УДК 004.9

А. Г. Титов, Е. П. Гордов, И. Г. Окладников

Институт мониторинга климатических
и экологических систем СО РАН,
пр. Академический, 10/3, Томск, 634055, Россия
E-mail: titov@scert.ru

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИЙ SEMANTIC WEB В ИНФОРМАЦИОННО-ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЕ ДЛЯ АНАЛИЗА ДАННЫХ ПО ОКРУЖАЮЩЕЙ СРЕДЕ *

В статье описана рабочая модель информационно-вычислительной системы для аннотации, хранения, семантического поиска, а также обработки и визуализации наборов пространственно-распределенных данных, содержащих результаты как метеорологических наблюдений, так и математического моделирования климатических процессов. Для создания пригодных к автоматизированной обработке описаний наборов данных, используемых в системе, разработан прототип стандарта метаданных в виде оригинальной RDF-схемы. В настоящее время система доступна в виде веб-сайта, реализующего функциональность работы с наборами геофизических данных, включая их семантический поиск, статистический анализ и визуализацию. Предложенная система представляет собой шаг в процессе разработки информационно-вычислительной инфраструктуры для поддержки мультидисциплинарных исследований окружающей среды на глобальном и региональном уровнях.

Ключевые слова: информационные системы, метаданные, веб-технологии, метеорология, изменение климата.

Введение

Особенностью наук о Земле является разнородность важных для исследователей наборов пространственно-распределенных геофизических данных, содержащих ряды наблюдений, данные реанализа, результаты математического моделирования климатических процессов. При этом значительное число наборов данных является результатом выполнения небольших научных проектов, что ограничивает доступ к ним и часто приводит к их фактической утрате по окончании исследований. Таким образом, для их эффективного использования необходима соответствующая инфраструктура, обеспечивающая их обработку, хранение, поиск и предоставление доступа [1]. В настоящее время в Интернете доступен ряд ресурсов, которые решают в той или иной мере подобные задачи: сайт Британского центра атмосферных данных (http://badc.nerc.ac.uk/home/index.html), портал GEON (http://www.geongrid.org/), портал ISDC (http://isdc.gfz-potsdam.de), сайт NERIN (http://nerin.scert.ru/), портал Genesi-DR (http://www.genesi-dr.eu/), и др.

Следует отметить, что ключевым элементом в реализации поисковых систем в Web является использование метаданных для описания веб-ресурсов. Первоначально для этой цели применялись специализированные HTML-теги и атрибуты (title, keywords и description). В связи со сложностью и специфичностью пространственно-распределенных данных, используемых в области наук о Земле, этот подход не приводит к удовлетворительным результатам. Значительно более продуктивным является использование XML для создания стандартизованного синтаксиса метаданных с целью адекватной каталогизации и последующего поиска данных по окружающей среде [2]. Еще более перспективным выглядит использование Semantic Web [1] для улучшения технологии поиска по метаданным путем использования семантических ссылок и соответствий между концепциями предметной области, в том числе при наличии ее нескольких формализованных описаний [3]. В рамках этого подхода также активно ведется исследовательская работа по созданию алгоритмов поиска пространственно-распределенной информации на основе соответствующих пространственных и терминологических онтологий [4; 5].

* Работа выполнена при частичной поддержке Программы фундаментальных исследований СО РАН № 4.5.2 и интеграционных проектов СО РАН № 4, 50 и 66.

[1] http://www.w3.org/2001/sw/.

Однако существующие в настоящее время инструменты, как правило, не обеспечивают возможность оперативной статистической обработки и анализа данных. Таким образом, актуальной является задача хранения, эффективного поиска и специфической обработки (с учетом особенностей предметной области) необходимых для научных исследований наборов геофизических данных, а также организации к ним оперативного доступа. Особенно важной эта задача является для регионов с наблюдаемыми быстрыми изменениями климата, к которым относятся Северная Евразия и, в частности, Сибирь. Последнее обстоятельство усугубляется тем, что для этого региона число доступных структурированных архивов данных относительно невелико.

Постановка задачи

Предлагаемый подход к универсальному решению поставленной задачи подразумевает построение на базе веб-технологий интегрированной информационно-вычислительной системы, обладающей необходимой для исследователей функциональностью:

- обработка данных с использованием методов математической статистики;
- графическая визуализация данных и результатов;
- работа с данными на стороне веб-клиента с использованием веб-ГИС технологий (выбор географического диапазона, масштабирование, использование слоев, а также спутниковых снимков).

