УДК 004

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

Институт вычислительных технологий СО РАН, Новосибирск
пр. Акад. Лаврентьева, 6, Новосибирск, 630090, Россия
Новосибирский государственный университет
ул. Пирогова, 2, Новосибирск, 630090, Россия
E-mail: fedotov@nsc.ru

ИНТЕГРАЦИЯ РАЗНОРОДНЫХ ДАННЫХ В ЗАДАЧАХ ИССЛЕДОВАНИЯ ПРИРОДНЫХ ЭКОСИСТЕМ

Рассматриваются вопросы интеграции разнородных данных для задач исследования природных экосистем. Описывается архитектура информационной системы, охватывающей широкий спектр ресурсов – от спутниковых данных и картографической информации, до так называемых электронных библиотек и алгоритмов обработки, предоставляемых в виде сервисов. Интеграция данных позволяет, с одной стороны, свободно группировать любые имеющиеся разнородные данные по произвольному признаку в реальные и / или виртуальные коллекции, а с другой – организовывать по всем массивам данных прозрачный для конечного потребителя сквозной поиск информации.

Ключевые слова: интеграция разнородных данных, спутниковая информация, информационные системы, Web-сервисы, электронные библиотеки, Z39.50.

Введение

Всестороннее изучение сложных природных экосистем, особенно характеризующихся большой протяженностью и труднодоступностью, невозможно без комплексного анализа всего объёма накопленной информации об объекте исследования, включая спутниковые данные, данные полевых наблюдений, картографическую информацию, известные литературные источники и др. Поэтому в последние годы проблемам совместного использования разнородных данных уделяется повышенное внимание (см., например, [1]). Интеграция разнотипных данных в единую информационную среду обеспечивает возможность их комплексного анализа и позволяет получить качественно новые знания об объекте исследования. Такая информационная среда должна объединять, как минимум, следующие типы информационных ресурсов:

- геоинформационные ресурсы (картографические материалы, спутниковые снимки, данные полевых наблюдений и т. п.), а также соответствующие базы метаданных;
- библиографические базы данных и электронные каталоги;
- полнотекстовые базы данных и электронные библиотеки;
- базы метаданных по различным цифровым архивам (цифровые изображения, аудио, видео) и собственно эти архивы;
- другие ресурсы (аудио- и видеозаписи, электронные презентации и др.), снабженные стандартизованными метаданными.

Под интеграцией данных следует понимать, с одной стороны, возможность свободно группировать любые имеющиеся данные по произвольному признаку в реальные и / или виртуальные коллекции, а с другой – возможность организовывать по всем массивам данных прозрачный для конечного потребителя сквозной поиск информации. Интеграция ресурсов, хранимых в различных информационных системах, возможна лишь при обеспечении унифицированного доступа к ним, максимально соответствующего принятым международным стандартам [2]. При этом стандартизации должны подлежать:

- протоколы и интерфейсы доступа к данным и их поиска;
- схемы и форматы представления данных;

- 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 создаваемые технологические решения в других проектах и информационных системах.

Интеграция пространственных данных

В последние годы в области создания и развития средств и технологий дистанционного зондирования Земли наблюдается стремительный прогресс. Пространственное разрешение снимков повысилось до десятков сантиметров, спектральное разрешение – до сотен каналов. Кроме того, с каждым годом растет число запускаемых спутников высокого и сверхвысокого разрешения. Как следствие, лавинообразно увеличиваются получаемые объемы спутниковых данных.

В то же время накоплен достаточно большой объем эмпирической информации об изучаемых объектах и явлениях (заданной таблицами разнотипных данных, временными рядами и экспертными знаниями, представленными в виде логико-вероятностных высказываний).

Изучаемые в ходе междисциплинарных исследований процессы и явления, а также их взаимосвязи настолько сложны, что только совместный анализ спутниковых и других пространственных данных позволяет получить качественно новые знания и построить адекватные модели природных экосистем. Пространственными будем называть любые цифровые данные, имеющие географическую привязку.

