УДК 004:528.8

В.М. Брыксин, А.В. Козлов

РАЗРАБОТКА ТЕХНОЛОГИИ ПУБЛИКАЦИИ ПРОСТРАНСТВЕННЫХ ДАННЫХ СВЕРХБОЛЬШИХ ОБЪЕМОВ С ИСПОЛЬЗОВАНИЕМ ОТКРЫТЫХ СИСТЕМ

Описана технология предварительной обработки и публикации пространственных данных сверхбольших объемов с использованием различных методов оптимизации доступа к исходным данным для достижения оптимальной производительности картографической системы, основанной на открытом программном обеспечении. Рассмотрены способы оптимизации сверхбольших пространственных данных через генерализацию разномасштабных данных и публикацию фрагментированных изображений.

The technology of large spatial data volumes pre-processing and publication using different methods to optimize access to the raw data to achieve optimal performance mapping system based on the open source software. The methods for optimizing very large scale spatial data across different scales of data generalization and publication of fragmented images.

Ключевые слова: карта, космоснимок, мозаика, сервис, каталог, метаданные, геопортал, дистанционное зондирование.

Key words: map, Space image, mosaic, service, catalogue, metadata, geoportal, remote sensing.

В настоящее время проходит процесс создания региональных, национальных и международных инфраструктур пространственных данных (ИПД). Цель ИПД — обобщение и стандартизация пространственных данных для предоставления доступа пользователям к информации различного назначения. При этом подразумевается, что поставщики пространственных ресурсов различного назначения должны размещать их в свободном или регламентируемом доступе в сети Интернет, а конечные пользователи могут искать и получать точные и актуальные пространственные данные из внешних источников для своих проектов. В подобном подходе к организации пространственных данных заинтересованы государственные органы, коммерческие и общественные организации, а также IT-сообщество в целом [1]. Согласно Концепции создания и развития инфраструктуры пространственных данных Российской Федерации [2] ИПД должна содержать информационные ресурсы, организационную структуру, стандарты и технологии.

В то же время бурное развитие технологий дистанционного зондирования Земли (ДЗЗ) накопление данных центрами приема и обработки информации привело к проблеме представления данных сверхбольших объемов (как по собственно объему данных, так и по их количеству). Каталоги метаданных имеются у каждого оператора спутнико-

вых систем, а также в организациях, имеющих крупные архивы пространственных данных. Однако существующие системы либо не решают большинство задач, возникающих при публикации данных, либо не в полной мере обеспечивают доступ к исходным данным, либо являются закрытыми и не предоставляются сторонним пользователям.

Для решения поставленных проблем в Югорского НИИ информационных технологий, при непосредственном участии авторов, разрабатывались технологии создания комплексного банка космических снимков Земли [3—5] и каталога метаданных распределенного банка пространственных данных регионального уровня [6—7].

Работы по созданию каталога метаданных банка пространственных данных Ханты-Мансийского автономного округа — Югры были организованы в соответствии с постановлением правительства Ханты-Мансийского автономного округа — Югры № 23-п от 27.01.2010 г. «О Регламенте формирования и функционирования банка пространственных данных Ханты-Мансийского автономного округа — Югры». Этот регламент относится к пространственным данным, созданным или приобретенным за счет средств бюджета автономного округа. В нем были определены основные участники организационной структуры поддержки распределенного банка данных (государственный и сторонний пользователи, уполномоченный и коллегиальный орган, оператор каталога метаданных), а также их полномочия и возможности. Для реализации положений регламента были разработаны стандартизованные формы для получения сведений об имеющихся в распоряжении государственных пользователей пространственных данных, созданных или приобретенных за счет средств бюджета автономного округа, а также форма заявки на приобретение пространственных данных [6].

Работы были продолжены в Балтийском федеральном университете им. И. Канта в рамках НИР «Разработка информационных технологий тематической обработки, каталогизации и представления данных ДЗЗ и другой пространственно-привязанной информации» и гранта РФФИ 11—07—12058 «Создание ГРИД-сервисов хранения, обработки и визуализации данных ДЗЗ для мониторинга добычи углеводородов». Одной из наиболее важных задач при разработке каталогов пространственных данных, геопорталов и порталов территориально распределенных исследователей в области пространственных данных является задача публикации сверхбольших объемов растровых данных ДЗЗ.

