Современные проблемы дистанционного зондирования Земли из космоса.
2012. Т.9. №3. С. 324–331

Инфраструктура приёма, распределённой обработки и поставки спутниковых данных Центре коллективного пользования Регионального спутникового мониторинга ДВО РАН

И.В. Недолужко [1], П.В. Бабяк [1], Г.В. Тарасов [1], В.С. Ерёменко [2]

[1] Институт автоматики и процессов управления ДВО РАН,
690041, Владивосток, Радио, 5,
e-mails: paulb@dvo.ru; ilya@dvo.ru; george@dvo.ru
[2] Дальневосточный федеральный университет
690060, Владивосток, Октябрьская, 27,
e-mail: vitaer@gmail.com

Работа посвящена проблемам развития инфраструктуры, реализующей концепцию заказа на обработку, в Центре коллективного пользования (ЦКП) регионального спутникового мониторинга окружающей среды ДВО РАН. Особое внимание уделяется соответствию рекомендаций и моделей, используемых Европейским космическим агенством (ESA).

Ключевые слова: спутниковые данные, инфраструктура, распределённая и параллельная обработка, архивы спутниковых данных, каталоги, OAIS, SSE, HMA, GRID, ESA

Введение

В настоящий момент спутниковые данные широко применяются в различных отраслях исследовательской и хозяйственной деятельности человека. Эффективность решения прикладных и теоретических задач напрямую зависит от развитости средств доступа потребителей к продуктам обработки спутниковых данных. Современной тенденцией развития таких средств является создание гетерогенных распределённых информационных систем с целью обеспечения взаимовыгодного взаимодействия поставщиков и потребителей различных видов данных и услуг. Особенную важность создание таких систем приобретает на межведомственном и международном уровне. Вопросы обеспечения интероперабельности и разграничения доступа в этом случае выходят на первый план.

Центр коллективного пользования Регионального спутникового мониторинга окружающей среды Дальневосточного отделения Российской академии наук (ДВО РАН) создан на базе лаборатории спутникового мониторинга Институт автоматики и процессов управления (ИАПУ ДВО РАН). Круглосуточно работают 4-антенный комплекс приёма спутниковой информации и Суперкомпьютерный вычислительный центр производительностью 16 Тфлопс. Созданы средства и методы автоматического приема, накопления, распределенной обработки и поставки через интернет базовых видов информации, принимаемой со спутников NOAA, FY-1D, MTSAT-1R, Aqua/Terra и «Метеор-М» № 1. Реализованы автоматические цепочки обработки данных спутников серии NOAA (температурные и структурные карты поверхности моря, профили температуры и влажности атмосферы), Aqua, Terra (около 200 параметров морской воды и атмосферы), MTSAT-1R (температура воды, облачности, мониторинг морского льда) на основе пакетов AAPP, SeaDAS, RTTOV, MetOffice-1Dvar и собственных программных разработок (Шокин и др., 2008; Левин и др., 2010).

Для предоставления пользователям услуг по поставке и обработке спутниковых данных Центр коллективного пользования регионального спутникового мониторинга окружающей среды ДВО РАН ориентируется (Алексанин и др., 2007) на применение средств, развиваемых в рамках европейских инициатив по созданию глобальных инфраструктур спутниковых данных. Целью данной работы является рассмотрение основных проблем реализации инфраструктуры Центра в соответствии с моделями, рекомендациями и стандартами, применяемых в Европе для решения подобных задач на международном уровне.

Инфраструктура Центра коллективного пользования

Сформулированные ранее требования к инфраструктуре Центра (Бабяк и др., 2011) основаны на сложившейся в Центре практике работы с пользователями (Alexanin et al., 2002). Специфика такой работы заключается в преобладании исследовательских задач, где традиционный подход с поставкой исключительно стандартной продукции не способен полностью покрыть запросы пользователя. Основным подходом к решению проблемы является концепция заказа на обработку, в рамках которой пользователь имеет возможность управления процессом получения интересующей его продукции. В то же время, наличия каких-либо специальных знаний в области информационных технологий от него не требуется.

