ГИС-ПОРТАЛ НА ОТКРЫТЫХ КОДАХ. ПОДХОД К ПРОЕКТИРОВАНИЮ И РАЗРАБОТКЕ

М.М. Кабанов, С.Н. Капустин, П.Н. Колтун,
В.А. Крутиков, Р.Ю. Малахов, П.Б. Милованцев
Институт мониторинга климатических и экологических систем СО РАН, Томск, Россия

29 января 2004 г.

The approach to the creation of GIS-portal as a uniform information base component, uniting the heterogeneous natural measurements and experiments data in environmental monitoring problems is considered. Realization of the specified project is carried out using only opensource systems, and data exchange with users is implemented via http over internet.

Введение

Одной из ключевых проблем научных исследований, особенно мультидисциплинарных комплексных, является обеспечение доступа исследователей к совокупности экспериментальных данных, накопленной в соответствующей области. К сожалению, приходится признать, что, не смотря на гигантский прорыв в области развития информационных технологий, произошедший в последнее время, проблема эта так и остается актуальной во многих научных областях. Причиной этого является не столько недостаток технологий и аппаратной базы, сколько отсутствие продуктивных связей между учеными предметниками и специалистами по информационным технологиям и системному анализу. Создание эффективной и полной информационной системы, в каждом конкретном случае наталкивается, с одной стороны, на естественную некомпетентность ученых конкретной научной области в области информационных технологий, с другой стороны, на столь же естественные проблемы установления единого общего языка понятий и парадигмы предметной области между исследователями и проектировщиками информационной инфраструктуры. Реальным способом решения такой задачи видится создание единой рабочей группы из экспертов в предметной области (в данном случае — ученых соответствующей специальности) и экспертов в области информационных технологий и построения информационных систем.

1. Общие положения

В данной работе рассматривается подход к созданию единой информационной базы для обеспечения комплексных научных исследований, проводимых в Институте мониторинга

климатических и экологических систем СО РАН. Эти исследования проводятся в институте, как в рамках основных направлений фундаментальных научных исследований, так и как соответствующие части российских и международных грантов и комплексных интеграционных проектов Сибирского отделения РАН [1, 2]. Указанная информационная база в настоящее время объединяет разноформатные данные натурных измерений и экспериментов, проводимых в институте, геоинформационные данные и слои, включая совокупность космических снимков и профилей измеряемых величин, а также многолетние ряды данных по метеорологическим измерениям, имеющиеся в свободном доступе по сети метеостанций Западной и Восточной Сибири [3 – 5]. К этому следует добавить необходимые для описания экологических процессов и их взаимосвязи с климатическими параметрами данные, полученные при выполнении комплексных проектов и программ, а также имеющиеся в литературе результаты многолетних биологических исследований, проведенных на научных базах и стационарах Западной Сибири и Алтая [6, 7].

С учетом описанной структуры данных и основной направленности исследований было принято решение о создании распределенной информационно-вычислительной системы (ИВС) с доступом по сети Интернет, использующей необходимые для анализа пространственно распределенных данных возможности геоинформационных (ГИС) технологий [8 – 10]. Указанная ИВС включает три взаимосвязанных, но в то же время имеющих все необходимые автономные средства и ресурсы, уровня:

- средства формирования и управления базами данных, а также комплексными проектами и программами различного уровня — информационный уровень;
- вычислительные средства и блок численных моделей исследуемых процессов — вычислительный уровень;
- средства пространственно-временного анализа и визуализации — ГИС–портал.

Далее будут изложены результаты разработки и создания собственно ГИС–портала, основными задачами которого являются:

- аккумулирование и унификация климато-экологической информации;
- предоставление широкого доступа исследователям к накопленной информации;
- обеспечение возможности отбора информации по комбинации критериев;
- простейшие операции моделирования и визуализации;
- обеспечение экспорта в основные пакеты математического и статистического анализа.