Если классифицировать публикуемые растровые данные, то можно выделить следующие виды растровых пространственных данных, участвующих в работе каталога:

1) обзорные изображения космических снимков и данных цифровых моделей рельефа;

2) исходные данные дистанционного зондирования Земли различного разрешения и цифровые модели рельефа;

3) разномасштабные мозаики, составленные из данных дистанционного зондирования Земли различного разрешения.

В зависимости от поставленной задачи объем публикуемых данных может сильно варьироваться. Так, если задача состоит в предварительном просмотре данных с целью визуального ознакомления с их качеством, то объемы публикуемых данных будут небольшими и измеряться десятками мегабайт. В случае публикации исходных данных ДЗЗ или цифровых моделей рельефа конечный объем данных будет варьироваться в зависимости от их пространственного разрешения или масштаба, а также площади области покрытия. Размеры таких данных могут составлять от сотен мегабайт до нескольких гигабайт. Наиболее сложной является задача публикации разномасштабных мозаик данных ДЗЗ, используемых в качестве растровой подложки при разработке тех или иных геоинформационных систем. В последнем случае речь идет о публикации десятков и сотен терабайт растровых данных. В качестве примеров представления сверхбольших объемов данных ДЗЗ можно привести такие системы, как «Яндекс-карты», «Карты Google», «Microsoft Bing» и др.

На основе проведенного анализа существующих систем и наборов исходных данных были выделены требования к публикуемым растровым данным. Во-первых, так как данные будут использованы в геоинформационных системах, необходимо обеспечить базовые геоинформационные функции при работе с такими данными, которые включают в себя возможность выбрать определенную область данных с заданным пространственным разрешением или масштабом (функции масштабирования и панорамирования). Во-вторых, требуется обеспечить возможность изменения проекции данных без дополнительной их обработки. В-третьих, доступ к данным должен быть организован по открытому протоколу, что обеспечит простоту интеграции таких пространственных данных в различные коммерческие (ESRI ArcGIS, Mapinfo и др.) и открытые (GDAL [8], QGIS и др.) геоинформационные системы, а также широкие возможности интеграции данных с картографическими веб-системами (Leaflet API, Google maps API).

Кроме того, в зависимости от назначения публикуемых растровых данных и области их применения могут возникать различные требования к производительности сервисов, предоставляющих растровые данные. Основные критерии, позволяющие оценить требуемую производительность сервиса, — количество пользователей, а также класс геоинформационной системы, работающих с сервисом. Зачастую эти критерии могут быть взаимозаменяемы. Требования к критерию количества пользователей можно дифференцировать на малопроизводительные сервисы, обеспечивающие одновременную работу с данными ограниченного круга лиц одного или нескольких подразделений организации, и высокопроизводительные сервисы, обеспечивающие одновременную работу сотен и тысяч пользователей из различных организаций. С точки зрения класса геоинформационной системы можно выделить настольные геоинформационные системы и публичные картографические системы с доступом из сети Интернет. В некоторых случаях с определенными ограничениями можно говорить, что класс сис-

темы может определять количество пользователей системы, а следовательно, и требования к ее производительности.

При публикации сверхбольших объемов растровых данных возникают вполне определенные трудности в обеспечении требуемой производительности сервиса. Эта проблема возникает в связи с необходимостью производить масштабирование, панорамирование и перепроецирование данных, так как, с одной стороны, эти операции требуют значительных объемов оперативной памяти для обработки данных, а с другой стороны, значительных вычислительных ресурсов системы. В случае, когда объем растровых данных составляет несколько гигабайт, производить такие операции в реальном масштабе времени становится невозможно, потому что не обеспечивается адекватное время реакции сервиса на действия групп пользователей, а в случае работы с данными сверхбольших объемов даже одного пользователя.

Задача обеспечения требуемой производительности сервиса может быть решена с использованием определенных методов оптимизации доступа к исходным данным. Оптимизировать доступ к растровым данным на разных масштабах можно путем предварительной генерализации данных для нескольких дискретных масштабов, это позволит исключить из цепочки обработки необходимость масштабирования данные. Другая проблема, возникающая при работе с большим объемом данных, — это перепроецирование данных в требуемую пользователю проекцию. Эта проблема решается путем ограничения пользователя в выборе проекций посредствам создания дискретного списка проекций и предварительным приведением данных в эти проекции. Однако, исключив указанные операции из цепи обработки данных, остается нерешенной задача выбора требуемой области данных. Оптимизировать операцию выбора можно путем разбиения исходных растровых данных на регулярные фрагменты. В этом случае необходимо выбрать нужные фрагменты, попадающие в заданную область, и склеить их, что позволит загружать в оперативную память не весь массив данных, а только требуемую пользователем часть.

