ИНТЕГРАЦИЯ РАЗНОРОДНЫХ ДАННЫХ ДЛЯ ЗАДАЧ ИССЛЕДОВАНИЯ ПРИРОДНЫХ ЭКОСИСТЕМ

Жижимов О.Л., Молородов Ю.И., Смирнов В.В., Пестунов И.А., Федотов А.М.

Институт вычислительных технологий СО РАН, Новосибирск

zhizhim@sbras.ru, yumo@ict.nsc.ru, pestunov@ict.nsc.ru, fedotov@sbras.ru

Аннотация

В работе рассматриваются вопросы, связанные с интеграцией разнородных данных для задач исследования природных экосистем. Рассматривается архитектура информационных систем, охватывающих широкий спектр информационных ресурсов – от спутниковых данных и картографической информации, до так называемых электронных библиотек. При этом под интеграцией данных понимается с одной стороны, возможность свободно группировать любые имеющиеся разнородные данные по любому признаку в произвольные реальные и/или виртуальные коллекции, а с другой – возможность организовывать по всем массивам данных прозрачный для конечного потребителя сквозной поиск информации.

Введение

Интеграция разнородных данных, в том числе и данных для задач исследования природных экосистем, реализует подход к изучению этих данных, содержащих информацию об исследуемом объекте, как единый взгляд сразу на весь спектр имеющейся информации. Эта информация накапливается в результате различных научных исследований и аккумулируется в некоторой информационной системе. Следует предполагать, что в такой информационной системе могут существовать, как минимум, следующие типы информационных ресурсов:

- Библиографические электронные каталоги и базы данных
- Библиографические базы данных научно-технической информации
- Полнотекстовые базы данных и электронные библиотеки
- Базы метаданных по различным цифровым архивам (цифровые изображения, аудио, видео) и собственно эти архивы
- Базы метаданных по ГИС-объектам
- Собственно ГИС-объекты (карты, космические снимки и т.п.)
- И др.

При этом под интеграцией данных следует понимать с одной стороны, возможность свободно группировать любые имеющиеся разнородные данные по любому признаку в произвольные реальные и/или виртуальные коллекции, а с другой – возможность организовывать по всем массивам данных прозрачный для конечного потребителя сквозной поиск информации. Но поскольку данные разного типа могут храниться в совершенно разных информационных системах, интеграция ресурсов этих информационных систем возможна лишь тогда, когда доступ ко всем информационным ресурсам во всех аспектах обеспечивается в максимальном соответствии с действующими международными стандартами и рекомендациями уважаемых организаций. При этом стандартизации должны подлежать:

- протоколы и интерфейсы доступа к данным
- поисковые языки и интерфейсы
- схемы и форматы представления данных
- интерфейсы визуализации однотипных данных
- правила кодирования информации
- правила контроля доступа к данным
- правила индексации данных и анонсирования сервисов

Стандартизация необходима для обеспечения интероперабельности, высокая степень которой в свою очередь позволяет не только взаимодействовать с другими информационными системами, но и использовать наработанные технологические решения в других проектах и информационных системах.

Интеграция картографической информации

Данные дистанционного зондирования – важнейшая составляющая информационной поддержки научных исследований в области экологии и природопользования. Особое значение они приобретают при региональном мониторинге обширных территорий Сибири и Дальнего Востока, поскольку они, зачастую, являются единственным источником объективной, независимой и актуальной информации. В ИВТ СО РАН разрабатывается прототип информационной системы для доступа к спутниковым данным (см., например, [1-2]). Работы выполняются в рамках междисциплинарной программы фундаментальных исследований СО РАН № 4.5.2. «Разработка научных основ распределенной информационно-аналитической системы на основе ГИС и веб-технологий для междисциплинарных исследований» (координатор – академик Ю.И. Шокин).

Доступ к информационной системе реализован посредством модуля Central Authentication Service (CAS)[1]. Он позволяет организовать многоуровневую систему разграничения прав доступа с централизованной базой пользователей на основе LDAP-каталога Сибирского отделения РАН и реализовать практически индивидуальные настройки доступа к любому защищаемому ресурсу. Модуль CAS позволяет легко создавать защищенные ресурсы как на основе Apache/Tomcat, так и при использовании технологий PHP/JavaScript на платформе Apache.

Система состоит из следующих функциональных блоков (см. рис.1).