В Институте вычислительных технологий СО РАН разрабатывается сервис-ориентированная геоинформационная система для обеспечения доступа к пространственным данным [3]. Система строится на основе каталога спутниковых данных Новосибирского научного центра СО РАН <mid:00000007/#_ftn1> [1] [4], который включает архивные данные со спутников серии LandSat на территорию РФ за 1982–2002 гг., и с 2008 г. регулярно пополняется оперативными данными SPOT 4. С апреля 2010 г. к каталогу подключена система структурного восстановления оперативных данных, поступающих с платформ TERRA/AQUA. Прием данных осуществляется Западно-Сибирским региональным центром приема и обработки данных в рамках соглашения о межведомственном сотрудничестве.

Система реализована в виде базового набора приложений, работающих в среде сервера приложений Tomcat. Подсистема пользовательских интерфейсов реализована с использованием технологий PHP/JavaScript. Доступ к системе реализован посредством модуля Central Authentication Service (CAS) [2]. 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.

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

Центральным блоком системы является подсистема картографических сервисов, реализованная на основе пакета GeoServer [3]. Подсистема обеспечивает доступ к картографической информации, хранящейся в системе (базовые подложки, векторные слои, построенные по базам данных и др.). Для публикации динамических данных используется пакет UMN MapServer [4], который обеспечивает доступ к данным, формируемым в оперативном режиме, а также к пользовательским наборам данных. На рис. 2 приведен пример интерфейса картографического сервиса.

[1] См.: http://sdc.esemc.nsc.ru/&q=node/12.
[2] См.: http://www.ja-sig.org/products/cas/index.html.
[3] См.: http://geoserver.org
[4] См.: http://mapserver.gis.umn.edu

