В настоящий момент реализован экспериментальный вариант работы технологической платформы интеграции источников данных в составе 3-х узлов: ИВТ СО РАН (центральный узел), ГПНТБ СО РАН и ТФ ИВТ СО РАН (г. Томск). Система предоставлен доступ к 60-и источникам данных (базы данных, электронные каталоги, информационные хранилища, файловые архивы и т.п.), содержащих более 45 млн. записей. Детали конфигурации этого варианта обсуждаются в докладе «Распределенный экспериментальный стенд для ZooSPACE – структура и состав».

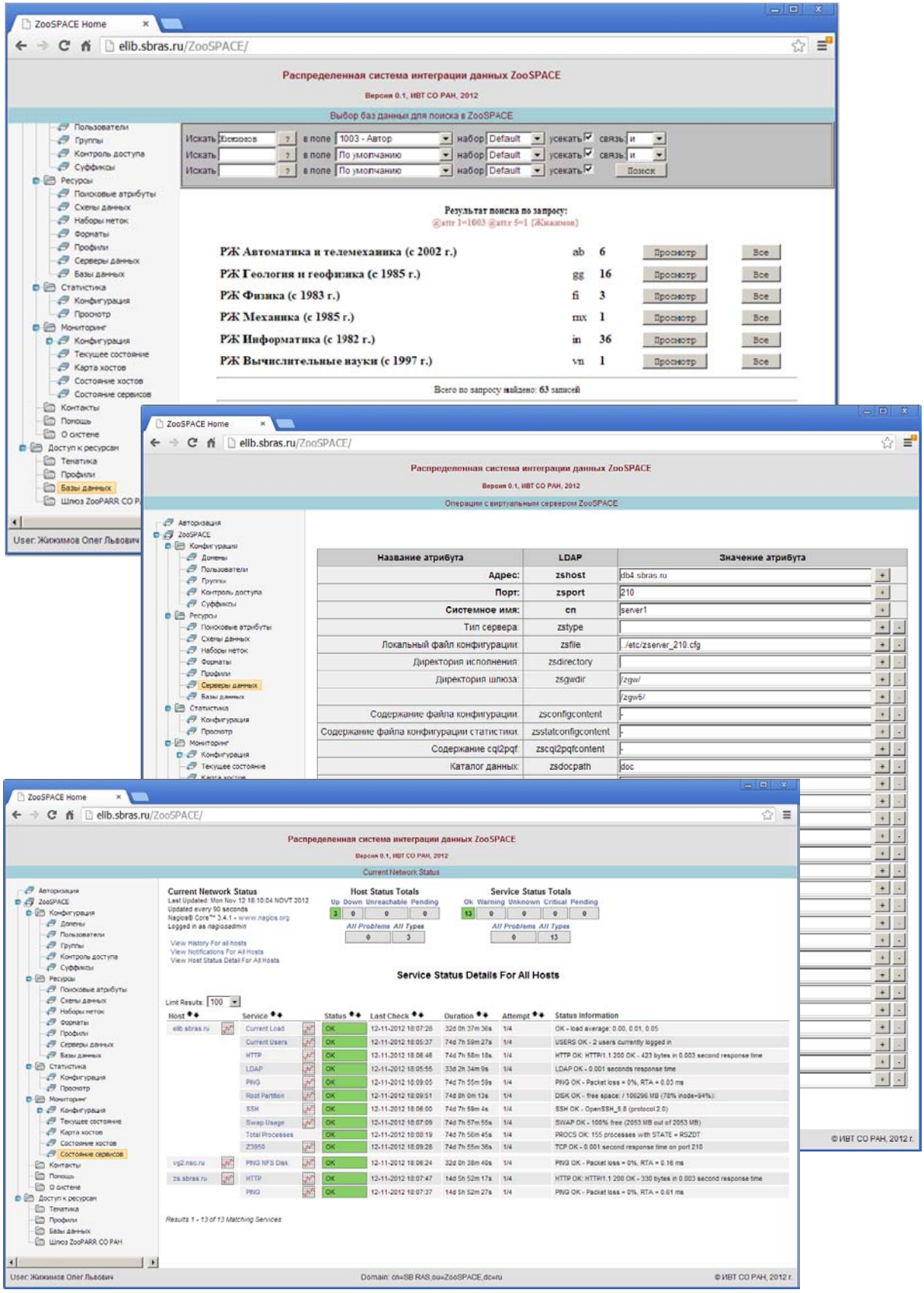

Рис.5 Пользовательские и административные интерфейсы ZooSPACE-W

Список литературы

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Жижимов О. Л., Лобыкин А. А., Турчановский И. Ю., Паньшин А. А., Чудинов С. А. Автоматизированная система сбора статистической информации о событиях в распределенной информационной системе // Вестник НГУ. Сер.: Информационные технологии. – 2013. – Т.11. – № 1. – С.42-52. – ISSN 1818-7900.