В качестве *HTTP-сервера* используется Apache (с расширением Tomcat) для платформы UNIX.

Важной составляющей системы являются картографические сервисы, которые также используются в других проектах. *Подсистема картографических сервисов* состоит из двух продуктов, распространяемых под лицензией GPL (GeoServer[2] и UMN MapServer[3]).

GeoServer предназначен для публикации набора векторных и растровых слоев. Приложение взаимодействует непосредственно с СУБД PostgreSQL/PostGIS, что позволяет построить высокопроизводительный и легкий в настройке сервис.

---

[1] http://www.ja-sig.org/products/cas/index.html
[2] http://geoserver.org
[3] http://mapserver.gis.umn.edu

Картографический сервер UMN MapServer содержит все необходимое для разработки картографических сервисов WMS/WFS, в соответствии со спецификациями OGC. Он позволяет формировать карты, одновременно используя информационные слои, размещенные как в локальных, так и в удаленных архивах.

В качестве базового *инструментария для обработки и анализа данных дистанционного зондирования* используются как пакеты программ с открытым исходным кодом, так и коммерческие продукты.

Для предварительной обработки поступающих данных используется специальный модуль, интегрированный в коммерческий пакет RSI ENVI[4] 4.5.

В качестве базового инструментария для тематической обработки и анализа данных дистанционного зондирования используется пакет программ с открытым исходным кодом GRASS GIS (Geographic Resources Analysis Support System[5]). Отличительные особенности пакета – полная интеграция в среду UNIX, поддержка основных типов пространственных данных, мощный процессор обработки растровых данных, модульность и наличие открытого инструментария для быстрой и эффективной разработки модулей расширения. По функциональности GRASS GIS не уступает коммерческим аналогам. Он позволяет разрабатывать модули расширения практически на всех языках программирования, для которых есть компилятор под UNIX (Perl, sh, C/C++, Java, Fortran и др.). Пакет позволяет выполнять ресурсоемкие алгоритмы на высокопроизводительных вычислительных системах. Он включает библиотеки для работы практически со всеми современными СУБД.

Для тематической обработки данных в систему интегрирован комплекс программ, основанный на эффективных непараметрических алгоритмах выбора информативных признаков и классификации [3, 4].

Для расширения функциональности системы используется *сервер приложений*. Он содержит интерфейсы для взаимодействия с внешними приложениями, описанные на языке XML.

Для обеспечения функционирования системы в распределенном режиме и интероперабельности по протоколам доступа к метаданным и их представлению, в нее интегрированы *модули поддержки протокола Z39.50* [5]. Поисковая система позволяет не только находить данные по метаданным, но и выполнять комплексные запросы.

В дальнейшем предполагается переход к адаптированным формам предоставления информации, что подразумевает стандартизованную и тематическую обработку «сырых» данных, а также предоставление их в режиме сетевых сервисов. Этот подход позволяет создавать неограниченное количество специализированных систем, базирующихся на одной информационной основе. При этом соответствующее программное обеспечение, установленное у пользователя, может быть максимально адаптировано к его деятельности и уровню квалификации. Подобная архитектура позволяет создавать действительно распределенные информационные системы.