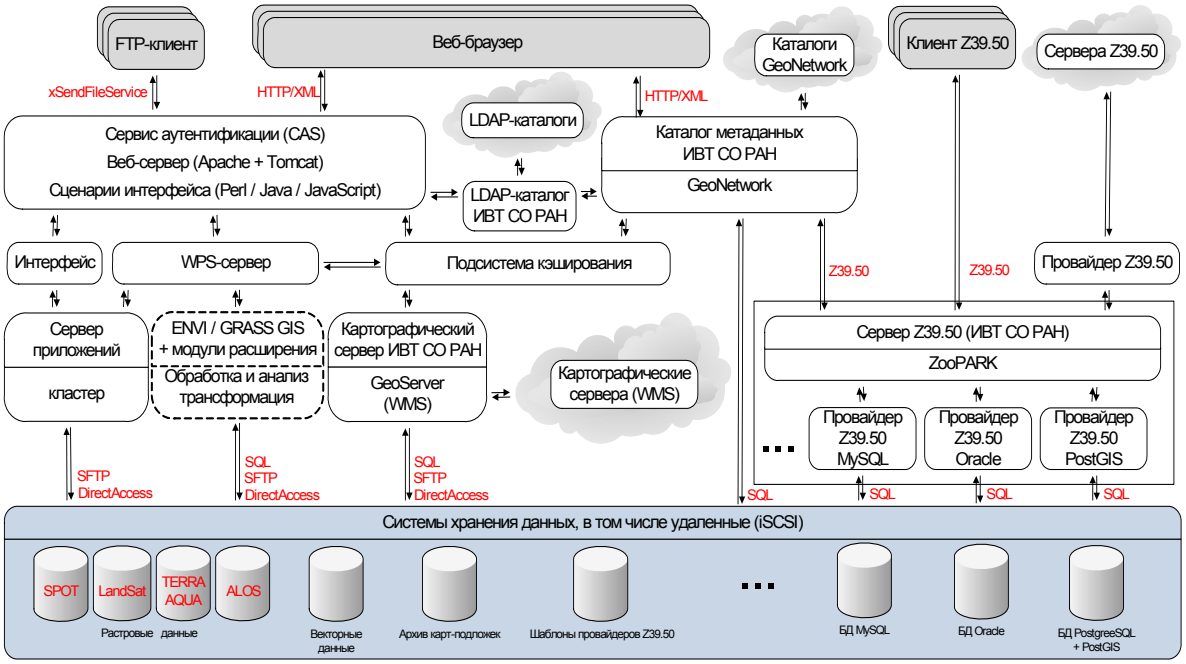

Рис. 1. Структура картографической информационной системы

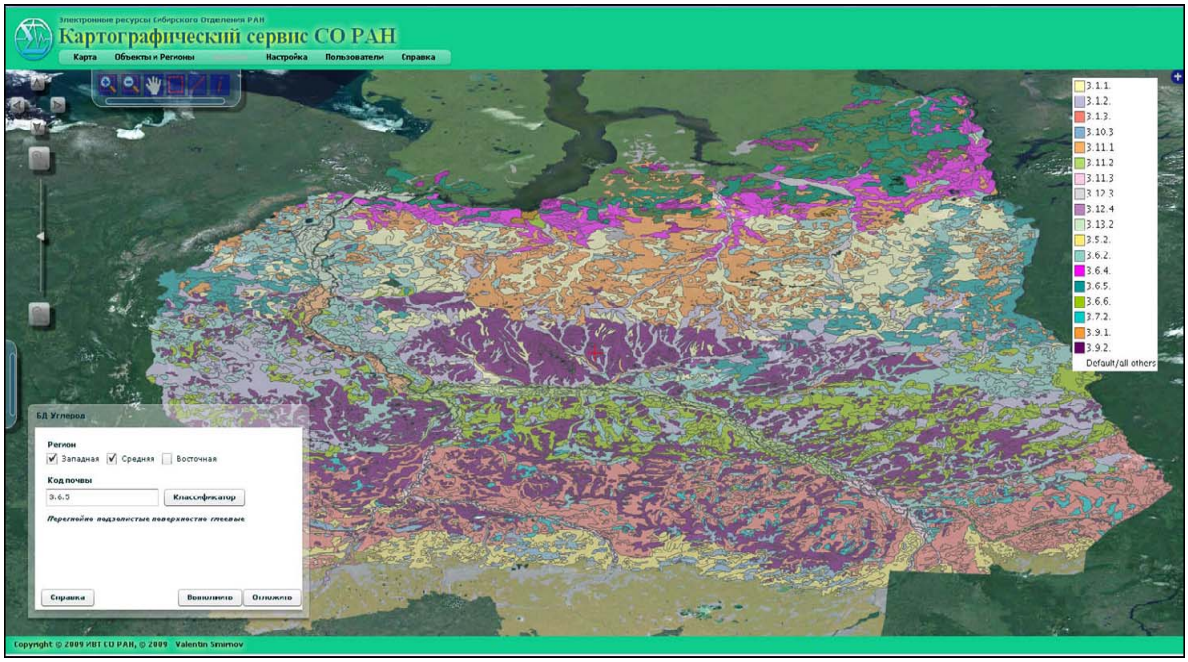

Рис. 2. Карта почв бореальной зоны Западно-Сибирской равнины

Для тематической обработки данных в систему интегрирован комплекс программ, основанный на эффективных логико-вероятностных алгоритмах выбора информативных признаков и классификации [5].

Для расширения функциональности системы используется подсистема сервисов. На основе WPS-сервера deegree, распространяемого по лицензии GPL, разработан модуль для интеграции в систему алгоритмов обработки пространственных данных. Он осуществляет интерпретацию входных и выходных данных согласно спецификации протокола WPS и выполняет функции контейнера для неограниченного числа WPS-процессов. Архитектура модуля представлена на рис. 3.

