ELEMENTS OF THE VIRTUAL RESEARCH ENVIRONMENT FOR BIG ENVIRONMENTAL DATA ANALYSIS

Evgeny P. Gordov[1,2], *Igor G. Okladnikov*[1,2], *Alexander G. Titov*[1,2], *Alexander Z. Fazliev*[2]

[1]Institute of Monitoring of Climatic and Ecological Systems SB RAS, Tomsk, Russia
[2]V.E. Zuev Institute of Atmospheric Optics SB RAS, Tomsk, Russia

Abstract

The description and the first results of developing a virtual computing and information environment for analysis, assessment and prediction of consequences of global climate changes for ecosystems and climate in the selected region are presented.

Keywords: virtual research environment, big environmental datasets, climate change

ЭЛЕМЕНТЫ ВИРТУАЛЬНОЙ ИССЛЕДОВАТЕЛЬСКОЙ СРЕДЫ ДЛЯ АНАЛИЗА БОЛЬШИХ ДАННЫХ ОБ ОКРУЖАЮЩЕЙ СРЕДЕ

Гордов Е.П.[(1)(2)], Окладников И.Г.[(1)(2)], Титов А.Г.[(1)(2)], Фазлиев А.З.[(2)]

[1] Институт мониторинга климатических и экологических систем СО РАН, Томск
[2] Институт оптики атмосферы имени В.Е. Зуева СО РАН, Томск

Представлено описание и первые результаты разработки виртуальной вычислительно-информационной среды для анализа, оценки и прогноза последствий глобальных климатических изменений для окружающей среды и климата в выбранном регионе.

Ключевые слова: виртуальная исследовательская среда, большие наборы данных об окружающей среде, изменения климата

Введение. Для понимания сложных механизмов изменения климата и его последствий для окружающей среды требуется сбор и последующий анализ геопространственных данных, получаемых в результате наблюдений и численного моделирования [1]. Увеличение разнообразия и объёмов наборов таких данных приводит к невозможности их сбора, обработки и анализа на рабочем месте исследователя с использованием традиционных подходов [2]. В тоже время, необходимость хранить, осуществлять поиск, обмениваться, обрабатывать, анализировать и визуализировать данные об окружающей среде, объём которых в настоящее время уже измеряется в петабайтах, приводит к появлению подходов и инструментов, разрабатываемых для областей науки с интенсивным использованием данных [3-7]. Объёмы, разнообразие и скорость появления современных климатических данных подпадают под модель 5V (Volume, Velocity, Variety, Variability, Veracity) [7] и позволяют уже говорить о них, с учётом их географической привязки, как о "больших геопространственных данных" [8].

Для комплексного использования больших наборов геопространственных метеорологических и климатических данных необходимо создать распределенную программную инфраструктуру [9, 10], основанную на инфраструктуре пространственных данных (ИПД) [11]. Геопортал ИПД [12, 13], при этом, представляет собой единую точку входа, предоставляющую функциональности поиска географических информационных ресурсов, выборки данных, согласно заданным параметрам (функциональность доступа к данным), а также обработки и картографической визуализации в виде соответствующих сервисов и клиентских приложений [14]. В настоящее время считается, что разработка клиентских приложений как элементов такой инфраструктуры должна выполняться с использованием современных веб- и ГИС-технологий [15-18]. Согласно требованиям директивы, INSPIRE к визуализации пространственных данных [19], приложение должно обеспечивать такие функциональные возможности, как просмотр данных, навигацию, прокрутку, масштабирование и наложение графических слоёв, а также отображение легенды и соответствующих метаданных, то есть – базовую функциональность стандартной ГИС.

В настоящее время существует несколько информационных систем и сервисов, предоставляющих подобную функциональность. Система GeoBrain Online Analysis System (GeOnAS) предоставляет доступ к данным спутниковых наблюдений (NASA, USGS) через сервисы Open Geospatial Consortium (OGC, **http://www.opengeospatial.org**), построенные на базе ПО с открытым кодом GRASS GIS, и оснащена веб-интерфейсом, основанным на библиотеке DHTMLX (**http://dhtmlx.com**). Сервис ncWMS [20] – это реализация сервиса OGC Web Map Service (WMS) для геопространственных наборов данных, представленных в формате netCDF. Он активно используется для визуализации данных в рамках геопорталов ИПД, но, к сожалению, слабо поддерживается стандартными ГИС. Портал Unidata THREDDS (**http://www.unidata.ucar.edu/software/thredds/current/tds/TDS.html**) предоставляет доступ к геопространственным данным и метаданным по OPEnDAP, OGC WMS и OGC Web Coverage Service (WCS). Этот продукт также поддерживает выборку данных с использованием ncWMS для визуализации результатов. Открытая распределённая архитектура Boundless / OpenGeo

