

Инфраструктура научных информационных ресурсов и систем

Кудашев Е.Б., Серебряков В.А.

Аннотация

Статья посвящена анализу работы Третьего Всероссийского Симпозиума «Инфраструктура научных информационных ресурсов и систем», состоявшегося в г. Сухум, Абхазия, 5-8 октября 2013 г. Лавинообразный рост объема электронного контента потребовал разработки новых подходов к хранению и непрерывному доступу к цифровым научным данным. Особый интерес вызывают актуальные научные задачи по созданию инфраструктур пространственных данных. На Симпозиуме традиционно обсуждаются вопросы, связанные с интеграцией геоинформационных ресурсов и свободного доступа к ним, исследования e-Infrastructures с целью формирования распределенных научных информационных ресурсов, развития взаимосвязанных каталогов и создания сети интегрированных интероперабельных баз данных. Развитие e-Science Infrastructures должно стать основой формирующихся систем коллективной работы исследователей на основе виртуального объединения информационных и вычислительных ресурсов. Основным направлением работы Третьего Симпозиума были вопросы применения современных подходов в технологии развития информационных систем к задачам информационной поддержки научных исследований.

Ключевые слова: цифровой контент, научные данные, формирование цифровой инфраструктуры, непрерывный доступ и длительное хранение данных, третий Симпозиум.

Всероссийский Симпозиум «ИНФРАСТРУКТУРА НАУЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ И СИСТЕМ» проводится ежегодно РАН и ВЦ РАН. Третий Симпозиум состоялся в Абхазии, г. Сухум при поддержке Академии Наук Абхазии с 05 октября по 09 октября 2013 г. В связи с лавинообразным ростом объема цифровых научных данных подходы к хранению и доступу к этим данным вызывают особый интерес. Актуальны научные задачи по созданию инфраструктур пространственных данных. На Симпозиуме традиционно обсуждаются исследования e-Infrastructures с целью формирования распределенных научных информационных ресурсов, развития взаимосвязанных каталогов и создания сети интегрированных интероперабельных баз данных. Развитие e-Science Infrastructures должно стать основой формирующихся систем коллективной работы исследователей на основе виртуального объединения информационных и вычислительных ресурсов.

Первый и Второй Всероссийские Симпозиумы «Инфраструктура научных информационных ресурсов и систем», проведенные в 2011 и 2012 гг., показали, что в ведущих научных центрах России накоплен большой опыт использования

геоинформационных технологий, реализованы многочисленные геоинформационные проекты, созданы информационные системы научных ресурсов. На Симпозиуме были заслушаны доклады из различных регионов России - это Тихоокеанский Институт географии ДВО РАН, Вычислительный Центр им. А.А. Дородницына РАН, Институт космических исследований РАН, Объединенный Институт высоких температур РАН, Институт вычислительных технологий СО РАН, Мировой Центр данных - Всероссийский НИИ гидрометеорологической информации, Институт оптики атмосферы СО РАНРАН, Институт мониторинга климатических и экологических систем СО РАН, Институт астрономии РАН, Специальная Астрофизическая обсерватория РАН (Архыз), ГАИШ МГУ, Географический факультет МГУ и другие ведущие научные центры России.

Основным направлением работы Третьего Симпозиума были вопросы применения современных подходов в технологии развития информационных систем к задачам информационной поддержки научных исследований.

Обзор основных проблем, возникающих в этой связи, и направлений в их решении был сделан в докладе **О.С. Бартунова** (ГАИШ МГУ) «Научные вызовы технологиям БД». Автор отметил, что в науке происходит процесс лавинного поступления информации, в первую очередь связанный с успехами в технологии создания приемных устройств. В современных крупных экспериментах, включая и численное моделирование, анализ терабайтов и даже петабайтов научных данных становится повседневной задачей. Из-за недостатков существующих СУБД, большинство научных проектов, в которых встает задача анализа больших объемов данных, осуществляют обработку и анализ исходных данных вне системы управления базами данных, что затрудняет решение вышеобозначенных задач. Современные суперкомпьютеры-вычислители не рассчитаны на работу с большими данными в режиме реального времени, они более ориентированы на пакетный режим выполнения задач, требующих в основном вычислительных мощностей.