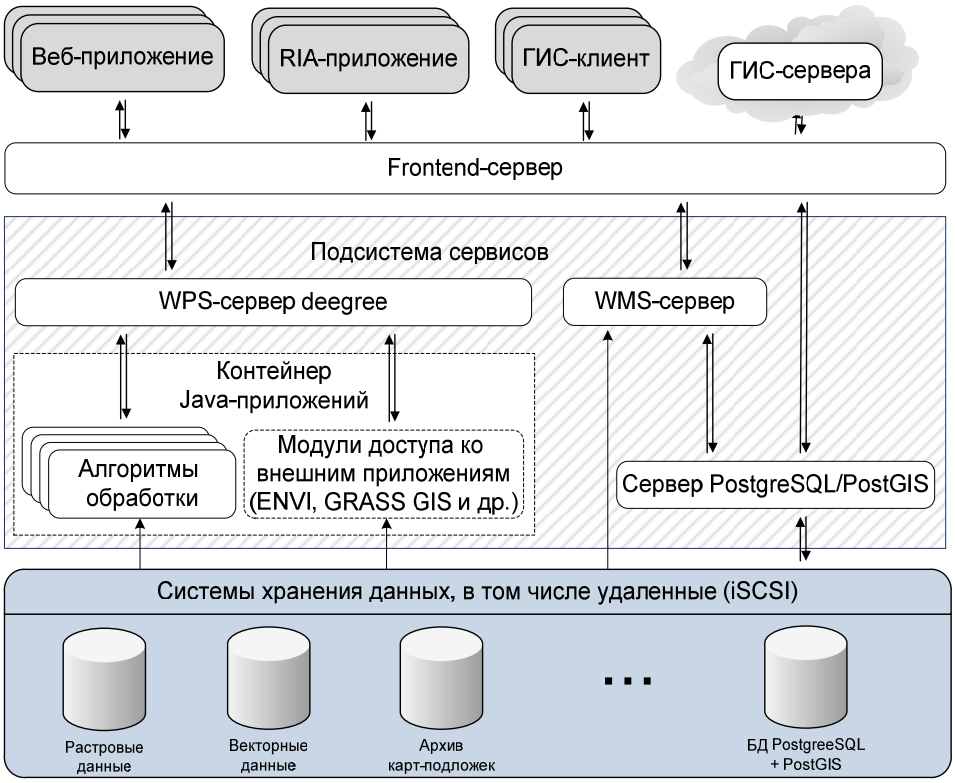

Рис. 3. Структурная схема подсистемы сервисов

Для обработки данных с помощью WPS-процесса пользователь вводит в клиентском приложении адрес WPS-сервера, после чего ему предоставляется список доступных процессов и их описания (метаданные на естественном языке). Выбрав необходимый алгоритм, пользователь указывает значения входных параметров в соответствии со спецификацией протокола WPS. Например, для алгоритмов классификации входными параметрами являются классифицируемое растровое изображение, обучающая выборка (для обучаемых и полуобучаемых алгоритмов), а также набор параметров, специфичных для конкретного алгоритма. Значениями параметров могут быть как данные, находящиеся на компьютере пользователя, так и результаты выполнения запросов к удаленным WPS/WMS-серверам. В этом случае запрос обрабатывается распределенно, без необходимости сохранения промежуточных результатов.

Эта технология позволяет обеспечить широкому кругу потенциальных пользователей доступ к хранилищу современных наукоемких алгоритмов и вычислительным ресурсам, необходимым для оперативной обработки больших массивов разнородных данных.

На данный момент в виде WPS-процессов реализован набор эффективных непараметрических алгоритмов, которые позволяют решать широкий круг задач, связанных с распознаванием образов и анализом многоспектральных изображений. В дальнейшем планируется подключение к системе модулей GRASS GIS [5] в виде WPS-процессов.

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*. Поисковая система позволяет не только находить данные по метаданным, но и выполнять комплексные запросы.

В настоящее время к системе подключено 26 институтов Сибирского отделения РАН. Система используется для выполнения междисциплинарных интеграционных проектов СО РАН.

В систему внедрено несколько информационных ресурсов, разработанных сотрудниками Сибирского отделения РАН в рамках различных проектов и грантов и предоставляемых в виде веб-сервисов, в том числе: 1) векторная карта растительности Западной Сибири и ланд-

[5] См.: http://grass.fbk.eu.

