

СЕТЕВЫЕ СЕРВИСЫ ПО СОЛНЕЧНО-ЗЕМНОЙ ФИЗИКЕ, МЕТЕОРОЛОГИИ, ГЕОДИНАМИКЕ И СЕЙСМОЛОГИИ В СТАНДАРТЕ GRID ДЛЯ ПРОГРАММЫ РАН «ЭЛЕКТРОННАЯ ЗЕМЛЯ»

М. Н. Жижин¹, Е. П. Харин¹, А. В. Андреев¹, Э. Кин²

¹ИФЗ и ГЦ РАН (г.Москва)

²НЦГД НОАА (г.Москва)

Базы данных по солнечно-земной физике, метеорологии, геодинамике и сейсмологии представлены в Интернет набором веб-сервисов, позволяющих получить метаданные и данные в реальном времени для сетевых приложений и информационных порталов. Описаны основы интерактивной работы с веб-порталами SPIDR, IDEAS, GPS, SeismRep и пиринговая технология запросов к веб-сервисам из приложений клиента по протоколу SOAP.

Database on solar-terrestrial physics, meteorology, geodynamics and seismology are presented in Internet by a number of web-services, which allow to get metadata and data in real time for network applications and information portals. The basis of interactive work with web portals SPIDR, IDEAS, GPS, SeismRep and piring technology of the requests to web services from client-applications on protocol SOAP are described.

Введение

Цифровые геофизические данные могут быть представлены в виде: 1) электронных научных публикаций; 2) исходных текстов и исполняемых программ; 3) текстовых или бинарных файлов с наблюдаемыми или смоделированными данными; 4) электронных карт для ГИС; 5) баз данных; 6) цифровых изображений. Программа «Электронная Земля» предполагает создание оптимальной сетевой среды для коллективного хранения, описания, поиска и распределенного доступа к этим категориям данных. При этом подразумевается в первую очередь объединение имеющихся данных в мета-ресурс(ы), а не разработка отдельных вычислительных моделей, баз данных или слоев ГИС. В основу этой сетевой среды заложены передовые ИТ продукты, форматы и языки: GRID, XML и Java.

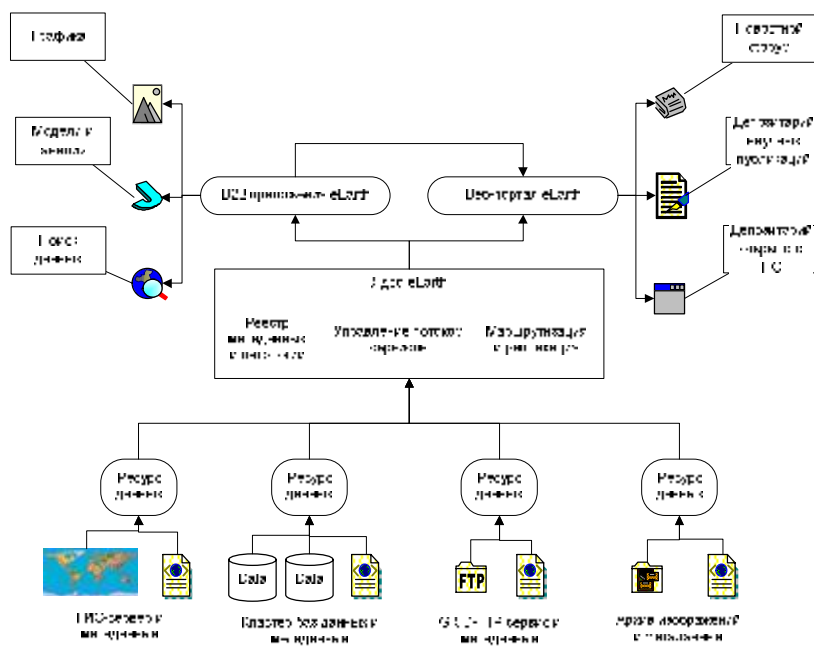
Очевидны трудности, стоящие перед программой: ограничения по времени и финансированию, недостаток кадров на стыке ИТ и геофизики, междисциплинарность предметной области, объем накопленных данных и число потенциальных участников, новизна самой постановки задачи и инструментов для ее решения. При наличии четкой концепции это – типичный набор граничных условий для open source проекта или серии пересекающихся подпроектов как, например, Jakarta, Globus, VisAD. Здесь под концепцией мы понимаем списки основных задач, которые будет решать система, готовых программных компонент и данных, из которых она будет строиться, и этапов реализации программы от прототипа до внедрения.

В части создания депозитариев электронных публикаций и программного обеспечения, новостных научных форумов и распределенных геоинформационных систем уже имеются открытые стандарты и программные реализации, которые можно применить в программе «Электронная Земля». Для нас наибольший интерес представляет та часть программы, где предполагается параллельный поиск, синтез (моделирование), визуализация и доставка пользователю карт, цифровых изображений, файлов и выборки из баз геофизических данных. Здесь цель программы можно будет считать достигнутой, когда по связям между этими данными будут строиться и проверяться геофизические теории без проведения новых измерений.