При формировании инфраструктуры Центра принимается во внимание развитие глобальных европейских инфраструктур данных дистанционного зондирования Земли (ДЗЗ). Для Центра основным ориентиром является среда SSE (Service Support Environment) Европейского космического агентства (ESA), и развиваемый на её основе проект HMA (Heterogeneous Mission Accessibility). Важным для Центра является не только реализации необходимых интерфейсов для интеграции Центра в HMA; но и следование моделям и методикам построения распределённых систем, которые определены в рамках этих и аналогичных проектов.

Основной стратегией реализации инфраструктуры Центра, позволяющей обеспечить перечисленные функции, является применение существующих решений и минимизация трудозатрат на каждом этапе её развития. Существующие решения могут включать как распространённые программные пакеты, так и собственные разработки Центра. При этом каждый компонент имеет чётко обозначенную функцию и может быть заменён либо продублирован другим, в зависимости от решаемой задачи. Основу инфраструктуры составляют следующие компоненты (рис. 1):

- репозиторий спутниковых данных как уровень абстракции поверх файлового архива и каталогов спутниковых данных; должен обеспечивать интерфейсы для усвоения данных и метаданных, управления, поиска и публикации;
- гетерогенная распределённая система обработки, имеющая возможность управления различными пакетами обработки спутниковых данных и связь с внешними вычислительными ресурсами (Дальневосточный вычислительный ресурс (ДВВР) ИАПУ ДВО РАН).
- система заказов на обработку спутниковых данных, предоставляющая пользователю возможность простого задания параметров обработки и получения результата в необходимом ему виде;
- система разграничения прав доступа ко всем сервисам и ресурсам Центра;
- веб-интерфейс для доступа пользователей к каталогам и системе заказов; реализуется как раздел глобального портала среды SSE Европейского агентства, так и в виде отдельного (местного) портала.

Репозиторий спутниковых данных

Основной функцией репозитория спутниковых данных Центра является как краткосрочное, так и долговременное хранение спутниковых данных различных уровней обработки. Краткосрочный (оперативный) архив, в отличие от долговременного, содержит как данные за последние месяцы, так и стандартную продукцию за этот период. При долговременном хранении предпочтение отдаётся данным более низких уровней обработки. Для эффективного управления данными архивами в Центре разрабатывается набор информационно-поисковых служб, обеспечивающих возможность извлечения наборов данных, соответствующих набору предъявляемых требований. Такой доступ может быть востребован как в автоматическом режиме,

так и в режиме диалога; как внутри Центра, так и извне. Основные требования к такому репозиторию спутниковых данных Центра рассмотрены ранее в публикации (Бабяк и др., 2011).