широко используется для разработки сложных геоинформационных приложений [21, 22]. Она состоит из трёх уровней (данные, сервер приложений и графический интерфейс) и опирается на следующее открытое ПО: ПО Geoserver и Geowebcache (**http://geoserver.org**), реализующее сервисы OGC WMS, WFS, Web Processing Service (WPS); JavaScript-библиотеку OpenLayers (**http://openlayers.org**), которая обеспечивает базовую функциональность "тонкого" веб-ГИС клиента; JavaScript-библиотеку GeoExt / ExtJS library [23] для разработки клиентских веб-приложений с интуитивно понятным графическим интерфейсом пользователя.

В данной работе приводится описание выбранного подхода и первых полученных результатов. В частности, рассматривается разработанная схема хранения больших наборов геопространственных данных, созданная база метаданных, а также графический веб-ГИС клиент пользователя.

Цели и задачи. Данная работа направлена на предоставление специалистам, работающим в смежных научных областях, ориентированных на изучение климатических изменений, оценку их влияния и разработку стратегий адаптации, а также лицам, принимающим решения, точных и подробных климатических характеристик, и надёжного, доступного инструмента для их углубленного статистического анализа, и проведения соответствующих исследований в выбранном регионе. Для достижения этой цели разрабатывается прототип программно-аппаратной платформы виртуальной исследовательской среды (ВИС) для всестороннего изучения наблюдаемых и возможных в будущем изменений климата и их влияния на окружающую среду выбранного региона. Он обеспечит получение корректной климатической информации, необходимой для изучения экономических, политических и социальных последствий глобального изменения климата на региональном уровне.

Подход к хранению данных. В настоящее время применяются два основных подхода к хранению геопространственных данных: геопространственные базы данных и наборы файлов. В качестве примеров использования геопространственных баз данных можно привести такие проекты, как Apache HBase, Esri Geodatabase, Paradigm4, SciDB, и т.д. При таком подходе данные необходимо вносить в базу данных до их непосредственного использования, что требует значительного времени и дополнительного дискового пространства. Второй подход опирается на использование обычных коллекций файлов с данными в рамках типовой файловой системы. В случае геопространственных данных обычно используются самоописательные форматы файлов, содержащие, помимо самих данных, их метаданные. Было показано [24], что скорость выборки фрагментов данных объёмом более 40 Мб из пространственной базы данных может быть ниже, чем при непосредственном чтении из набора файлов с данными. Хотя для работы с наборами файлов требуется разработка и использование дополнительных программных адаптеров, обеспечивающих интерфейсы (API) для записи, чтения и обработки распределённых файловых наборов, нами был выбран именно это подход за относительную простоту его реализации и более высокую скорость выборки больших фрагментов данных. В качестве основного самописательного формата файлов для хранения данных был выбран формат Network Common Data Form (netCDF), принятый различными научными организациями и OGC в качестве стандартного формата хранения и обмена геопространственными данными.

Таким образом, массивы данных хранятся в виде наборов netCDF-файлов и располагаются в строгой иерархии каталогов:

/<путь к корневому каталогу с данными>/
<название архива данных>/
<горизонтальное разрешение>/
<разрешение по времени>/
<набор файлов и каталогов с данными>

Здесь <путь к корневому каталогу с данными> определяется системным администратором, <название архива данных> задаёт имя каталога, содержащего все данные одного архива данных, <горизонтальное разрешение> задаёт имя каталога, содержащего данные с одним горизонтальным разрешением, <разрешение по времени> задаёт имя каталога, содержащего данные с одним шагом по времени. Далее по иерархии располагаются файлы с данными. Имена файлов и подкаталогов не регламентируются и определяются индивидуальными особенностями конкретного набора данных. Каждый файл содержит многомерный массив геопривязанных значений одного или нескольких метеорологических параметров.