Архитектура системы

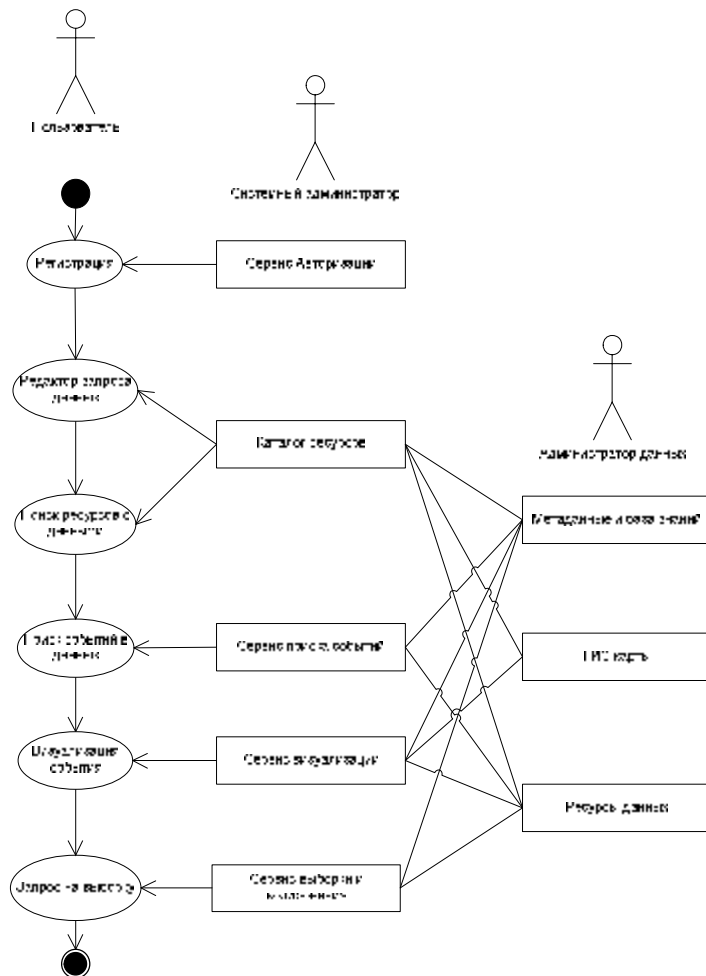
Опыт работы над подобными системами в международных проектах Space Physics Interactive Data Resource (SPIDR), Satellite Archive Browse and Retrieval (SABR), Investigation of Distributed Environmental Archive System, (IDEAS), Master Environmental Library (MEL) и Space Weather Reanalysis (SWR) позволяет нам предложить следующие «проекции» для описания концепции системы «Электронная Земля»: концептуальный интерфейс пользователя (CUI), уровень приложений (Application Layer), уровень ресурсов данных (Resource Layer), объектно-ориентированные модели данных (Data Model) и метаданных (Metadata). Основным языком программирования проекта предполагается Java.

Собственно задачи обнаружения, обработки данных, подготовки их для интерфейса пользователя и B2B взаимодействия с программами-«клиентами» решаются на уровне приложений, который представлен набором веб-сервисов и баз метаданных. Перечислим необходимые сервисы уровня приложений: каталог ресурсов данных (на базе UDDI и OWL-S), предметная онтология (общая терминология на базе OWL, прототип в ESG), картографический ГИС-сервис, сервис поиска событий в терминах нечеткой логики (на базе проектов ESG и SPIDR), онтология явлений и эффектов окружающей среды с проекцией на предметную онтологию и сервис поиска событий (на базе проекта ESG, сервис визуализации данных (на базе VisAD), сервис выборки, репликации и доставки данных в принятых научных форматах (NetCDF, HDF, и т.п. на базе GridFTP), сервис контроля доступа (на базе GRID), координационный сервис (на базе OWL-S). По мере развития системы сервисы могут быть расширены сервисами анализа данных (регрессионный, дисперсионный, кластерный, независимых компонент) и удаленного запуска приложений (параллельные программы расчета статистических и физических модели окружающей среды).



Обобщенная архитектура системы

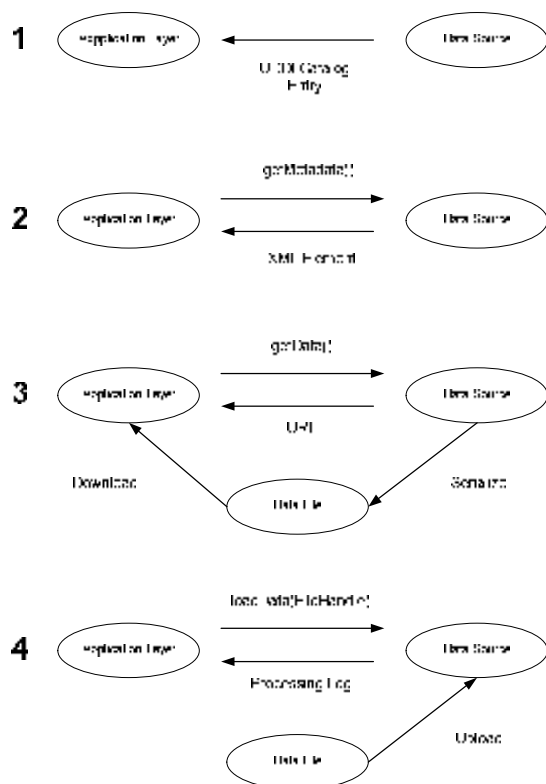
С помощью веб-интерфейса пользователя, построенного по технологии веб-портал,