В зависимости от начального объема растровых данных и решаемых пользователем задач методы оптимизации доступа к данным можно группировать. Так, в наихудшем случае, когда необходимо обеспечить доступ к растровым данным сверхбольших объемов большому количеству пользователей, возникает необходимость использовать все три метода, что обеспечит оптимальную производительность сервиса за счет введения в данные избыточности. Степень избыточности в этом случае будет двукратная для каждой требуемой пользователем проекции.

Задача публикации сверхбольших объемов растровых данных с использованием предложенных методов оптимизации доступа успешно решена в программном продукте Geoserver [9], который распространяется в открытых исходных кодах по лицензии GPL. Используя Geoserver, можно публиковать данные следующими способами:

1) растровые данных небольшого объема — в форматах World Image или GeoTIFF без применения методов оптимизации;

2) растровые данные среднего объема — в формате Image Mosaic, который применяется для публикации данных, разбитых на регулярные или нерегулярные фрагменты;

3) данные большого объема — в формате Image Pyramid, позволяющем публиковать предварительно генерализованные к дискретным масштабам данные с разбиением каждого масштаба на регулярную мозаику;

4) данные сверхбольшого объема — в формате Image Mosaic JDBC, который по методам оптимизации аналогичен Image Pyramid, но использует для индексации фрагментов пространственную базу данных PostGIS на базе СУБД PostgreSQL.

При создании каталога пространственных данных для публикации растровых изображений были использованы все вышеперечисленные методы, в зависимости от размеров данных. Так, для публикации обзорных изображений космических снимков высокого и среднего пространственного разрешения, имеющих небольшой объем, применялись форматы GeoTIFF и Image Mosaic. Для публикации обзорных изображений космических снимков низкого и сверхнизкого пространственного разрешения — формат Image Pyramid (рис. 1). Растровая спутниковая подложка интерактивной карты каталога была опубликована с использованием формата Image Mosaic JDBC (рис 2.)

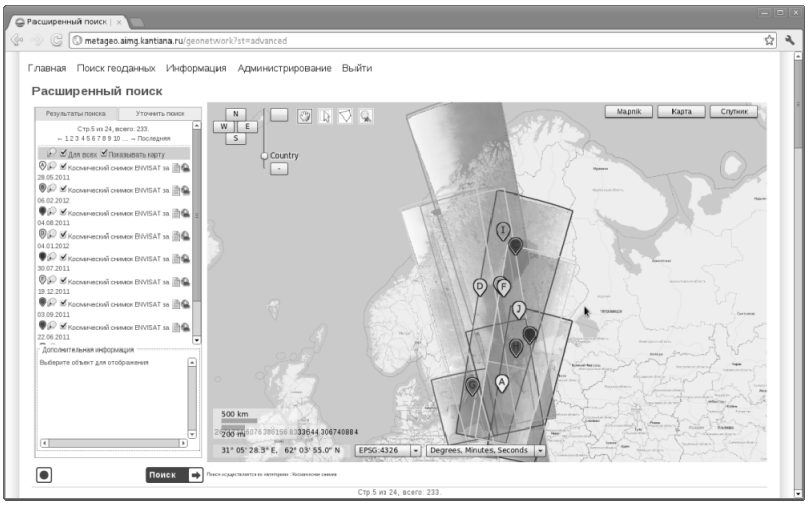
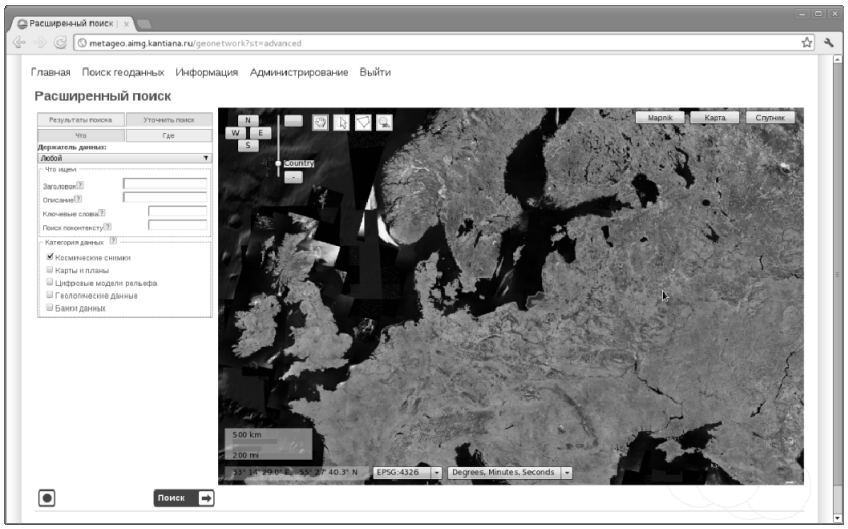