[4] http://www.envi.ru
[5] http://grass.itc.it

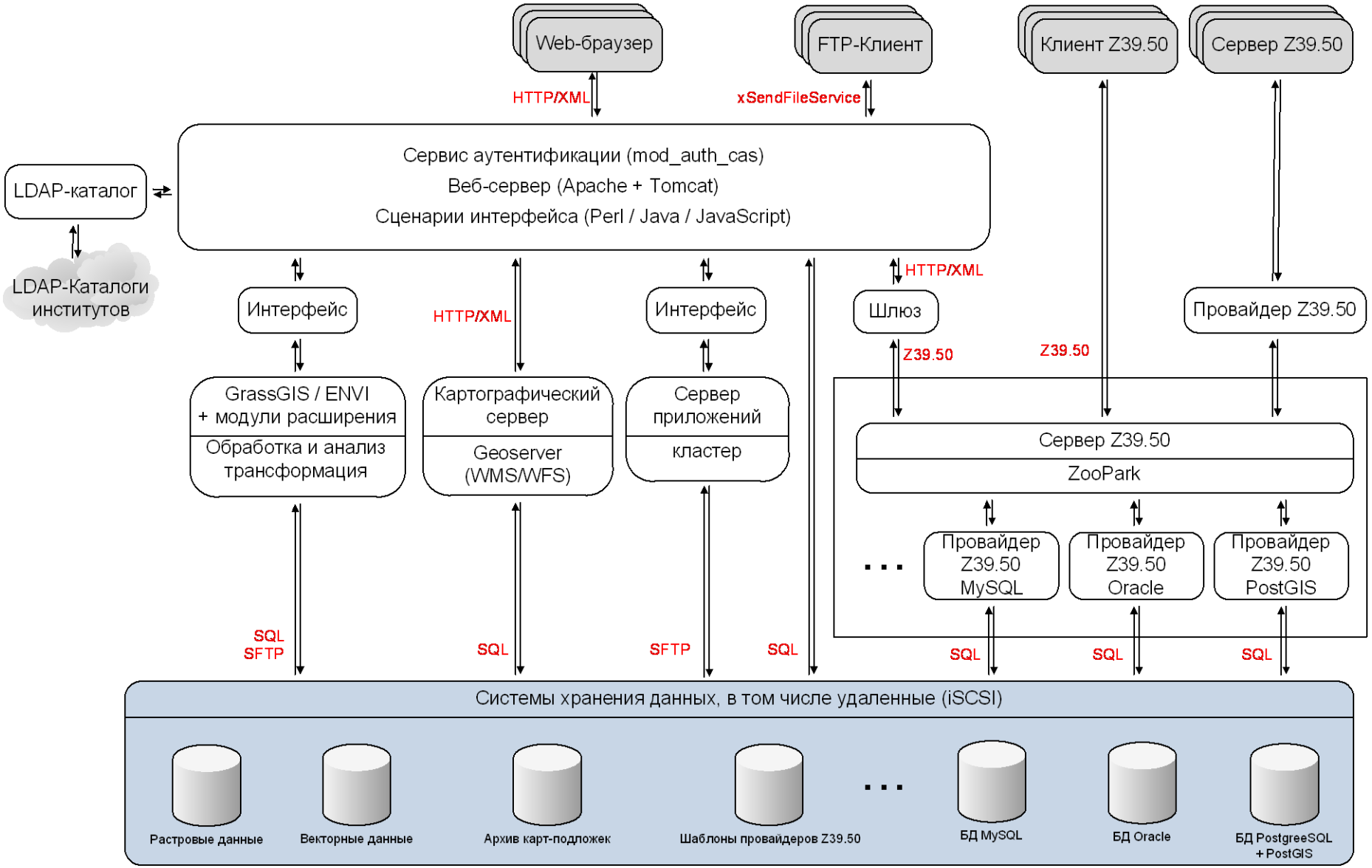

Рис. 1 Структура картографической информационной системы

В настоящее время к системе подключено 14 институтов СО РАН. Система используется для выполнения крупных интеграционных проектов, в том числе в проектах по исследовании. экосистем.

На рис. 2 приведен пример интерфейсов картографического сервиса.