можно будет решать интерактивные задачи обнаружения ресурсов (мета)данных, поиска событий окружающей среды, многомерной визуализации этих событий в пространстве и времени, выборки и доставки данных на компьютер пользователя. Условие интерактивности обусловлено темпами роста вычислительной мощности и пропускной способности сети, хотя в ряде случаев может потребоваться асинхронное выполнение запросов (напр., при работе с роботизированной библиотекой лент или при реанализе данных на суперкомпьютере). Рабочий поток в динамическом интерфейсе пользователя будет описываться конечным автоматом Jakarta Struts с возможной заменой на JavaServer Faces по мере утверждения стандарта.

Рабочий поток в динамическом интерфейсе пользователя

На уровне ресурсов данных решаются 4 основные задачи: регистрация в каталоге ресурсов, предоставление метаданных о ресурсе, выборка данных и пополнение ресурса. Ресурсы реализованы как *специализированные веб-сервисы* с простым расширяемым интерфейсом в контейнере Axis. При размещении ресурса данных в контейнере веб-сервисов он автоматически регистрируется в UDDI каталоге ресурсов уровня приложений. На запрос метаданных ресурс возвращает XML документ, динамически построенный в соответствии с общей схемой (XML Schema). Прототипы XML схемы реализованы в проектах IDEAS, SPIDR и SPDML. При запросе на выборку данных соответствующий метод веб-сервиса запрашивает данные из кластера баз данных, сериализует выборку в директорию, доступную на чтение по сети, и передает ссылку на данные клиенту (IDEAS, SPIDR). Частным случаем ресурса данных может быть файл с данными, который сопровождается XML документом с метаданными и зарегистрирован в каталоге ресурсов данных через реплицирующий сервис (SPDML).



Задачи уровня ресурсов данных: 1) регистрация в каталоге ресурсов; 2) предоставление метаданных о ресурсе; 3) выборка данных; 4) пополнение ресурса.

GRID технологии, основанные на веб-сервисах, сегодня находятся на этапе обсуждения стандартов. Прототип Globus 3.0 после полугодового тестирования в ближайшее время будет заменен на Web Services Resource Framework (WSRF) aka Globus 4.0. По мере утверждения стандарта наши сервисы уровня приложений и ресурсов данных будут приведены в соответствие со стандартом WSRF, что даст возможность воспользоваться средствами контроля доступа, репликации файлов, регистрации сервисов и удаленного запуска приложений, реализованными по технологии GRID.

Эта концепция интерфейса пользователя, основных веб-сервисов и модели (мета)данных уже отработана в международных проектах с участием нашей лаборатории: Investigation of Distributed Environmental Archive System, IDEAS,

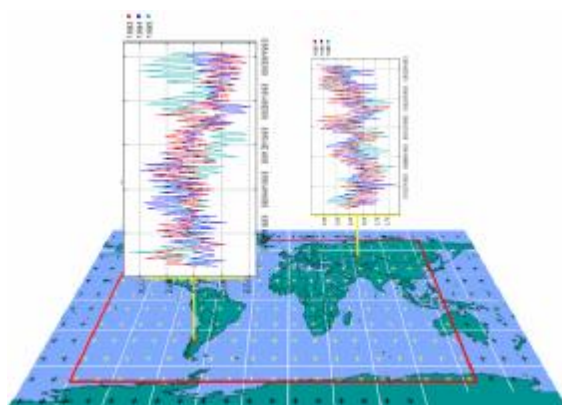
и Space Physics Interactive Data Resource, SPIDR.

Международная программа IDEAS

Экспериментальные серверы IDEAS установлены на параллельных кластерах в объединенном вычислительном центре Геофизического Центра и Института физики Земли РАН и в Национальном центре геофизических данных НОАА США. Основной ресурс данных IDEAS – это ре-анализ метеоданных за последние 55 лет с шагом 6 часов на сетке 2.5 градуса на кластере из 10 параллельных серверов баз данных объемом 1 Тбайт. Одна база данных содержит все параметры ре-анализа за один год.

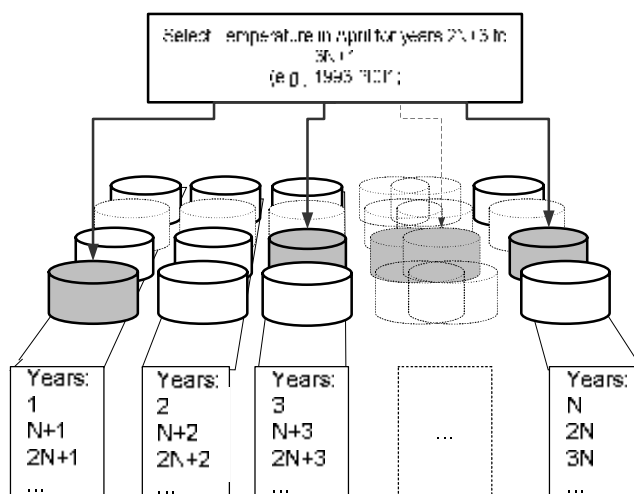
Архитектура системы позволяет легко наращивать список доступных ресурсов данных для параллельного поиска событий, визуализации и экспорта. К настоящему времени помимо ре-анализа исторических данных о погоде, в IDEAS доступны текущие данные и прогноз погоды на 2 недели вперед на глобальной сетке 1 градус, данные о космической погоде (геомагнитные индексы, солнечный ветер), и результаты расчетов по региональным климатическим моделям на сетках с высоким пространственным и временным разрешением (MM5, COAMPS).