шафтная карта Иркутской области; 2) векторная карта почв бореальной зоны Западно-Сибирской равнины и соответствующая ей карта растительности, содержащая 28 различных типов растительности (М 1 : 7 500 000); 3) база данных по содержанию органического углерода в почвах Сибири, 4) климатологические данные за период с 1989 по 2009 г. на ключевой участок с координатами: 53–70º с. ш., 59–93º в. д., подготовленные сотрудниками Института мониторинга климата и экологических систем СО РАН.

Климатологические данные содержат параметры, оказывающие влияние на растительный покров Западной Сибири и рассчитанные на основе декадных данных по температуре и осадкам реанализа Европейского центра среднесрочных прогнозов погоды с разрешением 0,25º× 0,25º за период с 1989 по 2009 г. К таким параметрам относятся: средняя температура по каждому из 12 месяцев; средняя годовая температура; среднемесячное количество осадков для каждого из 12 месяцев; годовое количество осадков; количество дней в году со среднесуточной температурой выше 5ºС и т. п.

Разработан подход к интеграции разнородных пространственных данных, позволяющий осуществлять их комплексный анализ. Для этого все необходимые данные приводятся к единому пространственному разрешению с использованием регулярной сетки. Разработанный подход применялся для построения информационных моделей бореальных экосистем Западной и Восточной Сибири.

Интеграция электронных библиотек

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, является появление разнородных документов:

- текстовые файлы – статьи, отчеты, доклады и т. п.;
- презентации выступлений;
- векторные и растровые изображения;
- аудио- и видеозаписи;
- и пр.

Часть этих ресурсов может быть позаимствована из других информационных систем, часть 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 не модифицировать эту систему для придания ей дополнительной функциональности.

[6] См.: GeoNetwork Opensource Community website: http://geonetwork-opensource.org/.
[7] См.: http:// www.dspace.org.

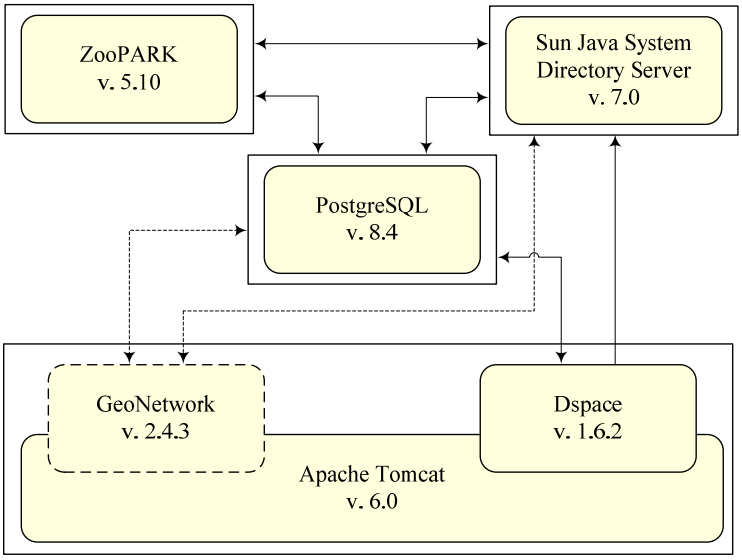

Рис. 4. Структура серверного программного обеспечения информационной системы

На рис. 4 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. 5 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.