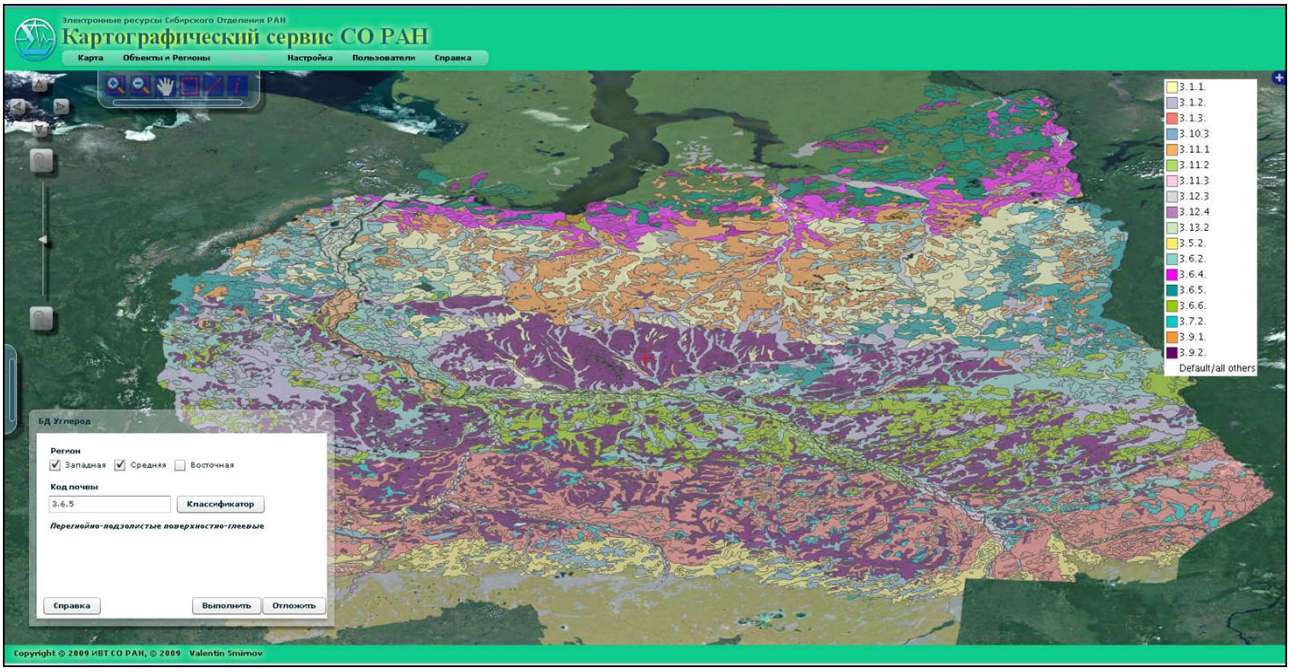

Рис. 2 Карта почв бореальной зоны Западно-Сибирской равнины.

Электронная библиотека

Если для интеграции спутниковых и картографических данных для минимальной функциональности информационной системы достаточно интеграции в рамках, например, каталога GeoNetwork [6], то для более широкого спектра информационных ресурсов функциональных возможностей этой системы становится явно недостаточно. Несмотря на то, что GeoNetwork при загрузке дополнительных схем данных и надлежащей настройки поддерживает каталогизацию и обеспечение доступа к ресурсам различного типа, способы управления ресурсами в этой системе оставляют желать лучшего. В этой системе отсутствует поддержка иерархических коллекций, включающих в том числе и разнородные информационные ресурсы, а также детализированное разграничение доступа к этим коллекциям на основе расширенных ролевых правил. Отсутствие этой функциональности в GeoNetwork существенно ссужает рамки ее использования для коллективной работы по созданию тематической информационной системы.

Результатом научной деятельности, как уже отмечалось выше, является появление разнородных документов

- Текстовые файлы – статьи, отчеты, доклады и т.п.
- Презентации выступлений
- Векторные и растровые изображения
- Аудио и видео записи

- и пр.

Часть этих ресурсов может быть позаимствована из других информационных систем, часть ресурсов требует полнотекстовой индексации для организации полнотекстового поиска. Появляется необходимость расширения возможностей информационной системы дополнительными функциями в части обработки вышеуказанных информационных ресурсов.

В качестве основы системы, интегрирующей данные указанного выше типа, была выбрана система DSpace [7]. Информационная система DSpace обладает широкими возможностями по управлению цифровым контентом, но не содержит интерфейсов для работы с географическими координатами. Учитывая, что DSpace широко используется для создания электронных библиотек, мы не могли пройти мимо соблазна модифицировать эту систему для придания ей дополнительной функциональности.

