«Электронная Земля»: интеграция ГИС-данных с использованием технологий Grid

© Вершинин А.В. Бездушный А.Н. Серебряков В.А.

Вычислительный Центр РАН

alan@ultimeta.ru anb@ccas.ru serebr@ccas.ru

Аннотация

В работе описывается подход к решению задачи интеграции геопространственных (ГИС) данных (на основе стандартов OpenGIS Consortium и ISO) и специализированных методов их обработки с помощью технологий Grid и Интегрированной Системы Информационных Ресурсов (ИСИР). Анализируются требования к составу метаданных для ГИС-данных и программных компонент для их обработки, применимость технологий Grid для осуществления распределенной обработки и хранения ГИС-данных и существующие решения в области данных технологий.

1 Введение

Проект ВЦ РАН «Электронная Земля» имеет своей целью создание программного решения на основе технологий ГИС и Grid с использованием системы ИСИР для интеграции геопространственных данных. На первый этапе проекта осуществляется интеграция метаданных ГИС-ресурсов и методов их обработки (размещенных в сегменте Grid-сети) и предоставление пользователю Web-интерфейса (Web-портала) для удобной работы с ГИС-данными. На второй этапе планируется осуществить интеграцию ГИС-данных на основе технологий Grid: с обеспечением надежного защищенного унифицированного доступа к ГИС-данным, их репликации, средств визуализации на машине пользователя в виде тонкого клиента, система безопасности GSI и т.д.

2 Технологии Grid

Зарождение технологий Grid проходило еще в середине 90-х, область исследований тогда получила название «метакомпьютинг». Метакомпьютером назвали объединение различных вычислительных машин (гетерогенных, распределенных географически, соединенных сетью, возможно Интернет) в виде одной виртуальной машины. Идеи метакомпьютинга были реализованы в многочисленных проектах, существующих и по сей день. Из самых известных стоит упомянуть поиск внеземных цивилизаций в проекте SETI@home, взлом шифрованной фразы RSA Challenges в рамках Distributed.net и один из самых известных локальных менеджеров виртуального пула ресурсов – пакет Condor.

В 1999 году двое американских учёных Ян Фостер (Ian Foster) и Карл Кессельман (Karl Kesselman) своей книгой “The Grid: Blueprint for a New Computing Infrastructure” 5 привлекли внимание к данной тематике сначала научных кругов, заинтересованных в получении доступа к вычислительным системам огромной мощности (производительность виртуальной машины, построенной на технологиях Grid в Японии, составила около 40TFlop), а затем и массовый интерес в коммерческих кругах. Они по праву считаются «отцами-основателями» Grid. Идеи книги приобрели четкие очертания после выхода в свет двух статей, “The Anatomy of the Grid” и “The Physiology of the Grid”, в которых описывается архитектура и требования к инфраструктуре Grid-сети. Также в них было сформулировано «официальное» определение: Grid – это гибкое, защищённое, координированное совместное использование ресурсов группами пользователей, организаций и других ресурсов. В 2001г. Фокс и Гэннон определили GRID как “скоординированное разделение ресурсов и решение проблем в динамической, многокомпонентной виртуальной организации ”, где виртуальная организация – это группа предприятий, объединяющих свои вычислительные ресурсы в единую GRID и совместно их использующая 4.

В начале 2000ых несколько крупных компаний, таких как IBM, Sun, Microsoft, были вовлечены в эту революцию вычисления. Было создано множество коммерческих и некоммерческих продуктов для построения GRID – инфраструктур. Например, инструментарий Globus, который был разработан проектом Globus, стал фактическим стандартом в области Grid middleware. В рамка проекта «Электронная Земля» используется локальный Grid-сегмент, построенный с помощью GT 4.0.1.

3 Гео-Информационные Системы

Геоинформационные системы (географическая информационная система, ГИС) — «комплекс аппаратно-программных средств и деятельности человека по хранению, манипулированию и отображению геопространственных данных» 6.

ГИС связывает информацию с местоположением (например, людей с адресами, месторождения с координатами и т.п.) и своей основной целью имеет организацию эффективного доступа к большим объемам информации об объектах, имеющих пространственную привязку 2.

С развитием Интернет все большее распространение получают Интернет-ГИС, предназначенные не только для распространения и публикации картографической информации, но и для её распределённой обработки, визуализации и обмена в режиме реального времени и т.д. Современные ГИС представляют собой новый тип интегрированных информационных систем, которые, с одной стороны, включают традиционные методы обработки данных, а с другой, обладают спецификой организации, обработки и отображения пространственно-временных данных. На практике это определяет их использование в качестве многоцелевых систем. На создание такой системы направлен проект ВЦ РАН «Электронная Земля».