Архитектура базы метаданных. Для описания наборов геопространственных данных и процедур их обработки, и обеспечения эффективного функционирования ВИС была разработана специализированная база метаданных. Эта база содержит описание пространственно-временных характеристик доступных для обработки наборов данных, расположение файлов с данными, а также описание выходных параметров программных компонент для анализа данных. Набор данных – это совокупность данных, заданных на единой временной и пространственной сетках, едином временном интервале и полученные при одних и тех же условиях моделирования или наблюдений (сценарии). Он может быть представлен как одним, так и несколькими однотипными файлами. Каждый файл содержит один или несколько метеопараметров в виде многомерных массивов, снабжённых метаданными. Состав метеопараметров и длина временного интервала, а также названия метеопараметров во всех файлах, входящих в один набор данных, одинаковые. Метеопараметр – это стандартизованное название некоторой метеорологической величины: температура, давление, влажность. Переменная – это собственное название многомерного массива в файле формата netCDF. Также, в netCDF-файле присутствуют особые переменные, содержащие горизонтальные и вертикальные сетки, а также сетку по времени.

Поскольку в рамках одной организации и одного проекта могут быть получены наборы данных с различным пространственным и временным разрешением, вводится понятие «коллекция данных». Коллекция данных – это совокупность наборов данных, полученных в одной организации в рамках одного проекта и заданных с разным пространственным и/или временным шагом, а также для различных сценариев. Коллекция может состоять из одного набора данных.

По назначению таблицы в БМД можно разделить на «технические» (содержат данные, необходимые для функционирования вычислительного ядра ВИС) и «интерфейсные» (содержат данные, используемые для наполнения элементов графического интерфейса пользователя). Некоторые интерфейсные таблицы могут содержать записи на различных языках.

Каждый набор климатических данных определяется совокупностью четырёх характеристик: названием коллекции, в которую он входит, горизонтальным разрешением, шагом сетки по времени и названием сценария (если применимо). Каждый набор климатических данных включает в себя один или несколько массивов данных. Каждый такой массив содержит значения какого-то метеопараметра, заданного на пространственной и временной сетках и определяется набором данных, переменной (метеопараметром) и вертикальным уровнем.

Для обработки данных с использованием вычислительного ядра ВИС необходимо подготовить и передать ему специализированный файл в формате XML (файл-задание). Этот файл содержит описание и уникальную для каждого вида обработки последовательность вы-

зова различных модулей обработки данных. В базе метаданных содержатся описания процедур-обработчиков данных, их выходные параметры и расположение шаблонных файлов, на основе которых подготавливаются задания на обработку данных для вычислительного ядра.

Веб-ГИС клиент. Разработанное картографическое веб-приложение (веб-ГИС клиент) основано на архитектуре Boundless / OpenGeo и может быть представлено в виде трёх основных функциональных уровней [25]:

- уровень метаданных netCDF в формате JSON;
- уровень промежуточного ПО, предоставляющего методы для работы с:
  - метаданными;
  - файлом-заданием в формате XML;
  - картографическими сервисами WMS/WFS.
- уровень графического интерфейса пользователя, представленного JavaScript-объектами, реализующими общую логику работы приложения.

Веб-ГИС клиент соответствует общим требованиям стандарта INSPIRE и обеспечивает запуск сервисов обработки данных для задач мониторинга окружающей среды и исследования изменений климата, а также отображения результатов обработки в виде картографических слоёв WMS/WFS в растровом (PNG, JPG, GeoTIFF), векторном (KML, GML, Shape) и двоичном (NetCDF) форматах.

Уровень метаданных netCDF. Уровень метаданных netCDF веб-ГИС клиента представляет собой набор взаимосвязанных JSON-объектов, созданных на основе MySQL базы метаданных, и содержащих информацию о наборах геопространственных данных (пространственное и временное разрешения, перечень доступных метеопараметров, перечень доступных процедур обработки и т.д.). В общем случае возможно два типа объектов:

- объекты, имеющие структуру, эквивалентную соответствующим таблицам и взаимотношениям в базе метаданных;
- объекты, созданные на основе сложных SQL-запросов к базе метаданных, позволяющие быстро получать необходимую информацию из базы метаданных, используя MySQL-индексы, как ключи в ассоциативном массиве.

Структура JSON-объектов была выбрана на основе следующих критериев:

- эффективность заполнения интерактивных форм в графическом интерфейсе пользователя;
- оптимизация процесса создания и редактирования XML-файла, описывающего конфигурацию обработки данных (XML файл-задание).

Таким образом, на данном уровне веб-ГИС клиента оптимизируются процессы взаимодействия пользователя с базой метаданных через графический интерфейс.