Исходя из перечисленных задач, были сформулированы следующие требования, предъявляемые к системе:

1. Максимальная простота ввода новых данных для исследователей-экспериментаторов. Возможность пополнения данных в режиме реального времени с помощью аппаратно-программных средств, включающих измерительное оборудование, АЦП, вычислительные комплексы. Удобство ввода данных вручную, как в режиме регулярных измерений, так и в случае «пакетного» ввода.

2. Максимально полное, наглядное и гибкое отображение данных для теоретических исследований. Возможность отбора по пространственным, временным критериям и их комбинации.

3. Обеспечение широкого доступа к данным, как в рамках института, так и в рамках всего Сибирского региона с минимальным набором требований к клиентскому месту в точке доступа.

4. Высокая скорость выборки информации. Зачастую для получения желаемой выборки исследователю необходимо проверить несколько вариантов и скорость получения информации в этом случае играет очень важную роль.

5. Низкая стоимость владения системой. Суть и причина требования очевидна. Очевидно так же, что данное требования оказывает ключевое влияние на выбор платформы реализации системы, а также выбор программных средств используемых в процессе реализации.

Таким образом, нами были сформулированы следующие основные положения по проекту создания ГИС-портала:

1. Проект полностью реализуется только с использованием систем на открытых кодах (GPL Лицензия) [11]. *Требование 5.*

2. Проект включает в себя модуль работы с ГИС-информацией и модуль работы с реляционной базой данных. Последнее объясняется тем, что работа с атрибутивной информацией в современных ГИС-платформах (причем как в коммерческих, так и в открытых) реализована лишь на примитивном уровне и не позволяет реализовывать сложные структуры данных непосредственно в рамках ГИС-приложения. *Требования 2,4.*

3. Обмен данных с пользователями осуществляется через сеть Интернет. Доступ к интерфейсу работы с ГИС-порталом осуществляется через HTTP протокол, в качестве ПО клиента выступает браузер пользователя. *Требование 3.*

4. Программная реализация интерфейса пользователя осуществляется с помощью собственного приложения, работающего на стороне сервера, с использованием технологий CGI (Common Gateway Interface) [12]. Реализация интерфейса — графическая, требующая от пользователя только минимальных навыков работы со стандартным ПО ПК. *Требования 2, 3, 4.*

5. Блок сбора данных реализуется отдельным модулем (независимыми модулями) и включает в себя как компоненты, работающие в автоматическом режиме, так и интерфейсы для ручного ввода. В рамках серверного CGI-приложения реализуется веб-интерфейс для администрирования картографической информации. *Требование 1.*

В общей структуре создаваемого ГИС-портала (рис. 1) можно выделить следующие основные блоки:

1. Хранилища данных.

2. Используемое ПО или библиотеки сторонних производителей.

3. Собственное ПО.