Схема данных предполагает хранение *годового временного ряда* в каждой точке сетки и для каждого параметра (элемента) ре-анализа *в виде бинарного объекта BLOB с индексированием параметр-широта-долгота*.



Модель данных IDEAS: индексированная сетка временных рядов в виде BLOB

Каждый сервер компьютерного кластера содержит несколько годовых баз данных с тем, чтобы выборка многолетнего временного ряда по параметру в точках сетки в интервале дат или по сезонам могла производиться параллельно сразу несколькими компьютерами кластера и частичные выборки из готовых базы данных «склеиваются» в общий результат уже на уровне приложения (application program interface, API).



«Распараллеливание» данных ре-анализа на кластере баз данных.

Для поиска событий в окружающей среде, описанных в терминах естественного языка, напр. «самый холодный и дождливый летний день в г. Петропавловск-Камчатский за последние 50 лет», нами разработаны веб-редактор перевода таких описаний событий в запросы к веб-сервису поиска на кластере баз данных по координатам, параметрам ре-анализа и временным интервалам и длительности с применением нечеткой логики.

Temporal Extent

Date from, inclusive (year month day): 19 9 7 Jan 1

Date to, inclusive (year month day): 19 9 8 Jun 30

Time window: 1 day

Fuzzy Search Criteria

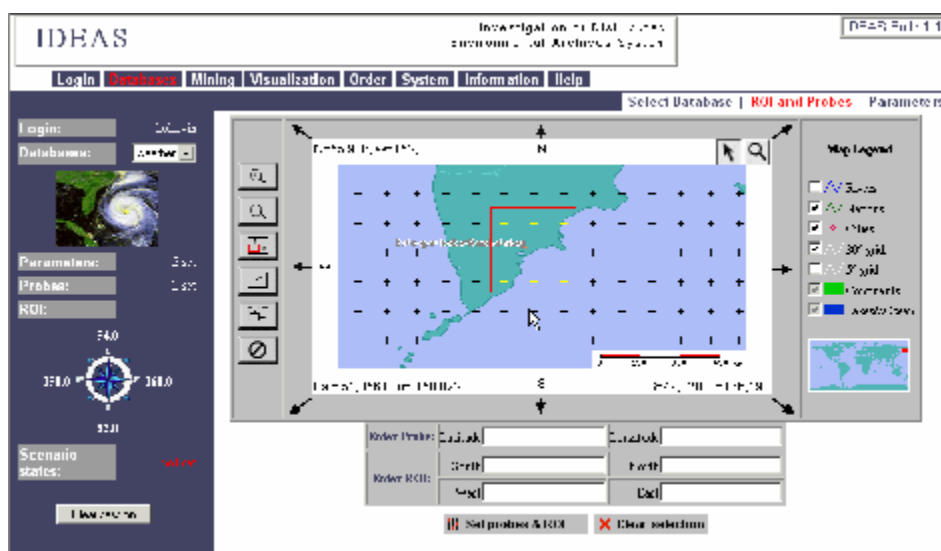
Fuzzy Membership Function

			Height	Very small	Small	Average	Large	Very large	W	U	Range	Any value	Threshold1	Threshold2	Importance
global	temperature		1000	☐	☐	☐	☐	☐	☐	☐	☐	☐	0.0	0.0	1.0
global	geopotential height		1000	☐	☐	☐	☐	☐	☐	☐	☐	☐	0.0	0.0	1.0
spidr	Kp index		KPAP	☐	☐	☐	☐	☐	☐	☐	☐	☐	0.0	0.0	1.0