Уровень промежуточного ПО. На этом уровне реализуются методы работы с метаданными netCDF, XML файлом-заданием и картографическими сервисами WMS/WFS. Он представляет собой промежуточное ПО, связывающее уровень представления метаданных в формате JSON с уровнем графического интерфейса пользователя. Методы, реализованные на этом уровне, обеспечивают:

- загрузку и обновление JSON-объектов метаданных, используя технологию AJAX;
- создание, редактирование и сериализацию объекта XML-задания;
- запуск и контроль выполнения задачи обработки данных на удалённом вычислительном узле;
- работу с картографическими сервисами WMS/WFS, а именно: получение списка доступных слоёв, отображение слоёв на карте, экспорт слоёв в различные форматы по запросу пользователя, получение и отображение легенды слоя с выбранным SLD-стилем.

Графический интерфейс пользователя. Этот уровень основан на объединении JavaScript-библиотек, таких как OpenLayers, GeoExt и ExtJS, и представляет собой набор программных компонент, как независимых (информационные панели, кнопки, списки слоёв, и

т.п.), так и реализующих общую логику реализации приложения (меню, панели инструментов, мастеры (wizards), обработчики сообщений мыши и клавиатуры и т.д.). Графический интерфейс выполняет две основные функции: предоставление функциональных возможностей для редактирования XML файла-задания и представление картографической информации конечному пользователю. Внешне он похож на интерфейсы таких популярных классических ГИС-приложений, как uDig, QuantumGIS и т. д. Основные элементы графического интерфейса пользователя представлены на рис. 1.