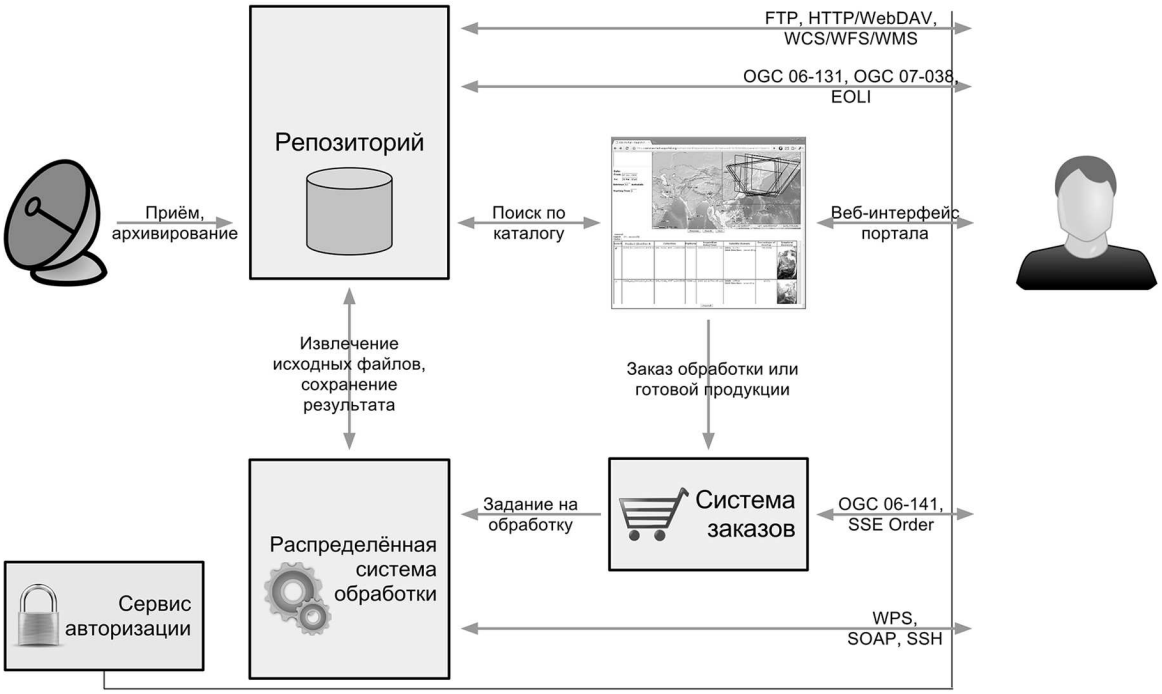

Рис. 1. Инфраструктура Центра коллективного пользования: внешние интерфейсы и взаимодействие базовых компонентов

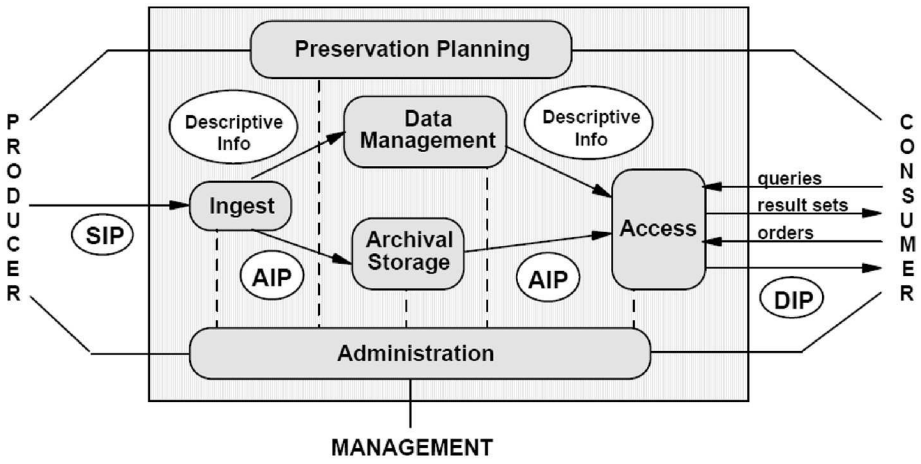

Рис. 2. Основные компоненты архива, реализующего модель OAIS

Наиболее общей моделью, описывающей различные аспекты создания и поддержки архивов данных различного типа, является OAIS (Open Archival Information System) (http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=24683). Модель подразумевает организацию системы (рис. 2), включающей в себя как автоматические компьютеризированные системы, так и обслуживающий персонал. Хранение данных может осуществляться как в цифровой форме, так и в физической.

Согласно стандарту OAIS, информация должна предоставляться пользователям в том виде, в котором он мог бы использовать её самостоятельно, без помощи экспертов (independently understandable information). В рамках предметной области ДЗЗ, зачастую происходит архивирование «сырых» (raw) данных; а получение необходимых пользователю продуктов происходит за счёт дополнительной обработки. Сырые данные никоим образом не могут рассматриваться как «самостоятельно понимаемая информация». Таким образом, архиви-

руемые данные и средства их обработки также должны рассматриваться при применении данной модели в области ДЗЗ. Необходимо также учитывать возможную эволюцию OAIS-системы: компромисс между архивированием продуктов и обработкой хранимых данных с целью получения этих продуктов может давать различные результаты в связи с совершенствованием технологий и расширением возможностей пользователей.