Новые СУБД для науки должны использовать возможности архитектуры масштабируемых хранилищ и распределенных вычислителей для обеспечения полноценной поддержки полного цикла работы с данными - хранение, обработка, анализ и обмен данными. Обмен данными подразумевает не только физический доступ, но и сохранение цифрового контента, работа с разнородными и распределенными данными, поиск информации о данных, доступность программного обеспечения, знание, как именно анализировать эти данные.

Примером применения подходов к решению задач, требующих больших данных, является работа, представленная в докладе **П.В. Кайгородова, Д.А. Ковалевой, О.Б. Длужневской, О.Ю. Малкова** «Архитектура Базы данных двойных звезд BDB». Представлено описание архитектуры базы данных двойных звезд (Binary star DataBase, BDB), разрабатываемой в Институте астрономии РАН. Целью создания BDB является объединение информации из множества разнородных каталогов двойных и кратных звезд, а также разработка удобного инструмента для работы с данными каталогов. Рассматриваются основные проблемы, возникшие в процессе реализации BDB, методики извлечения информации из исходных каталогов, а также методы кросс-идентификации объектов. BDB реализована на базе фреймворка Nagare (stackless Python/SQLAlchemy/Elixir) и СУБД

Postgresql, ее beta-версия доступна по адресу <http://bdb.inasan.ru>.

Еще одним важным и интенсивно развивающимся направлением информационной поддержки научных исследований являются электронные библиотеки, о чем говорилось в докладе **Федотова А.М., Жижимова О.Л., Пестунова И.А., Федотовой О.А.** «Электронная библиотека – путь к интеграции разнородных данных». Исходя из целей ЭБ и анализа существующих систем, направленных на поддержку научных исследований, сформулированы следующие функциональные требования к модели ЭБ по научному наследию: надежное долговременное и защищенное от исчезновения хранение информации; актуальность, полнота, достоверность происхождения документов; историчность информации; географическая привязка информации; наличие большого числа словарей-классификаторов (справочников), для обеспечения идентификации и классификации ресурсов; поддержка неоднородных и слабо структурированных информационных ресурсов; поддержка взаимосвязей информационных ресурсов; предоставление информации пользователю в виде, выбранном пользователем; наличие интеллектуальных служб обслуживания запросов пользователя; наличие программных интерфейсов для поддержки аналитической работы пользователя с помощью программных приложений; поддержка требований интероперабельности как на программном, так и на семантическом уровне; поддержка работы с внешними источниками. Рассмотренная модель информационной системы, работающей с материалами научного наследия, реализована на примерах: Научная школа А. А. Ляпунова – основателя теоретического программирования и российской кибернетики, ЭБ по моделям динамики биосферы, Сервис-ориентированная информационная система, обеспечивающая доступ к каталогам спутниковых данных, а также учебные пособия по курсам «Современные проблемы информатики и вычислительной техники», «Вычислительные системы», «Информатика» и «Экология».

Важным направлением развития современных подходов к информационным системам в науке является разработка онтологий различных предметных областей научных исследований и создание на их основе информационных систем. Пример такого рода приведен в докладе **О.М. Атаевой, А.О. Еркимбаева, В.Ю. Зицермана, Г.А. Кобзева, В.А. Серебрякова, К.Б. Теймуразова** «Представление данных по теплофизическим свойствам веществ с использованием концепций и методов Semantic WEB». Понятийный багаж таких наук как физика, химия, материаловедение и др. уже исходно достаточно формализован, чтобы лечь в основу онтологического описания. Тем самым открывается принципиальная возможность интеграции научных данных за счет спецификации содержимого разнородных источников. Среди других возможностей Semantic WEB - автоматизация обработки текстов, высокоточный информационный поиск, средства логического вывода и проч. Несмотря на множество примеров успешного применения этой концепции, вся технология и практика интеграции данных в ее рамках находятся еще в стадии становления. В этих условиях представляется особенно полезным отладить ее применение в относительно узкой предметной области, для которой характерны определенные типы данных и сложившиеся стандарты их представления в БД и печатных источниках. С учетом собственного опыта и круга интересов в данном исследовании авторы выбрали в качестве конкретной области теплофизические

свойства веществ и материалов.