На данный момент процесс стандартизации главным образом затронул методы доступа к геопространственным данным (Web Map Server, Web Feature Server) и формат их передачи (Geography Markup Language). Со стороны ISO (TC/211) были выдвинуты стандарты на ГИС-метаданные (ISO 19115, 19139). Основной акцент при реализации распределенной геоинформационной среды в рамках проекта «Электронная Земля» делается именно на использование существующих стандартов.

4 Проект «Электронная Земля»

В 2004 году стартовал проект информатизации академического сектора, относящегося к наукам о Земле (геология полезных ископаемых, физика Земли, экология), под названием "Электронная Земля: научные информационные ресурсы и информационно - коммуникационные технологии". Этот проект является одним из пунктов Программы фундаментальных исследований Президиума РАН «Разработка фундаментальных основ создания научной распределенной информационно-вычислительной среды на основе технологий GRID».

Проект имеет своей целью создание на основе технологий IT-инфраструктуры сетевой распределенной информационно-аналитической системы по наукам о Земле, обеспечивающей единообразный доступ к объединяемым в рамках проекта информационно-аналитическим и другим ресурсам. Создаваемая инфраструктура будет предоставлять возможность использования этих ресурсов для решения фундаментальных и прикладных задач, компьютерного моделирования и параллельных вычислений. Участниками проекта являются свыше двадцати институтов РАН, как гео-физической, - химической, картографической направленности, так и решающих задачи в сфере распределенных вычислений и телекоммуникаций.

Началом проекта «Электронная Земля» стал обширный анализ существующих решений по двум направлениям: Интернет-ГИС-системы и технологии Grid. На основе стандартов OGS и ISO, опыта наиболее популярных существующих ГИС-систем (как коммерческих - ArcIMS, так и свободно распространяемых - MapServer, GeoServer, etc.) и ГИС-порталов (на основе GIS Portal Toolkit от ESRI и множества единичных решений) был сформулирован набор требований к архитектуре системы. По направлению Grid очевидным решением было воспользоваться стандартом де-факто в области создания Grid-инфраструктур – свободно распространяемым пакетом Globus Toolkit 4.0.1.

Следующим шагом стало исследование существующих в мире наработок в области решения задачи интеграции и совместного использования ГИС и Grid. Основной интерес представляют проекты:

- GRID on-Demand Services and Infrastructure – GODIS
- Global Monitoring for Environment and Security
- The Solid Earth and Environment GRID
- The Natural Environment Research Council
- The Geosciences Network (GEON)

Направления движения разработок в этих проектах закономерно определяется потребностями ГИС-технологий – процесс накопления данных ускоряется с каждым днём с ростом использования спутниковых системы, позволяющих получать снимки высокого разрешения, он-лайн датчиков, поставляющих данные о различных показателях окружающей среды и т.п., поэтому необходимы эффективные способы передачи, хранения и обработки больших объемов данных. В решении этих задач перечисленные проекты идут по пути уже пройденному аналогичными проектами в области физики, биологии, химии (DataGRID) и т.п. – создаётся Grid-инфраструктура, обеспечивающая распределённое хранение данных и их надежную и скоростную передачу между узлами, далее заинтересованные в решении конкретных задач группы НИИ, учёных и коммерческих компаний создают «виртуальные организации». В рамках каждой ВО на построенной инфраструктуре разворачивается специфическое приложение,

зачастую используемое только участниками данной ВО.

Отличие идей, легших в основу проекта «Электронная Земля», от описанных выше заключается в существенном шаге, сделанном в сторону использования технологий Semantic Web 1 3, интеграции распределенных рабочих процессов и автоматизации их согласованного выполнения 2, унификации механизма создания и публикации вычислительных приложений в Grid. В этом смысле наше видение Grid совпадает с изначальными идеями родоначальников – Фостера и Кессельмана, Grid – это виртуальная вычислительная система, аппаратные и программные особенности реализации которой скрыты от конечного пользователя. Его интересует лишь решение его задачи, поэтому для решения он выбирает алгоритм, указывает входные параметры и исходные данные, оплачивает (или получает на основе «гранта») «вычислительное время» и ждёт результата – точная аналогия с функционированием электрических приборов, включенных в розетку.

На первом шаге реализации распределенной геоинформационной среды главной задачей является интеграция **ГИС-метаданных** из существующих ГИС участников проекта и предоставление механизма публикации и использования вычислительных приложений для обработки ГИС-данных в Grid. Схема среды изображена на Рис.1.