Update weather scenario state Undo changes

Редактор запросов на поиск событий окружающей среды в терминах нечеткой логики

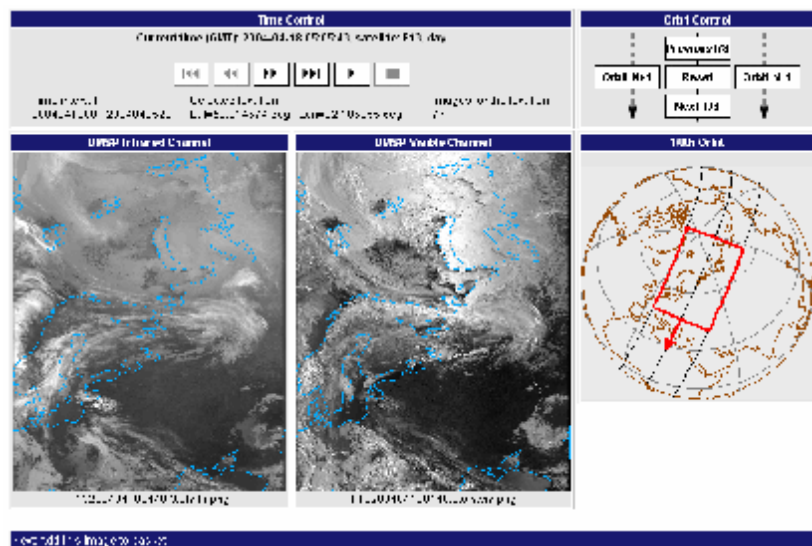
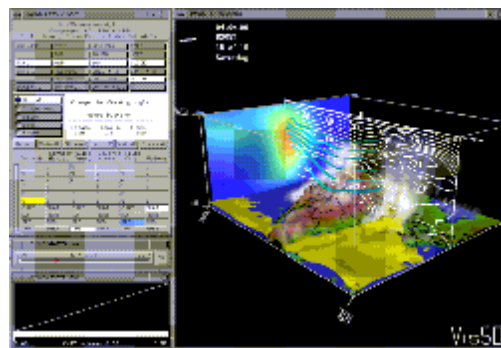
Результатом такого поиска, возвращаемым веб-сервисом, является ограниченный список событий-кандидатов заданной продолжительности (в нашем примере каждое событие длиной один день), упорядоченный по величине общей функции нечеткой принадлежности в виде логического «И» одномерных функций нечеткой принадлежности критериям по каждому из используемых параметров (в нашем примере «низкая температура на высоте 2 м над поверхностью» и «большой уровень осадков» в узле сетки, ближайшем к г. Петропавловск-Камчатский).



GIS-сервис IDEAS: выбор узла сетки прогноза погоды, ближайшего к г. Петропавловск-Камчатский

Данные ре-анализа по каждому из найденных событий можно экспортировать из баз данных в формате NetCDF и загрузить на компьютер пользователя для дальнейшего анализа или моделирования события с высоким пространственно-временным разрешением, а также визуализировать в веб-браузере в виде временных рядов, в виде объемной анимации по технологии Vis5d, или в виде последовательности изображений Земли из космоса над исследуемой точкой сетки (спутники программы DMSP), упорядоченных по времени в интервале дат найденного события.

Визуализация параметров ре-анализа найденного события окружающей среды



Визуализация последовательности изображений Земли в инфракрасном и видимом диапазонах над исследуемой точкой сетки со спутников DMSP

Оба сервера программы IDEAS (русский и американский) находятся в свободном доступе, хотя новому пользователю необходимо пройти простую процедуру автоматической регистрации, данные которой используются для ведения

внутренней статистики загрузки и использования ресурсов.

Международная программа SPIDR

SPIDR представлен несколькими "синхронными" серверами с кластерами баз данных (0.5 Тбайт) по космической погоде за последние 100 лет: индексы, геомагнитные вариации, ионосфера, солнечный ветер, космические лучи, радиотелескопы, изображения Солнца. Серверы SPIDR установлены в России, США, Китае, Японии, Австралии и Южной Африке. Серверы SPIDR получают данные с региональных обсерваторий и спутников в реальном времени и синхронную загрузку на другие узлы по подписке в рамках системы Мировых Центров Данных. Все серверы программы SPIDR находятся в свободном доступе, хотя для возможности загрузки данных из SPIDR пользователю необходимо пройти простую процедуру автоматической регистрации, данные которой используются для ведения внутренней статистики загрузки и использования ресурсов. На август 2004 г. американский узел SPIDR имеет 15 тыс. зарегистрированных пользователей при средней загрузке порядка 200 сеансов визуализации и экспорта данных в сутки.

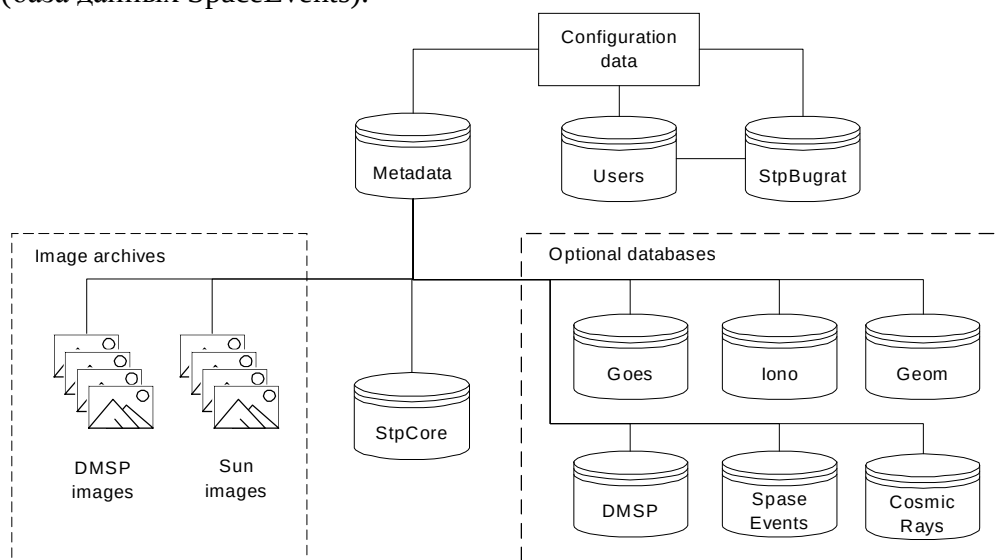
SPIDR объединяет распределенные тематические веб-сервисы ресурсов данных по различным направлениям солнечно-земной физики. Собственно базы данных могут быть установлены в локальной сети рядом с сервером приложений SPIDR и подключаться по протоколу JDBC или запрашиваться с удаленного веб-сервиса сервером приложений по протоколу SOAP, в зависимости от конфигурации конкретного узла SPIDR. Схема подключений баз данных в SPIDR изображена на рис. ниже. Основные разделы данных SPIDR:

- SSN – статистика солнечных пятен, с 1700 г.
- Геомагнитные вариации (минутные и часовые), с 1901 г.

- Геомагнитные и солнечные индексы, с 1932 г.
- Параметры ионосферы, с 1954 г.
- IMF – межпланетное магнитное поле и параметры солнечного ветра, с 1973 г.
- GOES – геостационарные спутники NOAA, с 1986 г.
- DMSP – метеорологические спутники, с 1991 г.
- Изображения Солнца в различных диапазонах и магнитограммы, с 1915 г.

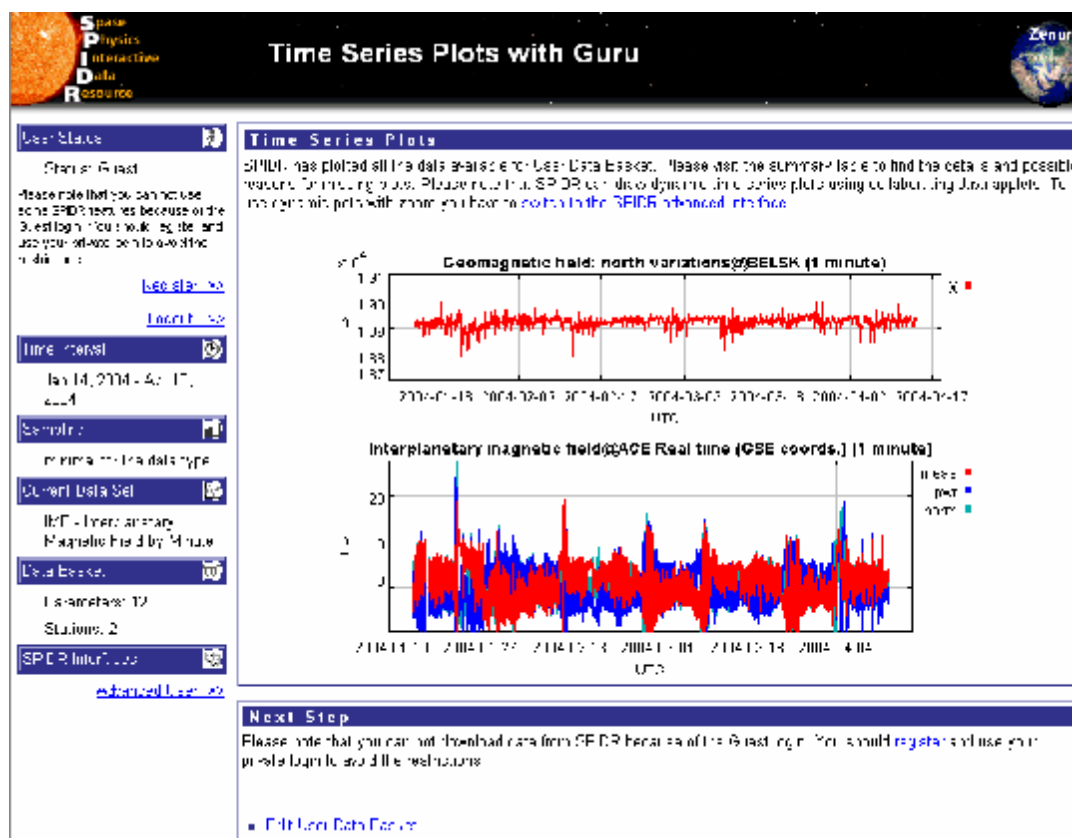
Конфигурационные файлы содержат адреса распределенных серверов баз данных (configuration data). Административная база данных пользователей регионального узла (Users) связана со встроенной базой данных отзывать и предложений по работе системы (StpBugrat). База мета-данных (Metadata) содержит имена Java-классов и/или адреса веб-сервисов для загрузки и экспорта данных из тематических баз, интервалы времени наличия этих данных, физические единицы измерений, заголовки и способы размещения графиков временных рядов, имена и координаты геомагнитных, ионосферных, солнечных обсерваторий и станций космических лучей, а также параметры орбиты и времена жизни научных спутников, данные от которых хранятся в системе.

Ядром SPIDR является база данных StpCore, которая содержит основные геомагнитные и солнечные индексы, параметры межпланетного магнитного поля (IMF), и обычно устанавливается на каждый узел SPIDR. Дополнительно можно подключить базы данных изображений Солнца с различных спутников и обсерваторий (Sun images), базу данных по параметрам околоземной среды со спутников GOES, базу данных с изображениями Земли и орбитальными данными спутников DMSP, временные ряды наблюдений ионосферных и геомагнитных станций (базы данных Iono и Geom), временные ряды наблюдений со станций космических лучей (база данных CosmicRays), и каталог событий космической погоды (база данных SpaceEvents).