Доклад А.М. Елизарова, Е.К. Липачёва был посвящен вопросам разработки сервисов структурирования математического контента и интеграция электронных математических коллекций в научное информационное пространство. В докладе обсуждаются вопросы интеграции в научное информационное пространство данных из научных электронных коллекций, имеющих различные организационную структуру и форматы представления. Предложены модель интеграции и комплекс методов, обеспечивающих автоматическое преобразование контента с применением систем конвертации данных, объединения схем данных, стилевого приведения данных из разных коллекций в единое представление. Описаны методы автоматического выделения метаданных из научного контента и система генерации служебных метаданных.

Охарактеризованы методы структурирования математического контента и способы организации хранения и автоматизации процессов управления естественнонаучными коллекциями на основе технологий Семантического веба. Акцент сделан на особенностях обработки математических текстов. С помощью сервисов, созданных по предложенной методике, выполнено структурирование достаточно большого по объему электронного хранилища, содержащего выпуски периодического журнала по математике и многотомных трудов конференций.

Представлены также современные информационные системы, предназначенные для автоматизации полного цикла подготовки и издания электронных научных журналов. Показаны преимущества использования журнальных систем открытого доступа. Обоснован выбор системы Open Journal System (OJS) в качестве платформы построения электронного хранилища научных журналов Казанского федерального университета (КФУ). Представлен опыт реализации пилотных проектов КФУ, выполненных на базе OJS.

Важное место в работе симпозиума было отведено пространственным данным. Вопросу эффективной работы с пространственными данными в PostgreSQL был посвящен доклад **О. С. Бартунова, С. В. Карпова, А.Е. Короткова**. Рассматриваются вопросы использования СУБД PostgreSQL в решении астрономических задач. Астрономия оперирует огромными объемами данных, поток которых непрерывно увеличивается вследствие успехов технологии производства высокочувствительных в разных диапазонах электромагнитного спектра и доступных сенсоров большого размера. Спецификой астрономических данных является их привязка к координатам на небесной сфере, что усложняет поисковые алгоритмы. Следует отметить, что с математической точки зрения астрономические данные не отличаются от данных, с которыми работают геоинформационные системы. Выбор СУБД для работы с астрономическими данными, таким образом, определяется (помимо стандартных требований) способностью эффективной работы с пространственными данными. В рамках проекта Виртуальной Обсерватории в ГАИШ МГУ были проведены исследования по применимости свободной СУБД PostgreSQL для хранилища астрономических каталогов и обеспечения основных сервисов. В процессе работы была серьезно улучшена производительность существующих реализаций алгоритмов и получены рекомендации по их наиболее эффективному использованию в зависимости от

вида нагрузки. Проведенные результаты показали применимость PostgreSQL для хранения пространственных данных и работы с ними.

Серия докладов посвящена вопросам развертывания геопорталов организаций.

Разработке Веб-приложений на основе технологий, ресурсов и сервисов геопортала ИВМ СО РАН был посвящен доклад **О. Э. Якубайлика, А.А. Кадочникова, А.В. Токарева**. Исследования в области проектирования и разработки комплекса программно-технологических решений для построения прикладных геоинформационных веб-систем и сервисов (геопорталов) ведутся в ИВМ СО РАН уже около 10 лет. В настоящее время системообразующим инфраструктурным элементом проводимых исследований является геопортал ИВМ СО РАН, создание которого началось около трех лет назад, при выполнении работ по междисциплинарным интеграционным проектам и программам фундаментальных исследований СО РАН. Геопортал ИВМ СО РАН постоянно развивается, «обрастает» новыми функциональными возможностями, являясь своего рода «экспериментальной установкой», на которой создаются и тестируются новые технологические решения, программные модули, веб-сервисы. После опытной эксплуатации на портале новые разработанные элементы могут быть внедрены в прикладные системы, находящиеся в промышленной эксплуатации. Одной из основных подобных систем является «Банк пространственных данных Красноярского края», которая сегодня рассматривается Министерством информатизации и связи Красноярского края как основа при формировании региональной инфраструктуры пространственных данных.