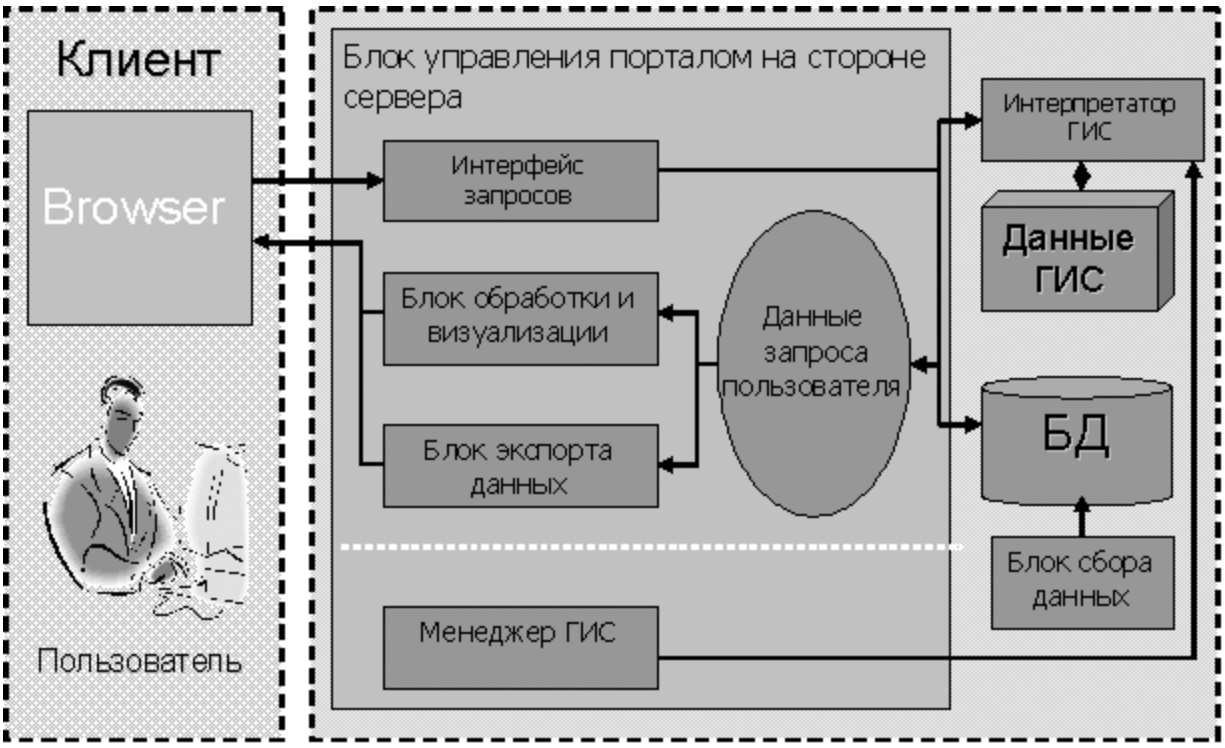

Рис. 1. Общая структура и основные блоки ГИС-портала.

Остановимся подробнее на структуре системы, функциональности и характеристиках ее основных частей. Система в целом представляет собой стандартное клиент-серверное приложение с «тонким» клиентом. В качестве сервера выступает CGI-приложение работающее на платформе Linux/Apache, реализованное на языке Perl. Серверное приложение формирует HTML с интерактивными формами для пользователя и обрабатывает ответ формы пользователя.

Данные, хранящиеся в системе, делятся на два основных вида — пространственно-распределенные данные (или данные ГИС) и данные измерений и экспериментов (как правило, также имеющие пространственную привязку к месту получения данных, но, по своей сути, являющиеся темпорально распределенными). Такое разделение обусловлено не только идеологически, но и технически — на сегодняшний день ГИС-системы не предоставляют возможности хранить атрибутивную информацию сложной структуры (включающую несколько связанных сущностей) в рамках встроенного механизма. В то же время большинство ГИС поддерживают механизм взаимодействия с реляционными СУБД, что позволяет прогнозировать дальнейшее развитие индустрии именно в направлении использования связки ГИС-РСУБД для проектов, требующих использования пространственной привязки в сложно-структурированных моделях данных.

2. ГИС Данные и Интерпретатор ГИС

В результате проведенных исследований, принимая во внимание ограничение, накладываемое положением о проектировании и реализацией системы только с использованием открытого ПО, а также учитывая накопленный на данный момент архив ГИС материалов и его формат, было принято решение в качестве основы для интерпретатора ГИС использовать разработку Университета штата Миннесота (США) (University of Minnesota, UMN) MapServer [13]. Данный продукт предназначен для обеспечения Интернет-доступа к геоинформационым данным. MapServer предоставляет разработчику API для доступа к пространственно-распределенным данным, как в векторном формате (ESRI shape files), так и в растровом (TIFF, JPG, GIF, ERDAS) форматах. Несомненными преимуществами этого продукта для использования в нашем проекте являются:

- MapServer может работать на платформах целиком состоящих из бесплатного ПО (Linux, Apache);
- MapServer поставляется бесплатно с открытым кодом. Таким образом, при отсутствии определенной функциональности в API, разработчик имеет возможность изменить непосредственно код продукта;
- MapServer поддерживает ESRI Shape files — формат, который на сегодняшний день является наиболее широко представленным в ГИС-данных проекта.