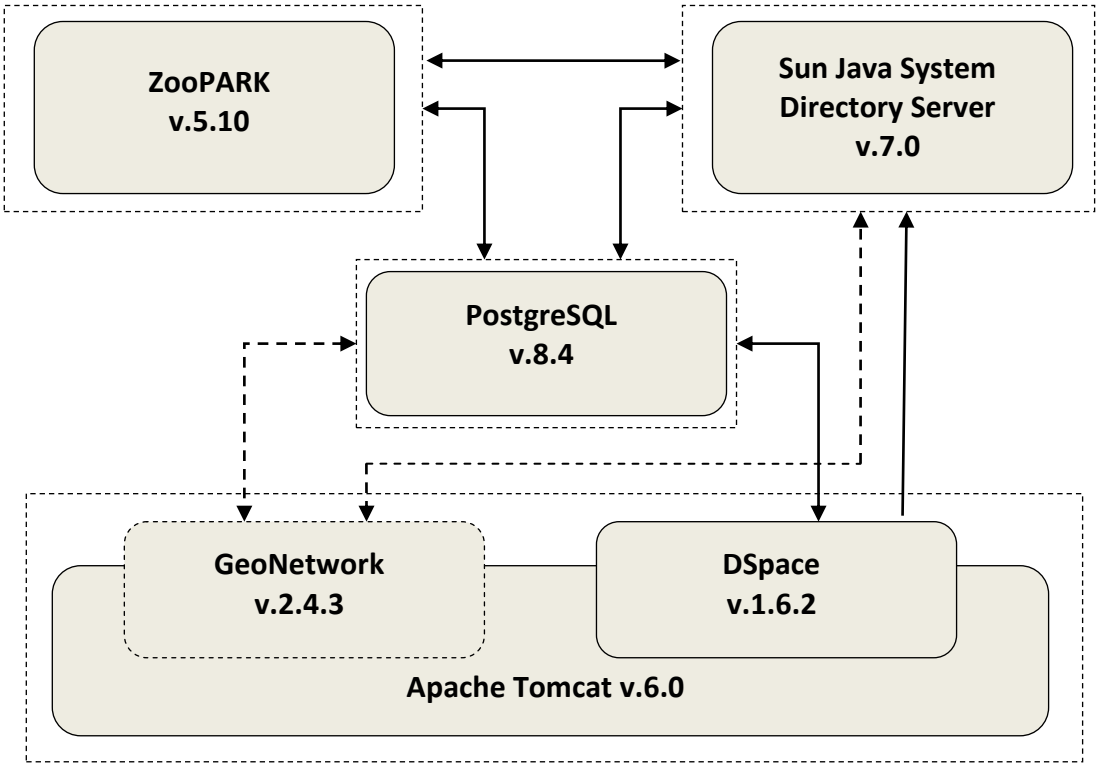

Рис. 3 Структура серверного программного обеспечения информационной системы

На рис. 3 показана общая схема информационной системы, реализующей электронную библиотеку. Наряду с компонентой, управляющей цифровым контентом (DSpace), система включает СУБД PostgreSQL для хранения метаданных, LDAP –сервер для обеспечения функции однократной аутентификации в корпоративном каталоге СО РАН, а также сервер ZooPARK для обеспечения географического поиска в ресурсах DSpace и обеспечения интеграции с другими информационными ресурсами, в том числе и с ресурсами ГИС.

На рис. 4 показаны пользовательские интерфейсы для ввода и редактирования географической информации в модернизированной системе DSpace. При этом достигнутая

функциональность системы позволяет реализовать географическую привязку как для контента, так и для контекста.