Среда условно делится на две части – ГИС и Grid (в т.ч. и ИС участников тоже относятся к одной или обеим частям), интеграция которых осуществляется с помощью технологий ИСИР 1 3, на которых построен центральный портал. ГИС-часть отвечает за предоставление доступа к распределенным ГИС-данным по стандартным интерфейсам, визуализацию карт, редактирование элементов; GRID-часть обеспечивает распределенное хранение (ГИС-)данных, поддержку ресурсоёмких вычислений; а главный портал - интеграцию, сбор, поиск ГИС-метаданных, каталогизацию ГИС и GRID ресурсов, управление потоками работ по обработке ГИС-данных в GRID и т.д.

Сначала рассмотрим ГИС-часть среды. Она состоит из ГИС-части главного портала и произвольного количества «участников». Выделяется три типа «участников»:

1. поставщики «сырых» данных – не имеют специализированного ГИС-ПО и собственных порталов, зачастую не имеют даже выхода в Интернет; хотят предоставить доступ к своим данным;
2. поставщик данных через стандартный интерфейс – имеет OGC-совместимое ПО и хочет предоставить доступ к своим данным через стандартные интерфейсы WMS, WFS;
3. обладатель самостоятельной ГИС с Web-порталом – имеет свой ГИС-портал (возможно, со средствами визуализации и обработки данных), за которым стоит не обязательно соответствующая стандартам OGC ГИС, хочет опубликовать метаданные своих ресурсов в общей геоинформационной среде.

Схема взаимодействия всех типов участников с главным порталом изображена на Рис.2.

Остановимся подробнее на функциональности ГИС-части главного портала, к ней относится:

- управление статическим содержанием (CMS)
- каталогизация и интеграция ГИС-участников различных типов
- сбор и каталогизация ГИС-метаданных
- атрибутный поиск и визуализация ГИС-данных
- кэширование и сохранение результатов обработки ГИС-данных

Для решения задачи интеграции ГИС-метаданных была сформирована общая минимальная схема на основе owl-реализации стандарта ISO 19115. Набор атрибутов слоя состоит из следующих обязательных элементов:

- Границы/масштаб
- Проекция
- Дата создания/последнего обновления
- Тематика/ключевые слова
- Название/описание
- Автор карты
- Источники, использованные при создании карты
- URL
- Формат исходных данных

Метаданные при загрузке на главный портал претерпевают преобразование к общей схеме – это должны обеспечивать адаптеры на стороне поставщика.

Grid-часть среды включается в себя локальный сегмент Grid – на данный момент состоящий из кластера ВЦ РАН и двух рабочих станций – и соответствующую часть главного портала, состоящую из так называемого Научно-вычислительного портала (НВП) и подсистемы взаимодействия с Grid.

Научно-вычислительный портал является самостоятельным решением на основе технологий ИСИР 1 3, предназначенным для информационного обеспечения деятельности учёных, нуждающихся в решении сложных вычислительных задач. К его задачам относятся:

- Каталогизация «вычислительных приложений», атрибутный поиск ВП в каталоге
- Запуск ВП, мониторинг выполнения и визуализация результатов

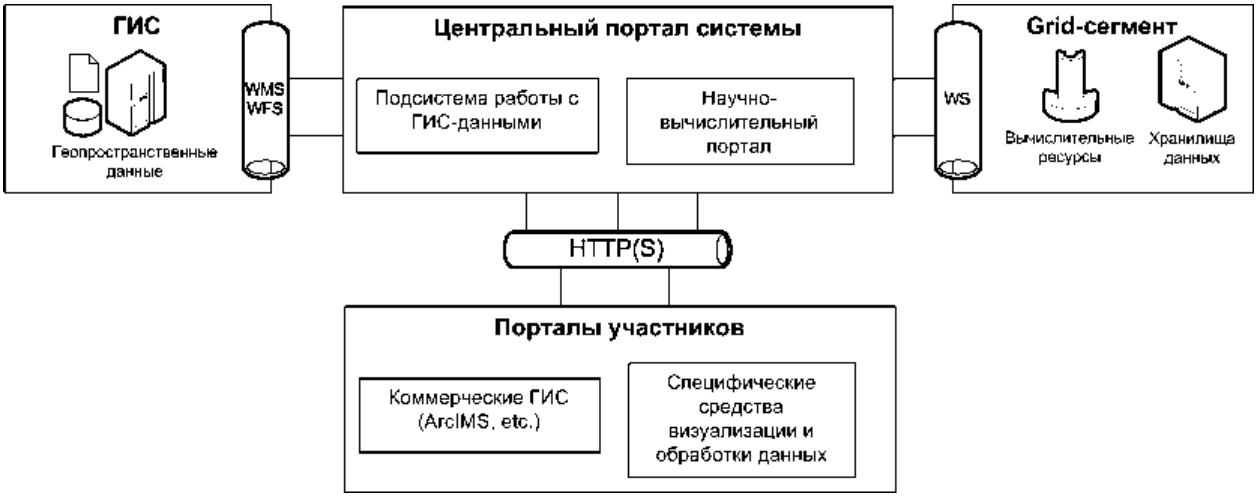

Рисунок 1. Схема распределенной геоинформационной среды.