Использование MapServer в проекте основано на вызове функций его API из CGI-приложения, реализующего интерфейс пользователя с ГИС-Порталом.

3. База данных по измерениям и экспериментам

Проектирование структуры базы данных осуществлялось в соответствии с соблюдением принципа оптимальности между двумя противоречивыми требованиями: универсальностью и открытостью для включения новой информации и эффективностью эксплуатации базы. Нас сегодняшний момент существуют три основных способа создания структуры базы данных для работы с разнородными данными и обеспечивающие открытость архитектуры для добавления новых данных:

1. Организация в базе структуры метаданных, и хранение самих данных в таблицах единой структуры. То есть, по существу, основная структура сущностей и связей БД описывает наиболее общие понятия моделей данных, а значения данных описывают конкретную структуру информации. Сама же информация при этом хранится в сущностях вида (ID, Поле_Метаданных, Значение_Поля), где Поле_Метаданных — значение поля в одной из таблиц, описывающих структуру, а Значение_Поля — непосредственное значение данных в этом поле. Как правило, при этом организуется множество таблиц значений, соответствующее количеству типов данных необходимых в системе. Подобный подход является своего рода примером «крайней универсализации», когда структура база данных практически независима от структуры данных предметной области. К сожалению, в нашем случае, этот подход вступает в противоречие с *требованием 4*, поскольку запросы к такой базе данных,

как правило, чересчур громоздки, блокируют слишком много данных общего пользования и потому обладают низкой скоростью выполнения, особенно в интенсивно многопользовательской среде.

2. Использование «разреженных» таблиц. Этот подход подразумевает выделение в структуре данных предметной области выделение нескольких основных сущностей, которые прогнозируются неизменными при появление новых видов информации. При этом наиболее интенсивно заполняемые (информационно-насыщенные) таблицы (сущности) искусственно денормализуются, чтобы хранить все возможные виды данных. При этом каждая запись, относящаяся к какому-либо виду информации, использует лишь небольшой процент полей таблицы, так как остальные предназначены для хранения иных видов данных. Например, в нашем случае это обозначало бы наличие в базе таблицы «Измерения», включающей поля и для результатов измерений по температуре, радиации, атмосферному аэрозолю и всех прочих учитываемых системой. Универсальность данного подхода достигается тем, что структура данных при возникновении новых видов данных изменяется лишь на уровне добавления новых полей в разреженную таблицу. Недостатком данного подхода является большой процент избыточной информации в системе. По универсальности он, при этом, первому подходу уступает. Тем не менее, его вычислительная эффективность в процессе гораздо выше.

3. Манипуляция структурой данных на уровне таблиц из кода программы. Данный подход, подобно второму рассмотренному, подразумевает выделение в структуре данных предметной области базового каркаса, устойчивого к изменениям при появлении новых видов данных, а также строгую типизацию данных уже имеющихся по раздельным сущностям. Таки образом, появление нового вида данных влечет за собой создание новой сущности по своим связям копирующей уже имеющиеся для других видов данных, но уникальной по своим атрибутам. Данный подход уступает в универсальности подходам 1 и 2, но превосходит их по эффективности, как с точки зрения скорости доступа, так и объема хранимой информации.

В результате проведенной оценки и изучения предметной области, нами было принято решение остановится на подходе 3, руководствуясь следующими соображениями:

- появление новых элементов метаданных не является в данном случае частым и регулярным, а потому отсутствует необходимость в избыточной универсализации;
- к системе выдвигаются высокие требования по эффективности хранения и обработки информации;
- в условиях многопрофильных исследований, с которыми работает система, разреженность таблицы данных при использовании второго подхода рискует превысить порог эффективности.