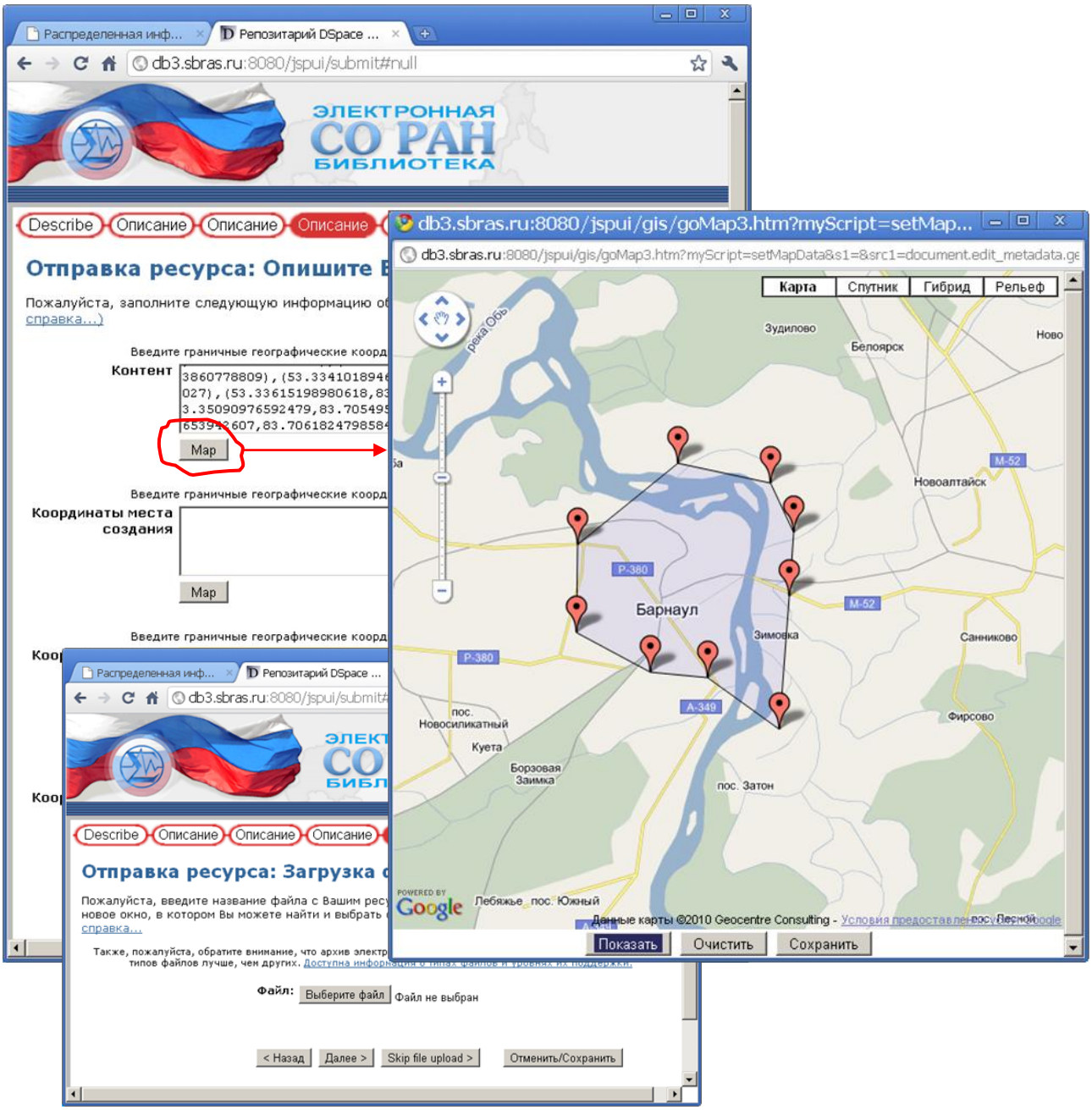

Рис. 4 Интерфейсы модернизированной DSpace для ввода и редактирования географической информации

Поиск информации по различным критериям осуществляется через интерфейсы ZooPARK, который напрямую связан с метаданными DSpace, хранящимися в СУБД PostgreSQL. На рис. 5 показаны интерфейсы шлюза Z-GW сервера ZooPARK для поиска информации. Существенно, что одновременно поиск может происходить по разным информационным источникам. При этом поисковые запросы формулируются в терминах Z39.50 или CIP (для географической информации). Это обеспечивает единый язык запросов для разных информационных систем, не привязанный к схемам и структурам данных конкретных целевых систем.