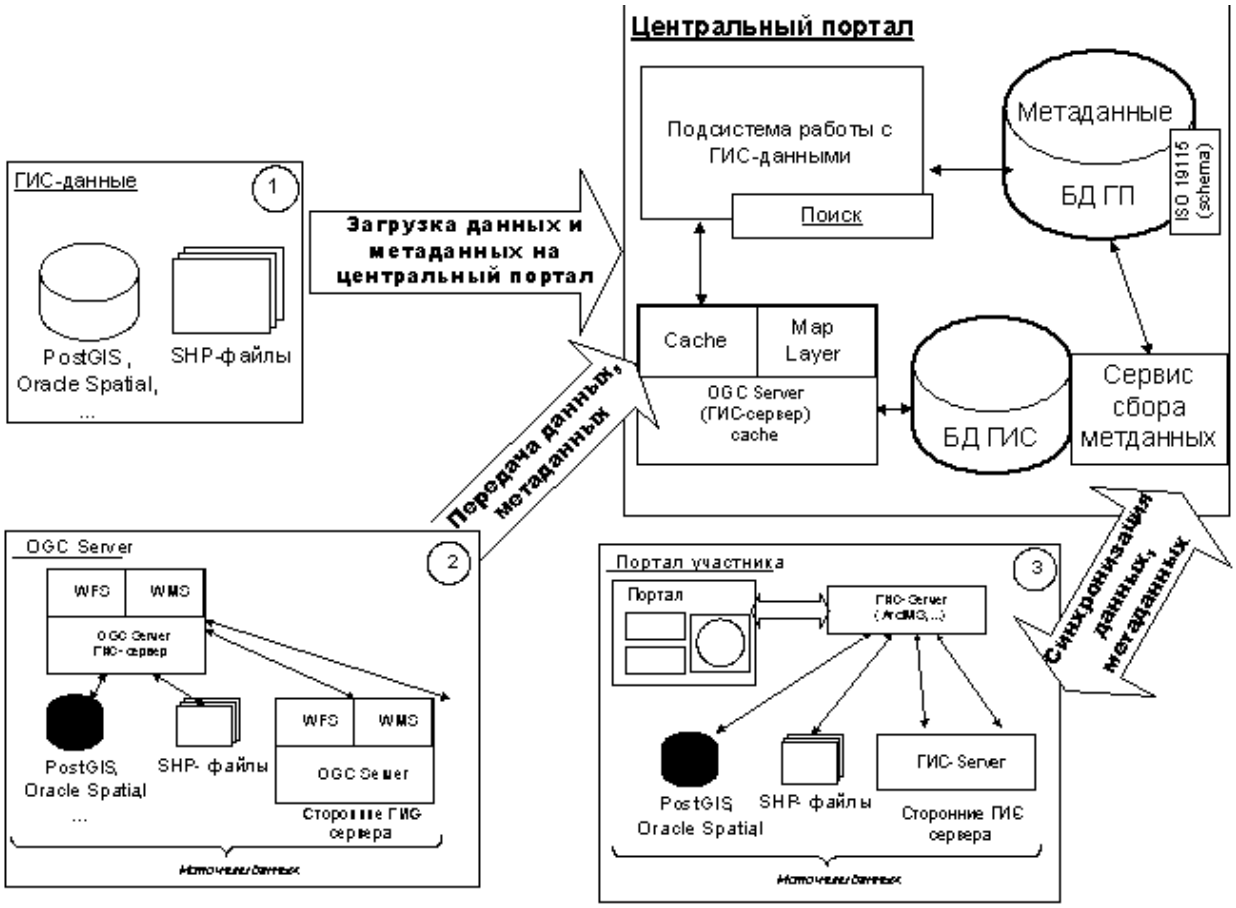

Рисунок 2. ГИС-часть распределенной геоинформационной среды.