В результате на основе вышеуказанных соображений и исследования предметной области была предложена схема данных, изображенная на рис. 2. База данных реализована в СУБД MySQL, что обусловлено требованием к скорости выборки информации — в отсутствие сложной структуры данных и изощренной бизнес-логики приложения этот выбор представляется наиболее логичным.

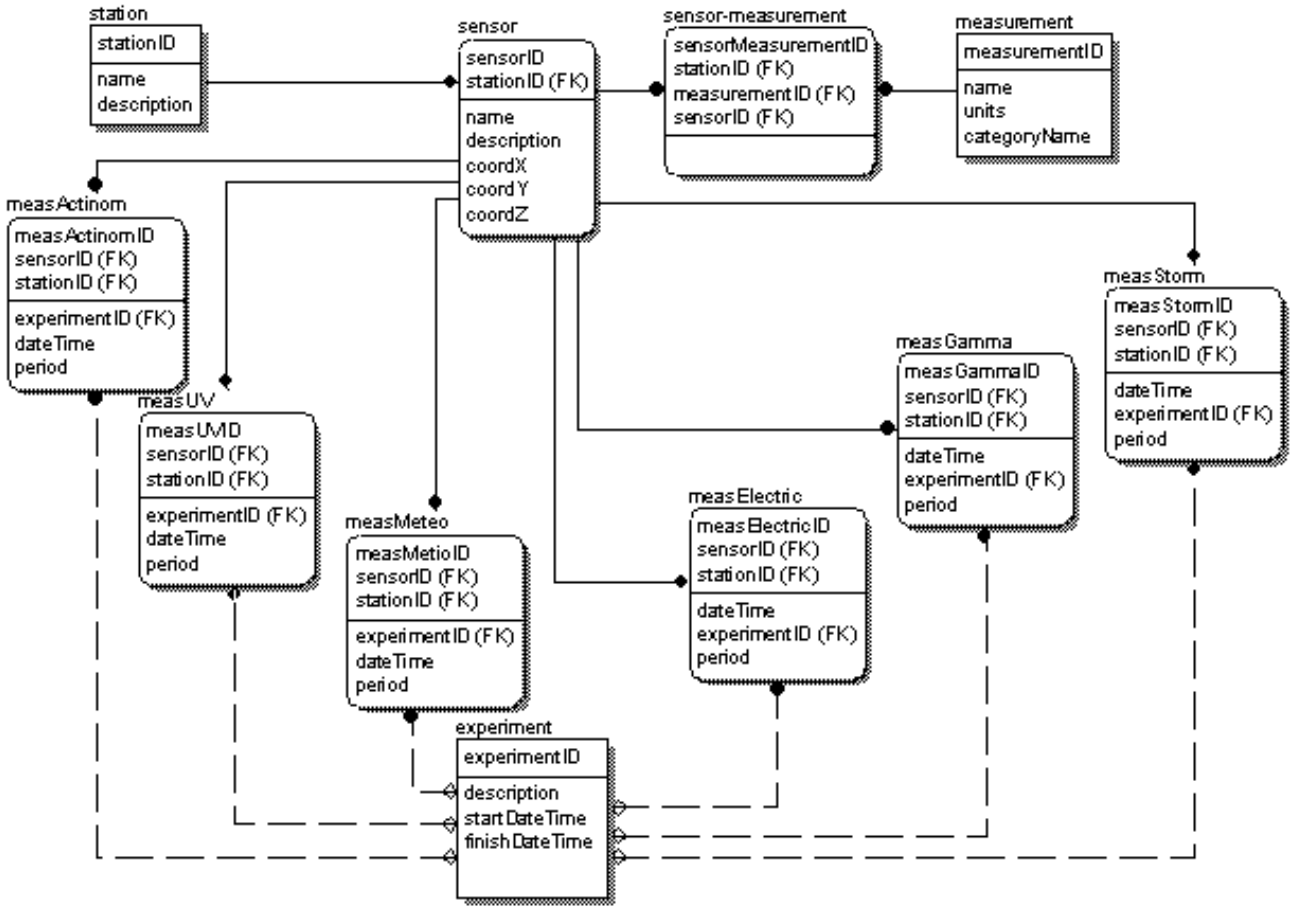

Рис. 2. Схема данных для ГИС-портала.