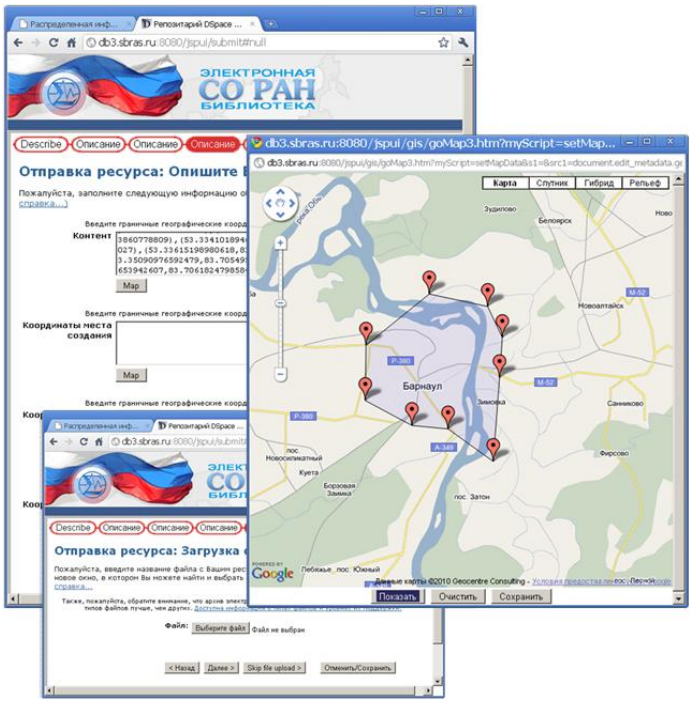

Рис. 5. Интерфейсы модернизированной DSpace для ввода и редактирования географической информации

Поиск информации по различным критериям осуществляется через интерфейсы ZooPARK, который напрямую связан с метаданными DSpace, хранящимися в СУБД Post-

greSQL. На рис. 6 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 [7] или CIP [8] (для географической информации). Это обеспечивает единый язык запросов для разных информационных систем, не привязанный к схемам и структурам данных конкретных целевых систем.

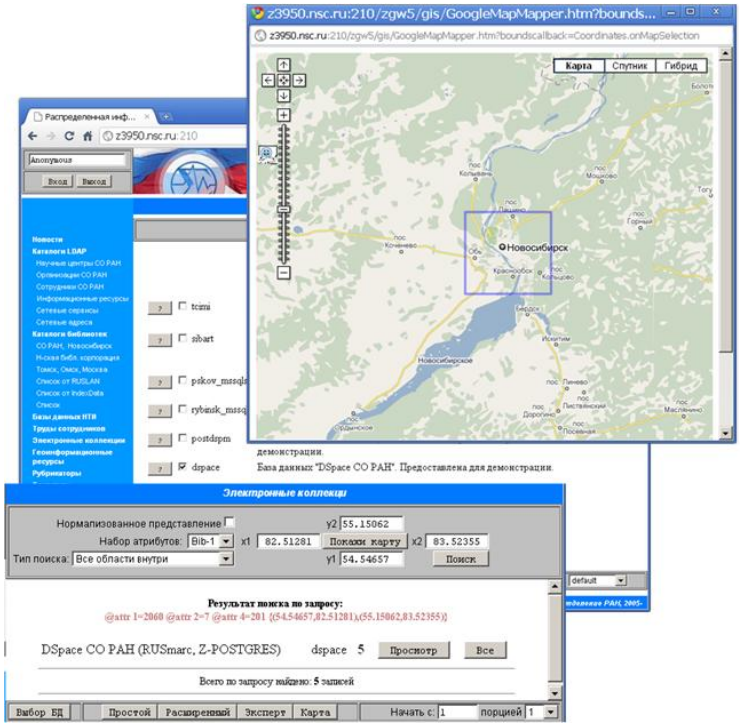

Рис. 6. Интерфейсы шлюза Z-GW комплекса ZooPARK для поиска географической информации

Функция интеграции разнородных данных при поиске в сервере ZooPARK реализуется через модель поиска Z39.50, которая изначально не связана ни с какой целевой системой. В соответствии с международным стандартом это обеспечивается через специальный язык запросов, основанный на использовании стандартизированных наборов поисковых атрибутов и правил их комбинации. В частности, запрос на поиск всех объектов в заданном прямоугольнике ((w,n),(e,s)) на поверхности может быть представлен (набор поисковых атрибутов GEO, поисковый запрос RPN в нотации PQF) в виде

```
@and @and @and @attr geo 1=2038 @attr 2=2 @attr 4=109 e
@attr geo 1=2039 @attr 2=4 @attr 4=109 w
@attr geo 1=2040 @attr 2=4 @attr 4=109 s
@attr geo 1=2041 @attr 2=2 @attr 4=109 n
```

Этот запрос может быть выполнен одновременно в нескольких указанных базах данных прозрачным для пользователя образом.