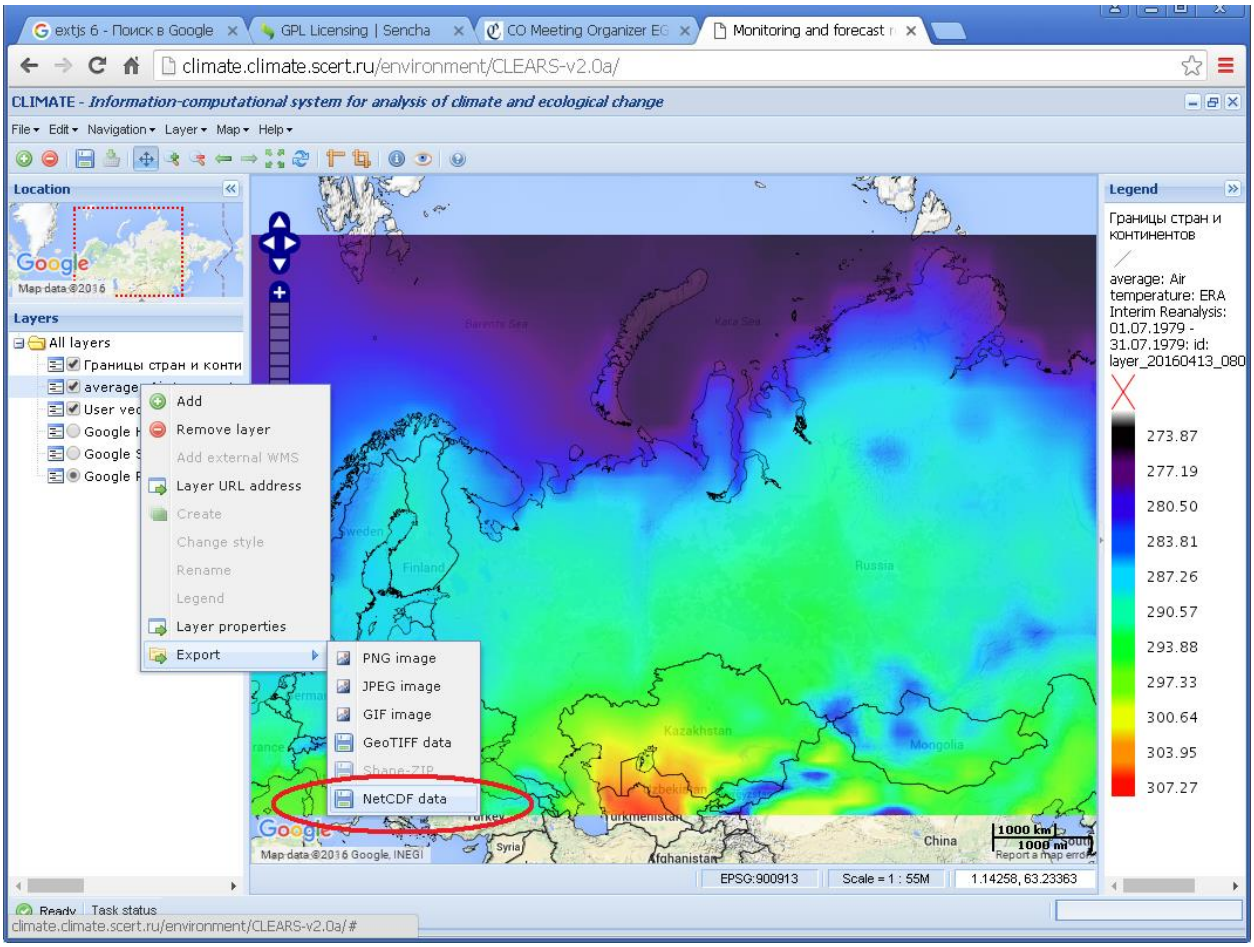

Рис. 1. Графический интерфейс пользователя веб-ГИС клиента.
Демонстрация экспорта слоя в формат netCDF.

Заключение. На сегодняшний день не существует общепринятого формализованного описания схемы базы метаданных больших наборов пространственно-привязанных климатических данных и представленная архитектура является, в своём роде, первой в мире попыткой решения данной фундаментальной задачи. Разработанная база метаданных решает три основные задачи: 1) содержит информационное наполнение для форм графического интерфейса пользователя; 2) предоставляет геопорталу информацию, необходимую для формирования корректного файла-задания для вычислительного ядра; 3) содержит информацию о структуре и расположении наборов данных, необходимую вычислительному ядру для их чтения и обработки. Применение этой базы данных систематизирует информацию об имеющихся наборах данных, облегчает автоматический поиск файлов данных и способствует повышению масштабируемости и гибкости вычислительной системы.

Разработанный веб-ГИС клиент основан на архитектуре Boundless / OpenGeo. Его первая версия основана на JavaScript-библиотеках OpenLayers, GeoExt и ExtJS, и представляет

собой набор программных компонент, реализующих как общую логику работы приложения, так и независимые элементы графического интерфейса пользователя.

Первое применение разработанной базы метаданных и веб-ГИС клиента в рамках геопортала показало, что их совместное использование унифицирует и упрощает процедуру расширения архива наборов данных, доступных для анализа, а также добавление новых функциональных модулей их обработки [26].

Полученные результаты показывают, что разрабатываемая ВИС, включая интерактивные инструменты анализа климатических данных, будет полезна как для лиц, ответственных за принятие решений, связанных с оценкой социально-экономических и экологических последствий, разработкой стратегий адаптации, выработкой научной политики, так и для профильных специалистов, работающих в областях науки, связанных с изучением климатических изменений. На разработанной основе данные категории пользователей получат корректные оценки климатических характеристик, необходимые для изучения экономических, политических и социальных последствий глобального изменения климата на региональном уровне.

Работа выполнена при финансовой поддержке РНФ (грант №16-19-10257).

ЛИТЕРАТУРА

[1] Lykosov V.N., Glazunov A.V., Kulyamin D.V., Mortikov E.V., Stepanenko V.M. Supercomputing Modeling in Physics of Climatic System. Moscow State University Publishing House, 2012, 402 p.

[2] Gordov E.P., Kabanov M.V., Lykosov V.N. Information-Computational Technologies for Environmental Science: Preparation of Young Researchers // Computational Technologies. Special Issue 1. 2006. V. 11. P. 3-15.

[3] MIKE 2.0. The open source standard for Information Management. Big Data Definition. **http://mike2.openmethodology.org/wiki/Big_Data_Definition** (дата обращения 29.06.2017).

[4] Dan Kusnetzky. What is "Big Data?". ZDNet. **http://www.zdnet.com/blog/virtualization/what-is-big-data/1708** (дата обращения 29.06.2017)

[5] Ashley Vance. Start-Up Goes After Big Data With Hadoop Helper. New York Times Blog. **http://bits.blogs.nytimes.com/2010/04/22/start-up-goes-after-big-data-with-hadoop-helper** (дата обращения 29.06.2017).

[6] Калиниченко Л.А. и др. Проблемы доступа к данным в исследованиях с интенсивным использованием данных в России // Информатика и её применения. М: ИПИ РАН. 2016. Т. 10, № 1. С. 3-23.

[7] Hilbert, Martin. "Big Data for Development: A Review of Promises and Challenges. Development Policy Review". **http://www.martinhilbert.net** (дата обращения 29.06.2017).

[8] Shekhar S. Spatial Big Data // Proc. AAG-NIH Symp. on Enabling a National Geospatial Cyberinfrastructure for Health Research. Minneapolis. USA, 2012.