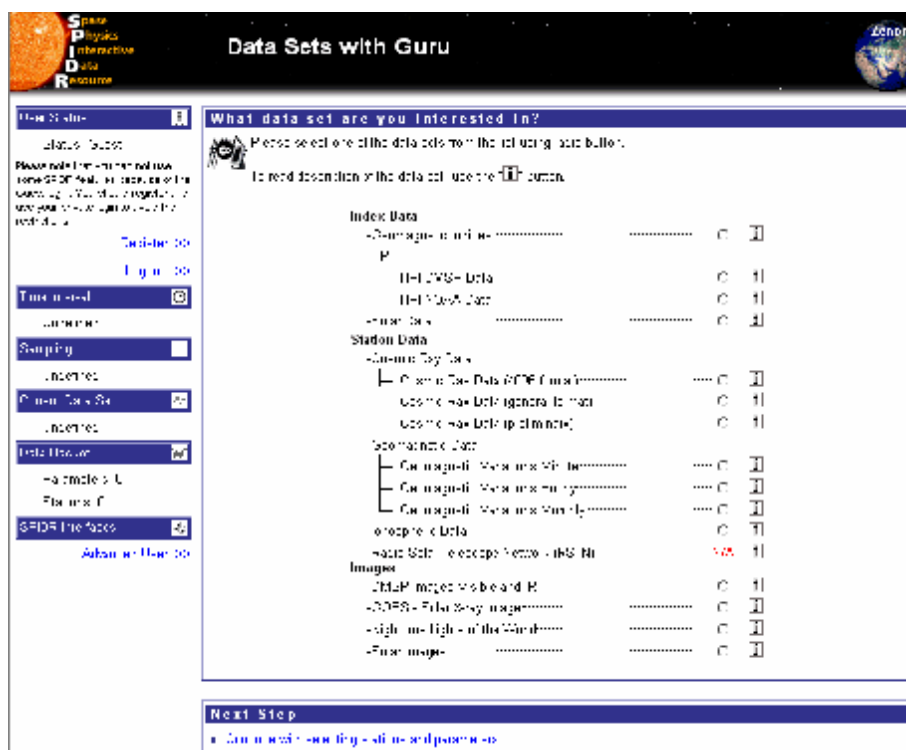
Конфигурация баз данных SPIDR

На наш взгляд, основным преимуществом объединения различных тематических баз данных в общий сетевой ресурс с единым интерфейсом доступа и форматами экспорта является возможность меж-дисциплинарного анализа данных. Для примера ниже мы приводим синхронные графики минутных вариаций магнитного поля на поверхности Земли в польской обсерватории Бельск и в солнечном ветре, поступающие в SPIDR в режиме реального времени. Задача построения подобных графиков и загрузки данных на локальный компьютер у нового пользователя SPIDR в среднем занимает порядка 5 мин.

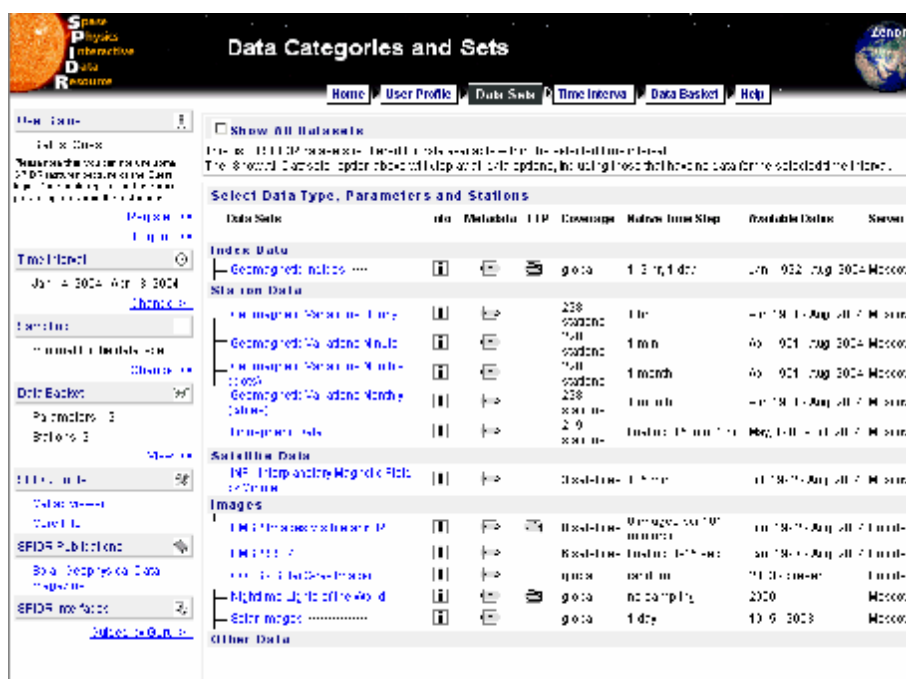


Параллельный анализ тематических ресурсов данных в SPIDR

В 2004 г. разработан новый интерфейс пользователя SPIDR с использованием новой технологии организации динамического веб-интерфейса пользователя Jakarta Struts. Фактически, SPIDR теперь имеет три различных интерфейса: для новых пользователей, для продвинутых пользователей, и для системного администратора. Ниже мы приводим примеры выбора раздела данных с использованием интерфейсов нового и продвинутого пользователей. В последнем случае помимо собственно выбора раздела пользователь может увидеть интервалы дат, количество станций, сервер, на котором расположен веб-сервис ресурса данных по разделу, а также получить метаданные на этот раздел в формате FGDC или перейти на FTP-репозиторий файлов исходных данных, из которых строились базы данных SPIDR.