При использовании наборов поисковых атрибутов CIP [7] приведенный выше запрос может быть упрощен (см. также рис. 6):

```
@attr cip 1=2060 @attr cip 2=7 @attr cip 4=202 {(n,w),(s,e)}
```

Результат выполненного запроса может быть сохранен как временная именованная коллекция (именованное результирующее множество) и повторно использоваться при поиске и просмотра информации.

Таким образом, использование модели работы с данными Z39.50 обеспечивает интеграцию разнородных данных как в виде произвольных коллекций на основе произвольного списка баз данных, так и в виде виртуальных коллекций на основе результатов выполнения поисковых запросов.

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 и подготовки высококвалифицированных кадров.

Список литературы

1. Материалы Всероссийской семинара «Современные информационные технологии для фундаментальных научных исследований РАН в области наук о Земле», 8–11 апреля 2010 г. Владивосток. URL: http://seminar2010.fegi.ru/.

2. *Жижимов О. Л., Федотов А. М., Юданов Ф. Н*. Модель управления информационными ресурсами организации // Вестн. Новосиб. гос. ун-та. 2010. Серия Информационные технологии. Т. 8, вып. 4. С. 81–95.

3. *Шокин Ю. И., Пестунов И. А., Смирнов В. В., Синявский Ю. Н., Добротворский Д. И., Скачкова А. П.* Корпоративная информационная система СО РАН сбора, хранения и обработки спутниковых данных // Горный информационно-аналитический бюллетень. 2009. Отдельный выпуск «Кузбасс 2». С. 3–10.

4. *Пестунов И. А., Смирнов В. В., Жижимов О. Л., Синявский Ю. Н., Скачкова А. П., Дубров И. С.* Каталог пространственных данных для решения задач регионального мониторинга // Вычислительные технологии. 2008. Т. 13. Вестн. Казан. гос. ун-та. им. аль-Фараби. Серия: Математика, механика, информатика. 2008. № 4 (59). Совместный вып. по материалам Междунар. конф. «Вычислительные и информационные технологии в науке, технике и образовании». Ч. 3. С. 71–76.

5. *Добротворский Д. И., Куликова Е. А., Пестунов И. А., Синявский Ю. Н.* Веб-сервисы для непараметрической классификации спутниковых данных // Сб. матер. VI Междунар. научн. конгресса «ГЕО-Сибирь-2010». Новосибирск: СГГА, 2010. Т. 1, ч. 2. С. 171–175.

6. *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.

7. ANSI/NISO Z39.50-2003. Information Retrieval (Z39.50): Application Service Definition and Protocol Specification // NISO Press, Bethesda, Maryland, U.S.A. 267 p.

8. Catalogue Interoperability Protocol (CIP) Specification – Release B // CEOS/WGISS/ICS/ CIP-B, Is. 2.4.75. April 2005.

Материал поступил в редколлегию 10.02.2011

O. L. Zhizhimov, Yu. I. Molorodov, I. A. Pestunov, V. V. Smirnov, A. M. Fedotov

HETEROGENOUS DATA INTEGRATION FOR NATURE ECOSYSTEMS RESEARCH

Questions of integration of the heterogenous data for research problems of natural ecosystems are considered. The architecture of the information system covering a wide spectrum of resources – from the satellite data and the cartographical information, to so-called electronic libraries and the algorithms for processing data as web services is described. Data integration allows to group freely, on the one hand, any available heterogenous data to any sign in real and / or virtual collections and, on the other hand, to organize transparent for the end user information search through all data resources.

Keywords: heterogenous data integration, spatial data, information systems, web services, electronic library, Z39.50 protocol.