[9] Gordov E.P., Lykosov V.N. Development of information-computational infrastructure for integrated study of Siberia environment // Computational Technologies. Special Issue 2. 2007. V. 12. P. 19-30.

[10] Stefano Nativi, Mohan Ramamurthy, Bernd Ritschel. EGU-ESSI Position Paper. **http://scert.ru/files/EGU-PositionPaper-final.pdf** (дата обращения 29.06.2017).

[11] Steiniger S., Hunter A.J.S. Free and open source GIS software for building a spatial data infrastructure. / In: Bocher E., Neteler M., (eds.), Geospatial Free and Open Source Software in the 21st Century, LNGC, Heidelberg, Springer, 2012. P. 247-261.

[12] Koshkarev A.V., Ryakhovskii A.V., Serebryakov V.A. Infrastructure of distributed environment of storage, search and transformation of geospatial data // Open Education. 2010. № 5. P. 61-73.

[13] Краснопеев С.М. Опыт развёртывания ключевых элементов инфраструктуры пространственных данных на базе веб-служб // Труды XIV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2011). Санкт-Петербург, 2011. С. 92-99.

[14] Koshkarev A.V. Geoportal as a tool to control spatial data and services. // Spatial data. 2008. № 2. P. 6-14.

[15] Yakubailik O.E. Geoformation geoportal // Computational Technologies. Special Issue 3. 2007. V. 12. P. 116-125.
[16] Dragicevic, S., Balram, S., Lewis, J. The role of Web GIS tools in the environmental modeling and decision-making process // 4th International Conference on Integrating GIS and Environmental Modeling (GIS/EM4): Problems, Prospects and Research Needs. Banff, Alberta, Canada, 2000.
[17] Frans J. M. van der Wel. Spatial data infrastructure for meteorological and climatic data // Meteorol. Appl. 12, 2005. Pp. 7-8.
[18] Vatsavai, Ranga Raju, Thomas E. Burk, B. Tyler Wilson, Shashi Shekhar. A Web-based browsing and spatial analysis system for regional natural resource analysis and mapping // Proc. of the 8th ACM int. symp. on Advances in geographic information systems. Washington, D.C., US. 2000. P. 95-101.
[19] Katleen Janssen. The Availability of Spatial and Environmental Data in the European Union: At the Crossroads Between Public and Economic Interests. Kluwer Law International, 2010. 617 p.
[20] J.D. Blower, A.L. Gemmell, G.H. Griffiths, K. Haines, A. Santokhee, X. Yang. A Web Map Service implementation for the visualization of multidimensional gridded environmental data // Environmental Modelling & Software. 2013. V. 47. P. 218-224.
[21] L. Becirspahic and A. Karabegovic. Web portals for visualizing and searching spatial data // Information and Communication Technology, Electronics and Microelectronics (MIPRO), 2015. Opatija, Croatia. P. 305-311.
[22] I.G. Okladnikov, E.P. Gordov, A.G. Titov, T.M. Shulgina. Information-computational System for Online Analysis of Georeferenced Climatological Data // Selected Papers of the XVII International Conference on Data Analytics and Management in Data Intensive Domains (DAMDID/RCDL 2015), 2015. Obninsk, Russia / Ed.: Leonid Kalinichenko and Sergey Starkov. CEUR Workshop Proceedings. Vol. 1536. P. 76-80.
[23] Shea Frederick, Colin Ramsay, and Steve Cutter Blades. Learning Ext JS. Packt Publishing, 2008. 299p.
[24] A. Santokhee, J. Blower, K. Haines. Storing and Manipulating Gridded Data In Spatial Databases // Reading E-science Center, University of Reading. **http://go-essp.gfdl.noaa.gov/presentations/06_06_05/Santokhee/Adit_Sank.ppt%20%5BRead-Only%5D.pdf** (дата обращения 29.06.2017).
[25] Титов А.Г., Гордов Е.П., Окладников И.Г. Разработка Веб-ГИС на основе сервисов обработки и визуализации пространственных данных для анализа и прогнозирования региональных климатических изменений // Информационные и математические технологии в науке и управлении. 2016. № 4-2. С. 96-109.
[26] Ryazanova A.A., Voropay N.N., Okladnikov I.G. Application of information and computing web system «Climate» for estimationn of aridity of South Siberia. // Proc. of International Conference and Early Career Scientists School on Environmental Observations, Modeling and Information Systems ENVIROMIS-2016, 2016. Tomsk, Russia. P. 358-362.