- Конструирование и сохранение сложных ВП из существующих

Под вычислительным приложением здесь понимается абстрактная сущность, описывающая некоторый исполняемый программный код, решающий конкретную вычислительную задачу. Причем с технической стороны этот код должен быть оформлен в виде Web-сервиса. Это позволяет осуществлять единообразное взаимодействие НВП с любым ВП, невзирая на особенности реализации. ВП обладает следующим набором атрибутов:

- Название
- Описание
- Создатель/лицо, опубликовавшее ВП
- Дата создания/публикации/модификации

- Ссылка на WSDL-описание или UDDI-репозиторий
- и т.д.

Запуск вычислительного приложения на исполнение обеспечивает динамический SOAP-клиент – по WSDL-описанию строится Web-форма, позволяющая ввести все необходимые параметры для запуска соответствующего Web-сервиса. Вызов сервиса производится асинхронно, что позволяет запускать вычисления любой длительности. При ответе от сервиса пользователю, запустившему вычисления, приходит оповещение о завершении выполнения.

Использование Web-сервисов, как «атомарной» единицы ВП, позволяет использовать технологии построения и выполнения рабочих процессов. В НВП для этой цели используется BPEL-редактор, позволяющий пользователю сконструировать рабочий процесс любой сложности из зарегистрированных в системе ВП и сохранить такой конструкт как **новый ВП.**

Также использование технологий Web-сервисов предоставляет возможность простой интеграции с Grid, так как последние несколько лет основным направлением движения там является переход на использование стандартов W3C. В итоге за Web-сервисом может скрываться сложное распределенное приложение, выполняемое на нескольких узлах Grid, но от клиента этого приложения все особенности реализации скрыты.

В данный момент ведется работа по разработке подсистемы, обеспечивающей управление Grid-сегментом прямо с главного портала. В частности, подсистема размещения ВП в Grid будет обеспечивать загрузку предоставленного клиентом исполняемого кода на указанный узел Grid, его установку и генерацию Web-сервиса для доступа к этому коду.

Также хотелось бы обрисовать дальнейшие перспективы развития описываемой среды. На втором этапе проекта предполагается сделать основной упор на решение задачи интеграции ГИС-данных с помощью таких технологий Grid, как OGSA-DAI и пр. Также существенным доработкам должна подвергнуться система безопасности ГИС-части – в данный момент стандарты OGC никак не регламентируют эту область ГИС. Системы, разработанные на первом этапе, также продолжат развитие – в частности сбор ГИС-метаданных будет переведен на полностью автоматический режим (в случаях конфликтов участие эксперта, конечно, все равно будет необходимо).

Литература

1. А. А. Бездушный, А.Н. Бездушный, А.К. Нестеренко, В.А. Серебряков, Т.М. Сысоев, "Архитектура RDFS-системы. Практика использования открытых стандартов и технологий Semantic Web в системе ИСИР", Пятая Всероссийская научная конференция: "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" – RCDL'2003, Санкт-Петербург, Россия, 2003. http://rcdl2003.spbu.ru/proceedings/J1.pdf
2. А.К. Нестеренко, А. А. Бездушный, Т.М. Сысоев, А.Н. Бездушный, В.А. Серебряков, " Служба управления потоками работ по манипулированию ресурсами репозитория ", Российский научный электронный журнал "Цифровые библиотеки", том. 6, выпуск 5, 2003. http://www.elbib.ru/index.phtml?page=elbib/eng/journal/2003/part5/NBSBS
3. A.N. Bezdushnyi, A.B. Zhizhchenko, M.V. Kulagin, and V.A. Serebryakov, "Integrated Information Resource System of the Russian Academy of Sciences and a Technology for Developing Digital Libraries", Programming and Computer Software, Vol. 26, No. 4, 2000, pp. 177–185
4. G. C. Fox, D. Gannon, "Computational Grids", IEEE Comput Sci Eng. Vol 3,No. 4, , pp. 74-77, 2001
5. Ian Foster and Carl Kesselman (Eds.), The Grid: Blueprint for a New Computing Infrastructure, Morgan-Kaufman, 1998.
6. Материалы Open Geospatial Consortium (OGC), http://www.opengeospatial.org/
7. Материалы открытой энциклопедии Wikipedia, http://www.wikipedia.org

"Electronic Earth": integration of geospatial data using Grid technologies

A.V.Vershinin, A.N.Bezdushny, V.A.Serebryakov

This paper describes our method for the task of integration geospatial (GIS) data (based on worldwide standards, e.g. OGC and ISO) and specialized algorithms for its processing. Using Grid technologies and ISIR (Integrated System of Information Resources). Also we list here the results of the analysis the metadata schema requirements and needed components for its processing, adaptability of Grid technologies for storing and processing geospatial data and short review of the existing projects in this area.