Рис. 1. Представление обзорных изображений космических снимков низкого пространственного разрешения ENVISAT/SAR в каталоге пространственных данных

Для предварительной обработки космических снимков, а именно перепроецирования из исходной картографической проекции в проекцию Google Mercartor, были разработаны программные утилиты на языке программирования IDL, имеющегося в составе программного комплекса ITT ENVI. Эти утилиты позволяют обрабатывать данные в автоматизированном пакетном режиме. При генерализации данных



исходное изображение разбивается на фрагменты размером 512 на 512 пикселей. Уровень масштабирования выбирается автоматически в зависимости от размеров исходного изображения. Генерация шейп-файлов по полученным фрагментам формата PNG осуществляется посредствам утилиты gdal_retile из пакета GDAL. Для автоматизации процессов была разработана утилита запуска программных компонент, что позволяет проводить обработку данных неквалифицированным пользователям.

Рис. 2. Представление мозаики космических снимков Landsat/ETM на территорию суши в качестве растровой подложка интерактивной карты каталога

Размещение и публикация обзорных изображений осуществляется путем простого копирования разномасштабных мозаик в каталог с данными Geoserver и их последующей регистрации в качестве слоя данных через веб-интерфейс управления Geoserver. Ввиду того, что количество публикуемых обзорных изображений было велико, была разработана программная утилита, позволяющая публиковать обзорные изображения в автоматизированном режиме с использованием программного интерфейса управления Geoserver REST RPC. После публикации данных в Geoserver они становятся доступны в виде wms- и wcs-сервисов по протоколу HTTP.

Подготовка данных мозаики космических снимков Landsat/ETM в формате MRSID с пространственным разрешением 28,5 м также вызвала трудности, так как объем получаемой мозаики, по предварительным расчетам, превышал 4 терабайта, что не позволило провести обработку данных стандартными способами. Для решения этой проблемы было предложено создать для каждого слоя набор мозаик размера 102400 на 102400 пикселей. Чтобы реализовать указанную методику, были разработаны программные утилиты, позволяющие в автоматизированном

пакетном режиме осуществлять обрезку исходных изображений и получение набора мозаик, а также подготовку генерализованных данных.

Размещение и публикация данных полученного супероверлея проводились в два этапа. На первом этапе размещались растровые разномасштабные данные в пространственной базе данных PostGIS с использованием штатной библиотеки из пакета библиотек geotools gt-imagemosaic-jdbc. Эта библиотека позволяет создавать метатаблицы фрагментов растровых данных и загружать в БД сами фрагменты изображений. Для последующего ускорения загрузки и инициализации растрового слоя в Geoserver в метатаблицах необходимо указать геометрические размеры пикселя данных и пространственной области, покрывающие данные для каждого масштаба. В процессе импорта данных в PostGIS формируется файл настроек растрового слоя данных, который необходимо скопировать в каталог данных Geoserver. На втором этапе осуществляется публикация данных в качестве растрового слоя в Geoserver с использованием веб-интерфейс управления.

Так как в проекте РФФИ необходимо было обеспечить доступ для научного сообщества к предобработанным и обработанным снимкам MODIS спутников TERRA и AQUA и фотопланам, содержащим растровые и векторные данные об объектах нефтедобычи, и результатам их мониторинга за 2004—2012 гг. на территории Западной Сибири общим объемом более 2 терабайт. Обработка исходных снимков до 2-го уровня осуществлялась в свободно распространяемом пакете IMAPP. Подготовка геопривязанных изображений производилась в пакете ITT ENVI. В связи с тем, что количество исходных снимков превосходило 20 тыс., были разработаны программные утилиты, позволяющие неквалифицированным пользователем проводить обработку в автоматизированном режиме.