4. Интерфейс запросов

Без сомнения, предоставление пользователю удобного и гибкого средства для отбора необходимого подмножества информации из данных, присутствующих в системе является одним из ключевых условий для создания востребованной системы. Именно поэтому основное внимание на данном этапе мы уделяем проектированию этого компонента. На сегодняшний день можно отметить следующие основные характеристики, которыми должен обладать этот инструмент:

1. Графический интерфейс
2. Возможность отбора информации по комбинации пространственных и временных критериев
3. Фильтрация выборки по атрибутам
4. Фильтрация выборки по значениям атрибутов

5. Блок обработки и визуализации

Основной функцией данного блока в системе является формирование и отображение информации, полученной либо вычислениями над исходными данными, либо графическое отображение исходных или расчетных данных для визуального анализа. Все вычисления и операции по визуализации осуществляются на стороне сервера средствами CGI-Приложения на языке Perl. Функции визуализации включают в себя построение графиков, как по исходным, так и по расчетным данным, а также построение изолиний в виде отдельного слоя ГИС по данным измерений в нерегулярных отсчетах (например, данные по измерениям температуры по сети метеостанций Западной Сибири). В качестве операций обработки планируется реализовать простейшие статистические функции, наиболее часто используемые в подобных исследованиях — скользящие средние, дисперсия, корреляция между двумя рядами данных и т.п.

6. Блок экспорта данных

Функциональность данного блока должна обеспечить пользователю возможность получить сформированную им через веб-интерфейс выборку на свой локальный компьютер в виде одного или нескольких файлов для дальнейшего анализа с помощью привычных для него программных средств. Очевидно, что попытка объединить в рамках одной системы портал доступа к данным и пакет их анализа высокого научного уровня — обречена на неудачу. Тем более что у большинства исследователей сформировался собственный набор инструментов, которыми они пользуется для расчетов. Таким образом, представляется естественным предоставить пользователю возможность получить данные сформированной им выборки в формате одного из наиболее распространенных пакетов анализа (Statitica, Surfer, MathCad, MathLab и т.п.), либо в одном из универсальных форматов (CSV, XML). Немаловажно также предоставить возможность получения данных несколькими способами, в зависимости от их объема и предпочтений пользователя (http-transfer, ftp, e-mail).

7. Блок сбора данных и Менеджер ГИС

Эти два компонента отвечают за пополнение информации в базе данных. Под блоком сбора данных подразумевается совокупность программно-аппаратных средств, обеспечивающая как автоматизированный, так и ручной ввод данных по измерениям. Данный блок в общем случае не связан с основным CGI-приложением, отвечающим за работу пользователя. В качестве способов пополнения информации в базе данных использованы модели с асинхронным доступом — пополнение базы через промежуточные файлы, выгружаемые АЦП оборудования, либо непосредственный доступ к базе управляющими программами АЦП. Менеджер ГИС является рабочим местом администратора базы ГИС. Этот блок предоставляет веб-интерфейс для добавления новых карт и слоев в портал, а также модификации параметров отображения текущих карт и слоев.