Таким образом, для достижения поставленной цели необходимо следующее.

1. Создание хранилища наборов данных, которое, по причине часто очень большого объема архивов, будет иметь распределенную структуру.
2. Разработка схемы метаданных для описания наборов данных. При этом необходимо учесть основные стандарты описания данных по окружающей среде, используемые в настоящее время (OGC – http://www.opengeospatial.org/).
3. Создание базы метаданных (каталога).
4. Разработка универсального набора программных инструментов для работы с метаданными, включая их создание, редактирование, визуализацию, индексирование, модуль для автоматического сбора метаданных сторонних разработчиков (metadata harvesting), а также поисковую систему.
5. Обеспечение онлайн-сервисов для работы с пространственно-распределенными данными.

Схема метаданных

В настоящее время в области наук о Земле для описания наборов данных и веб-ресурсов широко применяются стандарты метаданных, такие как Дублинское Ядро (Dublin Core – http://dublincore.org/), Ecological Metadata Language (http://knb.ecoinformatics.org/software/eml/), Directory Interchange Format (DIF – http://gcmd.gsfc.nasa.gov/User/difguide/difman.html), FGDC Content Standard for Digital Geospatial Metadata (http://www.fgdc.gov/metadata/csdgm/), ISO 19115:2003 «Geographic information – Metadata», ISO 19139 «Geographic information – Metadata – XML schema implementation», онтологии SWEET (Semantic Web for Earth and Environmental Terminology – http://sweet.jpl.nasa.gov/), и некоторые другие. За исключением онтологий SWEET и спецификации Дублинского Ядра, которая имеет версию для Semantic Web, эти стандарты представляют собой спецификации XML, жестко определяющие синтаксис метаданных; при этом любая предполагаемая семантика находится вне сферы действия спецификации XML [6]. Для обеспечения критерия семантической интероперабельности, а также возможности реализации элементов семантического поиска наборов данных, для повышения эффективности классических алгоритмов поиска в терминах точности (*precision*) и полноты (*recall*) [7; 8], при разработке схемы метаданных предлагается использовать средства технологии RDF (Resource Description Framework – http://www.w3.org/RDF/), являющейся на данный момент стандартом де-факто описания сетевых ресурсов. Под семантической интероперабельностью понимается возможность вычислительных систем обмениваться информацией и понимать ее точный смысл, т. е. информация всегда может быть представлена

системой в предназначенном для нее формате. Таким образом, семантическая интероперабельность связана с установлением соответствий между терминами, используемыми в данных различного формата [6; 9].

Технология RDF основана на простой, но устойчивой модели данных, в которой ресурсы описываются в терминах их свойств. Свойства могут принимать как атомарные значения, такие как строки текста и числа, так и представлять собой другие ресурсы, имеющие собственные свойства. Используя эту легко расширяемую и устойчивую логическую модель, можно создать структурированные метаданные для описания произвольных объектов вообще и сетевых ресурсов в частности. Для обеспечения механизмов описания групп связанных ресурсов (принадлежность к определенному типу, позиция в иерархической структуре) и отношений между ресурсами, используется язык описания словарей RDF, RDF Schema (http://www.w3.org/TR/rdf-schema/). Следует отметить, что конкретная RDF-схема фактически представляет собой простую онтологию, формализующую описание набора фактов о рассматриваемой предметной области.

Упомянутые выше XML и RDFS-стандарты для описания пространственно-распределенных данных, а также XML-спецификация, созданная в рамках Региональной информационной сети по Северной Евразии (NERIN – http://nerin.scert.ru/), были использованы в работе по созданию оригинальной RDF-схемы, первая версия которой описана в [10], предназначенной для описания наборов данных по таким предметным областям, как метеорология, климат и перенос атмосферного загрязнения (рис. 1). Очевидно, что, поскольку модель RDF базируется на триплетах, метаданные в формате, соответствующем произвольной RDF-схеме, могут храниться и эффективно использоваться путем применения унифицированных программных продуктов, таких как Jena [2], Kowari [11], 3store [12] и т. д., что невозможно в случае использования XML [13].