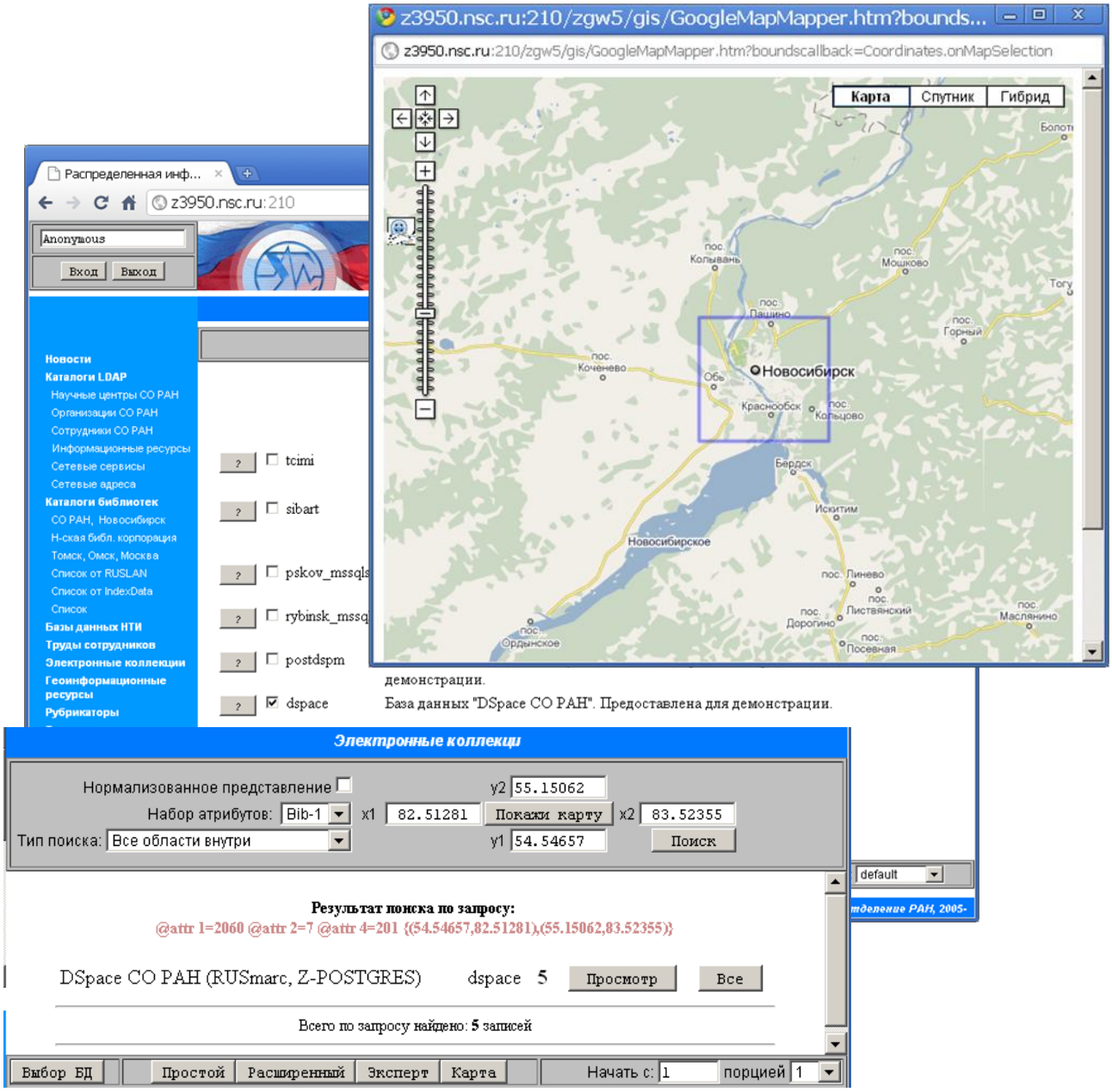

Рис. 5 Интерфейсы шлюза Z-GW комплекса ZooPARK для поиска географической информации

Заключение

Предложенная архитектура интеграции разнородных данных для задач исследования природных экосистем реализована в виде работающего прототипа информационной системы. Дальнейшее наполнение системы информационными ресурсами и активная работа с ними, мы надеемся, позволит эффективно использовать эту систему для научных исследований.

Литература

1. *Пестунов И.А., Смирнов В.В., Жижимов О.Л., Синявский Ю.Н.* Каталог пространственных данных для решения задач регионального мониторинга // Вычислительные технологии. - 2008.- Т. 13. - С. 71-77.

2. *Шокин Ю.И., Жижимов О.Л., Пестунов И.А., Смирнов В.В., Синявский Ю.Н.* Распределенная информационно-аналитическая система для поиска, обработки и анализа пространственных данных // Вычислительные технологии. - 2008. - Т. 12. - Вып: спецвыпуск № 3. - С. 108-115.
3. *Пестунов И.А., Синявский Ю.И.* Непараметрический алгоритм кластеризации данных дистанционного зондирования на основе grid-подхода // Автометрия. 2006. Т. 42, № 2. С. 90–99.
4. *Куликова Е.А., Пестунов И.А.* Классификация с полуобучением в задачах обработки многоспектральных изображений // Вычисл. технологии. 2008. Т. 13 (совместный вып. по матер. Междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании»). Вестн. КазНУ им. аль-Фараби. Сер.: Математика, механика, информатика. 2008. № 3 (58), ч. II. С. 284–290.
5. *Жижимов О.Л., Мазов Н.А.* Принципы построения распределенных информационных систем на основе протокола Z39.50.Новосибирск: ОИГГМ СО РАН; ИВТ СО РАН, 2004. 361 с.
6. GeoNetwork Opensource Community website. – http://geonetwork-opensource.org/.
7. Система DSpace, домашняя страница. – http:// www.dspace.org.