В докладе О.Э. Якубайлика (Институт вычислительного моделирования СО РАН) «Геосервисы веб-системы «Экологический атлас Красноярска», приводятся данные о состоянии разработки атласа. В настоящее время атлас состоит из 39 тематических слоев, которые можно условно разделить на две части. В первой из них показаны параметры климата, плотность и фитопатологическое состояние растительного покрова в городе, ареалы загрязнений некоторыми, в том числе и канцерогенными, веществами почв, снежного покрова и атмосферы. В других слоях находится информация о населении города. Это плотность жилищно-промышленной застройки, динамика показателя заболеваемости злокачественными новообразованиями жителей Красноярска. В этих слоях обозначены области различной заболеваемости по районам города. Карты позволяют сопоставить динамику заболеваемости для различных возрастных групп и некоторых форм злокачественных новообразований, выделить особенности расположения таких областей. Рассматриваемая информационная система «Экологический атлас Красноярска» основана на технологиях, ресурсах и сервисах геопортала ИВМ СО РАН.

В докладе С.М. **Краснопеева, С.С. Пашинского, Е.В. Шулькина** (Тихоокеанский институт географии ДВО РАН) рассматривается реализация аналитических сервисов ИПД ТИГ. 5 лет назад в ТИГ ДВО РАН начались работы по развертыванию ключевых элементов ИПД. На настоящий момент реализуется функционал, обеспечивающий решение круга задач, которые можно отнести уже к категории «бизнес для бизнеса» (B2B):

- Запущена в тестовую эксплуатацию транзакционная Служба публикации пространственных объектов (WFS-T, Web Feature Service Transactional);
- Расширен функционал клиентского приложения Веб-службы доступа и визуализации, обеспечивающего поддержку доступа и визуализации векторных классов объектов, а также возможность удалённого создания/удаления/редактирования пространственных объектов;
- Развёрнута и запущена в тестовую эксплуатацию Веб-служба обработки пространственных данных (OGC WPS), предоставляющая доступ к процедурам GRASS GIS (218), библиотеки Sextante и др.;
- Разработано клиентское Веб-приложение, предоставляющее пользователю интерфейс формирования сценария и выполнения удалённой обработки ПД;
- Продолжается формирование корпоративного ресурса базовых пространственных данных, данных дистанционного зондирования Земли высокого разрешения.

Создание подобной инфраструктуры позволит обеспечить доступ к возможностям ГИС через Интернет, работу с актуальным динамическим контентом, реализовывать проекты с элементами «краудсорсинга», существенно расширить круг потребителей геопространственной информации и геопространственных услуг.

В докладе **Жижимова О. Л., Скачкова Д. М.** (Институт вычислительных технологий СО РАН) «Реализация географического поиска информации в "негеографических" информационных системах на платформе ZooSPACE обсуждаются алгоритмы географического поиска в "негеографических" информационных ресурсах и их реализация для платформы интеграции гетерогенных данных ZooSPACE.

Работа выполнена при финансовой поддержке РФФИ (грант № 14-07-00032).

Об авторах

Кудашев Е.Б. – академик Российской Инженерной Академии, д.т.н., ведущий научный сотрудник Института космических исследований Российской академии наук, профессор Механико-Математического Факультета МГУ им. М.В. Ломоносова, Москва, Российская Федерация. e-mail: fmkdshv@gmail.com

Серебряков Владимир Алексеевич – Вычислительный центр им. А.А. Дородницына РАН, Москва e-mail: serebr@ultimeta.ru