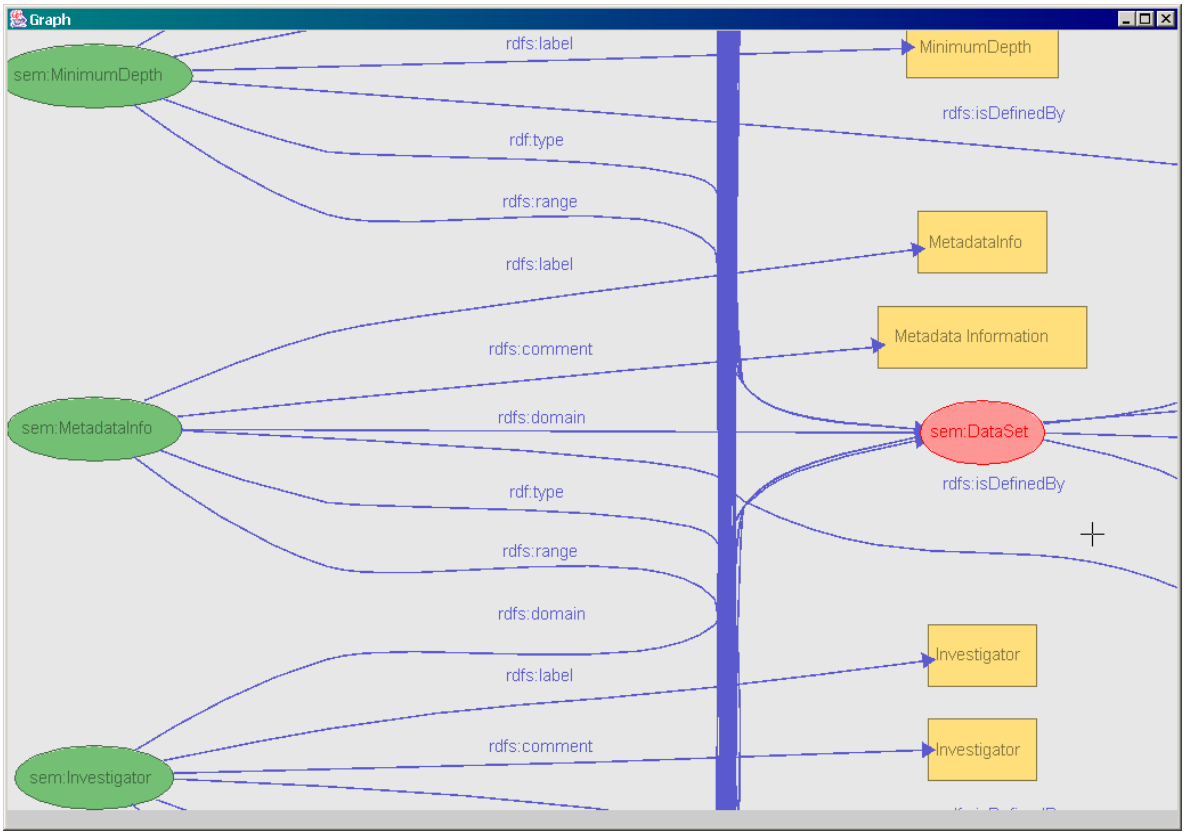

Рис. 1. Фрагмент RDF-схемы в графическом представлении (редактор IsaVIZ)

[2] A Semantic Web Framework for Java – http://jena.sourceforge.net/.

Информационный поиск в Semantic Web

Проблема поиска информации не теряет своей актуальности со времени возникновения Web. Такие системы для индексирования и поиска информации, как Yahoo, Alta-Vista, Google значительно повлияли на методологию доступа и представления сетевых информационных ресурсов. Минусом традиционных поисковых систем является то, что они сконструированы для поиска текстовой информации в слабо-структурированном распределенном хранилище, каким является традиционный Web, и практически не учитывают семантического содержания индексируемых документов, что заметно снижает эффективность поиска [14]. Следует отметить, что получившим наибольшее распространение видом поиска является текстовый поиск по ключевым словам, когда сложность поисковой системы скрыта за внешне элементарным графическим интерфейсом, представленным в виде единственного поля ввода. Аналогичный подход, наряду с предоставлением и более сложного интерфейса, отражающего особенности используемой RDF-схемы или онтологии, используется и при реализации семантического поиска информационных ресурсов [15–17], метаданные которых, как и в частном случае наборов данных по окружающей среде, имеют логически достаточно сложную структуру.

В настоящее время общая процедура семантического поиска имеет следующий вид, согласно результатам анализа 35 существующих систем [18].

1. Создание запроса пользователем.

2. Выполнение поискового алгоритма, включая процедуры синтаксического сравнения (точное совпадение, совпадение подстроки, и т. д.) и семантического сравнения, таких как алгоритм обхода RDF-графа, логический вывод на основе RDFS/OWL [3], и т. д.