Рост объёмов хранимых данных ДЗЗ в Европе и необходимость в их эффективном совместном использовании различными агентствами и организациями, потребовали разработки общего подхода организации долговременных архивов. В связи с этим получил развитие проект LTDP (Long Term Data Preservation), использующий модель OAIS в качестве основы (http://earth.esa.int/gscb/). Модифицированная модель OAIS, рассматриваемая в руководстве LDTP, является эталоном для развития репозитория спутниковых данных Центра по двум причинам. Первой является опыт успешного применения ряда технологий и инструментов для реализации компонентов инфраструктуры Центра, а также факт регистрации ряда сервисов Центра на портале среды SSE (Services Support Environment) Европейского космического агентства. Второй, и наиболее важной причиной, является факт соответствия основных представлений о принципах работы с архивами спутниковых данных в Центре и в рамках рассматриваемых европейских инициатив; и хорошая согласованность с развиваемой в Центре концепцией заказа на обработку. В частности, поднимается вопрос о хранении данных возможно более низкого уровня обработки и получения требуемого продукта с применением актуального в момент обращения алгоритма обработки спутниковых данных.

Руководство LTDP определяет три уровня соответствия согласно критериям безопасности, целостности, доступа и интероперабельности. Соответствие минимальному уровню (Level A) требует от Центра решения следующих задач:

- *Регулярное обновление хранимой продукции в соответствии усовершенствованием технологий её получения*. Поскольку время хранения продукции в оперативном архиве Центра исчисляется единицами месяцев, а в долговременном хранятся только «сырые» данные, требование выполняется автоматически.
- *Периодическое обновление компонентов платформы хранения*. Основной сервер хранилища работает под управлением Gentoo Linux и получает периодическое обновление используемых пакетов. Аппаратная часть модернизируется ежегодно за счёт введения в строй дополнительных дисковых массивов, подключаемых через интерфейсы SCSI и SaS. Рекомендуемая частота обновления составляет 5…6 месяцев, поэтому данное требование можно считать частично соблюдаемым.
- *Содержание оборудования в условиях*, *рекомендованных производителем*. Требование выполняется.
- *Контролируемый доступ к программной и аппаратной части для исключения возможности постороннего вмешательства*. Требование выполняется.
- *Готовность к воздействию внешних факторов* (*стихийные бедствия*). Требование выполняется частично (стандартные меры пожарной безопасности).
- *Хранение копий архивных данных*. Дублирование выполняется за счёт резервного копирования на носители DVD. Для соответствия требованию необходимо организовать хранение в отдельном здании.
- *Разграничение прав доступа для персонала*. Требование выполняется.
- *Усвоение данных системой архивирования согласно соответствующим стандартам*. Анализ соответствия подобным стандартам (например, ISO 20652) в Центре пока не проводился.
- *Генерация метаданных в соответствующих стандартах* (*OGC 06-080 GML*). Требование выполняется (Недолужко, Коробкова, 2012).
- *Автоматическая проверка данных до помещения в архив*. Проверка производится, однако вопрос соответствия стандартам не исследовался.

- *Проверка корректности сохранения данных на носителе*. Требование выполняется.
- *Доступность данных для поиска и заказа*, *генерация конечного продукта для предоставления пользователю*. Требование выполняется на основе исторически сложившихся решений и средств, используемых в SSE/HMA.
- *Информация об условиях предоставления данных пользователям*. Выполняется частично на основе исторически сложившихся решений.
- *Должно гарантироваться качество данных и продуктов в течение всего срока спутниковой миссии*. Центр следует рекомендациям организаций, ответственных за миссию, и не занимается поставкой данных, не соответствующих их требованиям.

Как следует из анализа, в настоящее время репозиторий спутниковых данных Центра удовлетворяет более чем половине требований уровня А. При этом выполняется также и ряд требований более высоких уровней. В частности, присутствует реализация интерфейса каталога OGC 06-131 (Недолужко, Коробкова, 2012). Ведущиеся в Центре работы позволят обеспечить соответствие большей части требований уровня А в ближайшем обозримом будущем.

Распределённая обработка данных

Основной задачей распределённой системы обработки является интеграция в единую гетерогенную систему различных программных комплексов обработки данных — как разработанных в рамках Центра, так и пакетов сторонних производителей — для обеспечения оперативной обработки спутниковых данных и обработки по заказу.