Таким образом, при выполнении заявленной НИР и работ по гранту были проведены исследования и определены способы публикации сверхбольших объемов растровых данных, в том числе обзорных изображений космических снимков низкого пространственного разрешения Landsat-5/TM, Landsat-7/ETM, ENVISAT/SAR, ERS-2/SAR в количестве более 40 тыс. кадров и более 20 тыс. кадров TERRA/MODIS и AQUA/MODIS. Также была опубликована спутниковая подложка интерактивной карты каталога разрешением 28,5 м на всю поверхность суши Земли общим объемом генерализованных данных 1,8 Тб. Все опубликованные данные космических аппаратов были внесены в метакаталог. При тестировании системы при полной нагрузке было получено, что время отклика при обращении до 10 пользователей без использования кэша не превышает 5 с, а с использованием кэша — 1 с, что позволяет говорить об успешном внедрении системы в опытную эксплуатацию.

Работа выполнена при поддержке гранта № 11 – 07 – 12058 РФФИ «Создание ГРИД-сервисов хранения, обработки и визуализации данных ДЗЗ для мониторинга добычи углеводородов».

Список литературы

1. *Кошкарев А.В.* Аналитический взгляд на GSDI-11// Пространственные данные. 2010. № 1. С. 6—12.

2. *Концепция* создания и развития инфраструктуры пространственных данных Российской Федерации // Там же. 2006. № 3. С. 6—9.

3. *Брыксин В.М.* Система интерактивного поиска и копирования космоснимков в локальной сети Intranet (BaseImages) // Свидетельство о гос. регистрации программы на ЭВМ № 2009611711. Федеральная служба по интеллектуальной собственности, патентам и товарным знакам, 31 марта 2009 г.

4. *Брыксин В.М.* База географически распределенных данных дистанционного зондирования Земли (ДЗЗ). Свидетельство о гос. регистрации базы данных № 2009620133. Федеральная служба по интеллектуальной собственности, патентам и товарным знакам, 31 марта 2009 г.

5. *Брыксин В.М., Евтюшкин А.В., Филатов А.В.* Технология создания комплексного банка космических снимков Земли // Известия Алтайского государственного университета. 2011. № 1—1 (69). С.55—59.

6. *Назаров И.В.* Каталог метаданных для распределенного банка пространственных данных / Алсынбаев К.С., Брыксин В.М., Козлов А.В. // X International Conference on Geoinformatics. Theoretical and Applied Aspects. Kiev, Ukraine, 10-13 May 2011. URL: http://earthdoc.eage.org/publication/publicationdetails/?publication=51587.

7. *Алсынбаев К.С., Брыксин В.М., Евтюшкин А.В., Козлов А.В. и др.* База метаописаний пространственных данных Ханты-Мансийского автономного округа — Югры (МПД). Свидетельство о гос. регистрации базы данных № 2012620171. Федеральная служба по интеллектуальной собственности, патентам и товарным знакам. 8.02.2012.

8. *GDAL* — Geospatial Data Abstraction Library [Офиц. сайт]. URL: http://www.gdal.org/ (дата обращения: 30.01.2013).

9. *GeoServer* [Офиц. сайт]. URL: http://geoserver.org/display/GEOS/Welcome (дата обращения: 30.01.2013).

Об авторах

Виталий Михайлович Брыксин — канд. техн. наук, вед. науч. сотр. Балтийский федеральный университет им. И. Канта, Калининград.

E-mail: VBryksin@kantiana.ru

Антон Владимирович Козлов — зав. лаб., Балтийский федеральный университет им. И. Канта, Калининград.

E-mail: AnKozlov@kantiana.ru

About authors

Dr Vitaliy Bryksin — PhD, senior research fellow, I. Kant Baltic Federal University, Kaliningrad.

E-mail: VBryksin@kantiana.ru

Anton Kozlov — head of the laboratory, I. Kant Baltic Federal University, Kaliningrad.

E-mail: AnKozlov@kantiana.ru