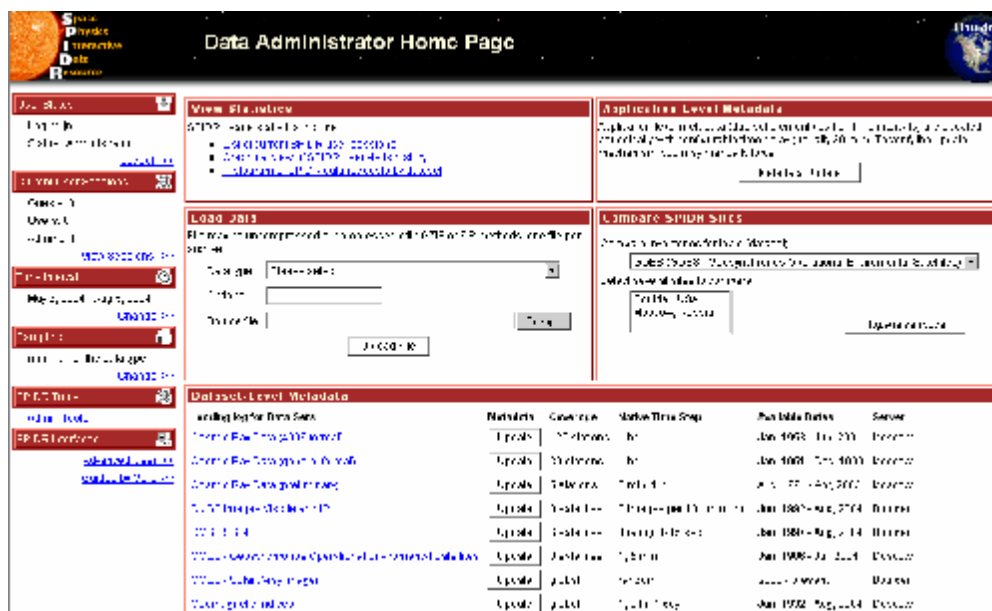
Упрощенный выбор раздела данных из каталога SPIDR в интерфейсе для новых пользователей



Выбор раздела данных из каталога SPIDR, интерфейс для продвинутых пользователей

Интерфейс системного администратора SPIDR позволяет загружать данные в базы данных из файлов, расположенных на компьютере пользователя, анализировать статистику обращений к серверу по пользователям и разделам данных (выборка и загрузка данных), сравнивать содержимое баз данных на разных серверах SPIDR по количеству на-

блюдений параметра на станции в интервале дат, и выполнять синхронизацию баз данных в случае их различия на локальном и удаленном серверах SPIDR.



Интерфейс системного администратора SPIDR

Программа РАН «Электронная Земля»

На первом этапе программы «Электронная Земля» (2004 г.) мы предполагаем определить ресурсы (базы) данных и карты, которые составят «ядро» информационной системы, разработать открытую схему метаданных и создать собственно метаданные, которые бы включали в себя все «ядерные» ресурсы и карты, связать их основными сервисами уровня приложений и построить рабочий прототип интерфейсов пользователя и внешних веб-сервисов. На втором этапе (2005 г.) мы предполагаем совмещение веб-сервисов уровня приложений и ресурсов данных с технологией GRID. На заключительном этапе мы планируем сетевую оптимизацию размещения сервисов, создание новых ресурсов данных и сервисов анализа данных, и окончательную отладку всей системы.

Ссылки

Серверы проекта IDEAS:

- в России
- в США

<http://clust1.wdcb.ru/ideas>
<http://ideas.ngdc.noaa.gov>

Серверы проекта SPIDR:

- в России
- в США

<http://clust1.wdcb.ru/spidr>
<http://spidr.ngdc.noaa.gov>

Используемые технологии

- GRID Globus toolkit
- Java programming language
- Java Server Pages (JSP) Tomcat, Struts
- Web services Axis
- Vis5d/VisAD visualization

<http://www.globus.org>
<http://sun.java.com>
<http://jakarta.apache.org>
<http://xml.apache.org>
<http://www.ssec.wisc.edu/~billh/vis5d.html>,
<http://www.ssec.wisc.edu/~billh/visad.html>

Метаданные и онтологии

- FGDC metadata
- SPDML metadata
- SWEET ontologies

<http://www.fgdc.gov/metadata/metadata.html>
<http://sd-www.jhuapl.edu/SPDML/>
<http://sweet.jpl.nasa.gov/ontology/>