Система обработки данных построена на распределённых вычислениях, где каждый вычислительный ресурс (в том числе задействованные внешние вычислительные ресурсы) является независимым вычислительным узлом, связанным с другими узлами посредством сети. Все задачи, запускаемые в рамках системы, являются независимыми асинхронными процессами. Схема является формальным описанием того как и на каких обрабатывающих машинах должна запускаться та или иная задача. Вся логика построения цепочек алгоритмов обработки данных строится на основе событий, обрабатываемых триггерами. Срабатывание триггера определяется событием, к которому привязан данный триггер. Как правило, это появление новых данных или окончание работы определённой схемы. При срабатывании триггер осуществляет запуск установленных схем с определёнными параметрами. При запуске схемы она получает уникальный идентификатор, который в дальнейшем используется для контроля процесса выполнения данной задачи (рис. 3).

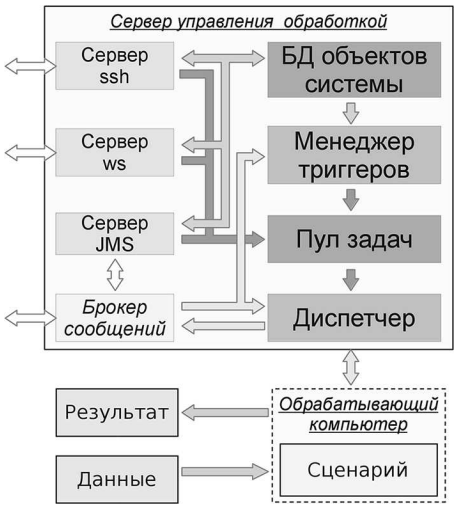

Рис. 3. Структурная схема управления системой обработки

Запуск и контроль заданий на вычислительных узлах осуществляет посредством протокола SSH2, что обеспечивает простоту взаимодействия и высокий уровень безопасности устанавливаемых соединений. Однако в данный момент в рамках системы обработки не формализована передача данных, но, как правило, она осуществляется посредством стандартных протоколов типа SMB, FTP или SFTP.

Такая организация распределенной системы позволяет достичь нескольких целей:

- полностью автоматический запуск необходимых задач с возможностью их ручного запуска;
- оперативность обработки поступающих данных;
- автоматическое распределение вычислительных ресурсов между задачами;
- устойчивость к выходу из строя тех или иных обрабатывающих машин;
- все поставленные в очередь задачи в конечном итоге будут обработаны в порядке их поступления.

Внешний интерфейс доступа к системе обработки реализуется через протоколы SSH, SOAP и WPS (Web Processing Service). SSH доступ применяется в первую очередь при интерактивной работе либо для доступа из интерпретатора команд операционной системы (shell-скриптов), а протоколы SOAP и WPS будут использованы для програмного доступа к системе обработки.

Система заказов на обработку спутниковых данных является уровнем абстракции над системой обработки, обеспечивающим реализацию специализированных низкоуровневых интерфейсов. При реализации этих интерфейсов в Центр ориентируется на стандарт заказа среды SSE (http://services.eoportal.org/massRef/documentation/icd.pdf), а также более новый стандарт проекта HMA, предлагаемый в качестве стандарта OGC 06-141 (http://wiki.services.eoportal.org/tiki-download_wiki_attachment.php?attId=168&page=HMA-E%20Baseline). Интерфейс заказа допускает задание дополнительных параметров, которые могут быть использованы для выбора способа и параметров обработки. Сервис заказа может быть интегрирован с каталогом продуктов, что позволяет пользователю осуществлять заказ обработки данных, выбранных по каталогу. В настоящий момент в Центре реализованы прототипы обоих интерфейсов, ведутся работы по организации их взаимодействия с распределённой системой обработки спутниковых данных Центра.

Увеличивающийся поток заданий на обработку требует доступа к дополнительным вычислительным мощностям. Три года назад Центр начал развивать направление внедрения GRID-технологий для получения доступа к потенциально неограниченному объему сторонних вычислительных ресурсов. Первые работы (Бабяк, Тарасов, 2009) по интеграции вычислительных мощностей были проведены совместно с ЦКП ДВВР, работа которого поддерживается сотрудниками ИАПУ ДВО РАН. В рамках проекта по интеграции вычислительных ресурсов ДВВР в систему обработки данных Центра спутникового мониторинга была разработана и внедрена следующая схема взаимодействия узлов объединенной GRID-сети (рис. 4).