Заключение

На данном этапе еще не все работы по данному проекту завершены. Однако сделан существенный шаг вперед — проведены исследования предметной области, приняты основные проектные решения, реализован ряд блоков и модулей — в частности большая часть CGI-приложения ответственного за работу с комбинацией пространственных и темпоральных данных через Интернет, менеджер ГИС, реляционная база данных. В процессе проектирования и разработки большое внимание участниками проекта уделялось масштабируемости и переносимости получаемых решений – одной из основных целей являлось создание технологии, которая не только позволит решить специализированную задачу в рамках одного института, но и распространить полученные решения, как на другие регионы, так и на другие области исследований.

Работа выполнена в рамках целевой программы СО РАН «Развитие информационных ресурсов Сибирского отделения РАН» при финансовой поддержке гранта Президента Российской Федерации для поддержки ведущих научных школ за 2003 год (проект НШ-1008.2003.5), интеграционных проектов СО РАН «Сибирская геосферно-биосферная программа» и «Комплексный мониторинг Большого Васюганского болота», проекта INTAS 00-00189 ATMOS: A Scientific WWW Portal for Atmospheric Environment.

Список литературы

[1] Тематический выпуск, посвященный 30-летию Института оптического мониторинга СО РАН // Оптика атмосферы и океана. 2002. Т. 15, № 1. С. 1–120.

[2] Большое Васюганское болото. Современное состояние и процессы развития / Под общей редакцией чл.-корр. РАН М.В.Кабанова. – Томск: изд-во ИОА СО РАН, 2002. 230 с.

[3] Крутиков В.А., Полищук Ю.М. Геоинформационное обеспечение мониторинга окружающей среды и климата // Оптика атмосферы и океана. 2002. Т. 15, № 1. С. 12–20.

[4] Дюкарев Е.А., Ипполитов И.И., Кабанов М.В. и др. Региональные особенности современных климато-экологических изменений в Сибири // Большое Васюганское болото. Современное состояние и процессы развития. Томск, 2002. С. 104–110.

[5] Крутиков В.А., Полищук Ю.М., Алексеева М.Н. и др. Применение космической информации в мониторинге ландшафтной структуры Васюганской болотной системы. // Большое Васюганское болото. Современное состояние и процессы развития. Томск, 2002. С. 180–186.

[6] Прейс Ю.И. Криогенез болотообразовательного процесса на территории Большого Васюганского болота. // Большое Васюганское болото. Современное состояние и процессы развития. Томск, 2002. С. 45–63.

[7] Бляхарчук Т.А., Кирпотин С.Н., Воробьев С.Н. Динамика субарктических плоскобугристых болот Западно-Сибирской равнины как индикатор глобальных климатических изменений. // Вестник ТГУ. 2003. № 7. С. 122–134.

[8] Гордов Е.П., Родимова О.Б., Фазлиев А.З. Атмосферно-оптические процессы: простые нелинейные модели. Изд-во ИОА СО РАН, Томск, 2002, 246 с.

[9] Гордов Е.П., De Rudder А., Ипполитов И.И. и др. Веб-портал для представления информации об окружающей среде в Сибири. // Тезисы докладов Международной конференции “Измерение, моделирование и информационные системы как средства снижения загрязнений на городском и региональном уровне” (ENVIROMIS 2002) 6–12 июля 2002 г., Томск. С. 102–103.

[10] Крутиков В.А. ГИС-технологии и Интернет. // Тезисы докладов Международной конференции «Вычислительно-информационные технологии для наук об окружающей среде Cites 2003». 8–11 сентября 2003., Томск. С. 34–36.

[11] Open Source Initiative (OSI). http://www.opensource.org/

[12] Гулич С., Гундаварам Ш., Бирзнекс Г. CGI Программирование на Perl. Символ-Плюс, Спб, 2001, 480 с.

[13] MapServer Home Page. http://mapserver.gis.umn.edu/home.html.

[14] Уолл Л., Кристиансен Т., Орвант Д. Программирование на Perl. «Символ-Плюс», СПб, 2002, 1152 с.

[15] Дейт К. Введение в системы баз данных. Вильямс, Киев, 2001, 1072 с.