Созданная к настоящему моменту GRID-сеть состоит из шести основных компонент, две из которых непосредственно связаны через GRID-сервисы (веб-сервисы: GRAM: Grid Resource Allocation and Management и RFT: Reliable File Transfer). Центральным звеном сети является два узла: один узел размещен на стороне Центр приема и обработки, второй узел расположен на стороне вычислительного центра и имеет прямой доступ к внутренним ресурсам вычислительного кластера.

Схема прохождения задания через GRID состоит из девяти основных шагов. Шаги 1–3 образуют стадию формирования задания, результатом которой является запрос от системы обработки к стороннему ресурсу на выполнение вычислений. В данной GRID-сети систему обработки представляет некоторый выделенный сервер, на котором установлено соответствующее промежуточное программное обеспечение поддержки работы GRID-сервисов

(или веб-сервисов). Основу запроса составляет *паспорт задания*, согласно которому происходит вся обработка. В паспорте задается адрес ресурса, на котором следует выполнять задание, скрипт задания, параметры параллельной обработки (количество требуемых процессоров, временны*е* ограничения), дополнительные параметры со ссылками на необходимые файлы с данными и результатами обработки. Подготовка паспорта задания выполняется системой обработки автоматически. Шаги 4–6 образуют стадию выполнения паспорта задания. На данной стадии происходит копирование входных файлов данных с использованием сервиса RFT на сторону вычислительного ресурса, последующая автоматическая генерация паспорта задания уже для локальной системы управления вычислителя и запуск этого паспорта на выполнение. Шаги 7–9 образуют завершающую стадию, на которой выполняется периодический опрос состояния выполнения задания, и по его окончанию происходит обратное копирование результатов расчетов от вычислителя до системы обработки.

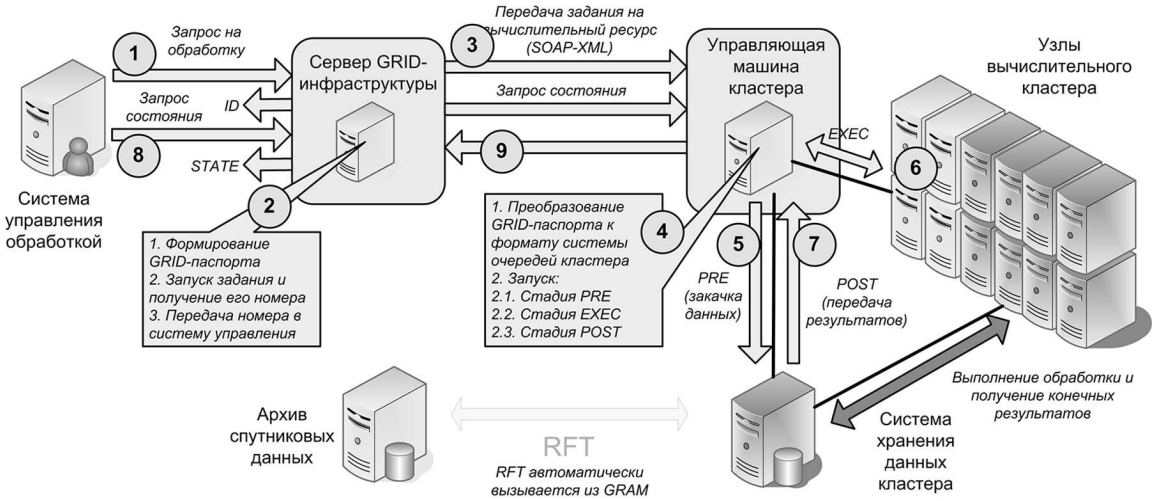

Рис. 4. Схема интеграции сторонних вычислительных ресурсов в распределенную систему обработки спутниковых данных

Работоспособность данной схемы прохождения заданий и всей GRID-сети в целом неоднократно тестировалась при разных режимах обработки и на разных приложениях (расчет полей доминантных ориентаций термических структур, построение профилей температуры и влажности атмосферы и другие задачи).

Заключение

Ведущиеся в Центре работы по развитию инфраструктуры, реализующей концепцию заказа на обработку спутниковых данных, позволят обеспечить потребности пользователя в полном объёме. При проектировании делается акцент на соответствие применяемых стандартов, методик и архитектурных решений требованиям и рекомендациям, выработанным в Европе при создании глобальных гетерогенных инфраструктур, таких как SSE и HMA (разрабатывающихся под руководством Европейского космического агентства). Полученный опыт, включая опыт интеграции создаваемых в Центре сервисов в SSE/HMA, может представлять особый интерес в свете вопроса по созданию единой межведомственной инфраструктуры пространственных данных в России.

Работа выполнена при поддержке РФФИ (проекты № 11-01-12107-офи-м-2011, 11-07-00511-а) и интеграционного гранта ДВО РАН 12-II-СО-01И-004; реализована на оборудовании Центра коллективного пользования регионального спутникового мониторинга окружающей среды ДВО РАН при финансовой поддержке Минобрнауки России.

Литература

1. *Алексанин А.И.*, *Алексанина М.Г.*, *Бабяк П.В.*, *Недолужко И.В*. Организация информационного обеспечения и телекоммуникационные технологии в спутниковом центре ДВО РАН // Тр. 10-й Санкт-Петербургской Междунар. конф. «Региональная информатика — 2006». СПб.: СПОИСУ, 2007. С. 329–333.
2. *Бабяк П.В.*, *Тарасов Г.В*. Опыт использования Grid-Технологий в системе обработки данных Спутникового центра ДВО РАН // Современные проблемы дистанционного зондирования Земли из космоса. 2009. Т. 6. № 1. С. 71–80.
3. *Бабяк П.В., Недолужко И.В., Фомин Е.В.* Подход к предоставлению услуг по обработке спутниковых данных в Центре коллективного пользования регионального спутникового мониторинга окружающей среды ДВО РАН // Материалы XIV Всероссийской объединённой конференции “Интернет и современное общество” (IMS-2011). СПб.: МПСС, 2011. С. 27-32.
4. *Левин В.А.*, *Алексанин А.И.*, *Алексанина М.Г.*, *Дьяков С.Е.*, *Недолужко И.В.*, *Фомин Е.В.*. Разработка технологий спутникового мониторинга окружающей среды по данным метеорологических спутников // Открытое образование. 2010. № 5. С. 41–49.
5. *Недолужко И.В.*, *Коробкова О.О*. Средства интеграции каталогов в современных европейских инфраструктурах данных ДЗЗ // Рос. науч. электрон. журн. «Электронные библиотеки». 2012. № 3
6. *Шокин Ю.И.*, *Пестунов И.А.*, *Смирнов В.В.*, *Синявский Ю.Н.*, *Скачкова А.П*, *Дубров И.С.*, *Левин В.А.*, *Алексанин А.И.*, *Алексанина М.Г.*, *Бабяк П.В.*, *Громов А.В.*, *Недолужко И.В*. Распределенная система сбора, хранения и обработки данных для мониторинга территорий Сибири и Дальнего Востока // Журн. Сибирского федерального ун-та «Техника и технологии». 2008. Т. 1. № 4. С. 291–314.
7. *Alexanin A.I.*, *Babyak P.V.*, *Herbeck F.E.*, *Levin V.A*. Satellite information support of scientic researches and economic applications // Proc. Science and Technical information STI-2002. M.: VINITI, 2002. P. 17–18.

Infrastructure for receive, distributed processing and delivery of satellite data at Multiple Access Center of Regional Satellite Monitoring FEB RAS

I.V. Nedoluzhko [1], P.V. Babyak [1], G.V. Tarasov [1], V.S. Eremenko [2]

[1] Institute of Automation and Control Processes FEB RAS,
690041, Vladivostok, Radio, 5,.
e-mails: paulb@dvo.ru; ilya@dvo.ru; george@dvo.ru
[2] Far Eastern Federal University,
690060, Vladivostok, Oktyabrskaya,
e-mail: vitaer@gmail.com

The paper is devoted to problems of development of infrastructure that implements «order-for-processing» concept at the Multiple Access Center of Regional Satellite Monitoring FEB RAS. Strong accent is put on adequacy with recommendations and models used by European Space Agency (ESA).

Keywords: satellite data, infrastructure, distributed and parallel processing, satellite data archives, cataloguesm OAIS, SSE, HMA, GRID, ESA.