3. Представление результатов поиска, включая алгоритмы упорядочивания, такие как PageRank [19], и визуализацию.

Таким образом, задача построения системы для семантического поиска пространственно-распределенных данных сводится к выбору хранилища RDF-метаданных и комбинации поисковых алгоритмов, а также к технической реализации элементов системы на языке программирования высокого уровня.

Полученные результаты

На данный момент модель системы для хранения, поиска и аналитической обработки данных по окружающей среде доступна по адресу http://climate.risks.web.scert.ru/metadatabase/ в виде веб-приложения. Реализована следующая функциональность для работы с наборами данных: добавление / редактирование метаданных, поиск по авторам метаданных и по исследователям, входящим в соответствующий проект, семантический поиск по ключевым словам, а также статистический анализ и визуализация выбранных архивов данных [20]. Общая структура модели системы представлена на рис. 2.

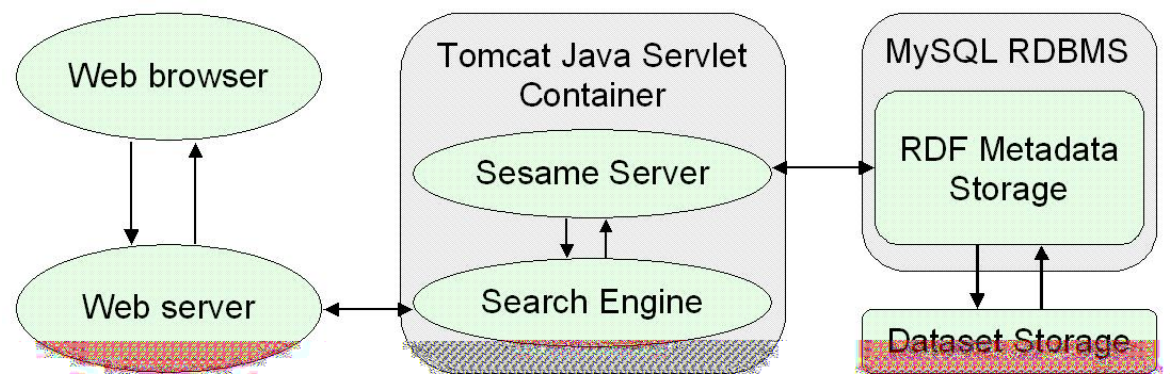

Рис. 2. Архитектура модели системы

[3] RDF Semantics, W3C Recommendation. 10 February 2004, http://www.w3.org/TR/rdf-mt/.

Модель системы реализована на платформе веб-сервера Apache (http://httpd.apache.org/), в связке с Java-сервлет контейнером Tomcat (http://tomcat.apache.org/). Сервер репозитария RDF-метаданных Sesame [21] функционирует в рамках среды Tomcat и обеспечивает необходимый инструментарий для анализа, интерпретации, создания запросов и хранения RDF-метаданных. Поскольку Sesame может использовать внешнюю СУБД для хранения данных RDF, для реализации хранилища была выбрана хорошо зарекомендовавшая себя СУБД MySQL (http://dev.mysql.com/). Выбор открытого ПО Sesame был обусловлен его богатыми возможностями для работы с RDF-данными, встроенной поддержкой протокола HTTP, поддержкой логического вывода на основе RDF [22], а также простотой настройки и наличием удобного интерфейса администрирования.

Собственно веб-приложение разработано на основе веб-портала ATMOS [23], который является универсальной программной средой для быстрой разработки приложений научной тематики. Для интеграции ПО Sesame и ATMOS был использован продукт Phesame [24]. Алгоритм семантического поиска реализован на базе запросов на языке SeRQL (Sesame RDF Query Language, [22]), который используется для конструирования на основе введенных ключевых слов множества запросов к хранилищу RDF-метаданных, при этом результаты поиска (рис. 3) упорядочиваются по вычисляемой степени релевантности узлов RDF-графа, соответствующих запросу пользователя. На правой панели рис. 3 в качестве примера представлены RDF-метаданные, полученные из репозитария Sesame, для набора данных «реанализ ECMWF ERA-40». Следует отметить, что в ряде случаев в описании будет иметься поле «Data Processing URL», которое ссылается на ПО для обработки конкретного архива данных. Например, для выбранного набора данных «реанализ ECMWF ERA-40» мы можем перейти непосредственно к специализированному веб-приложению, реализующему процедуры статистического анализа [25].

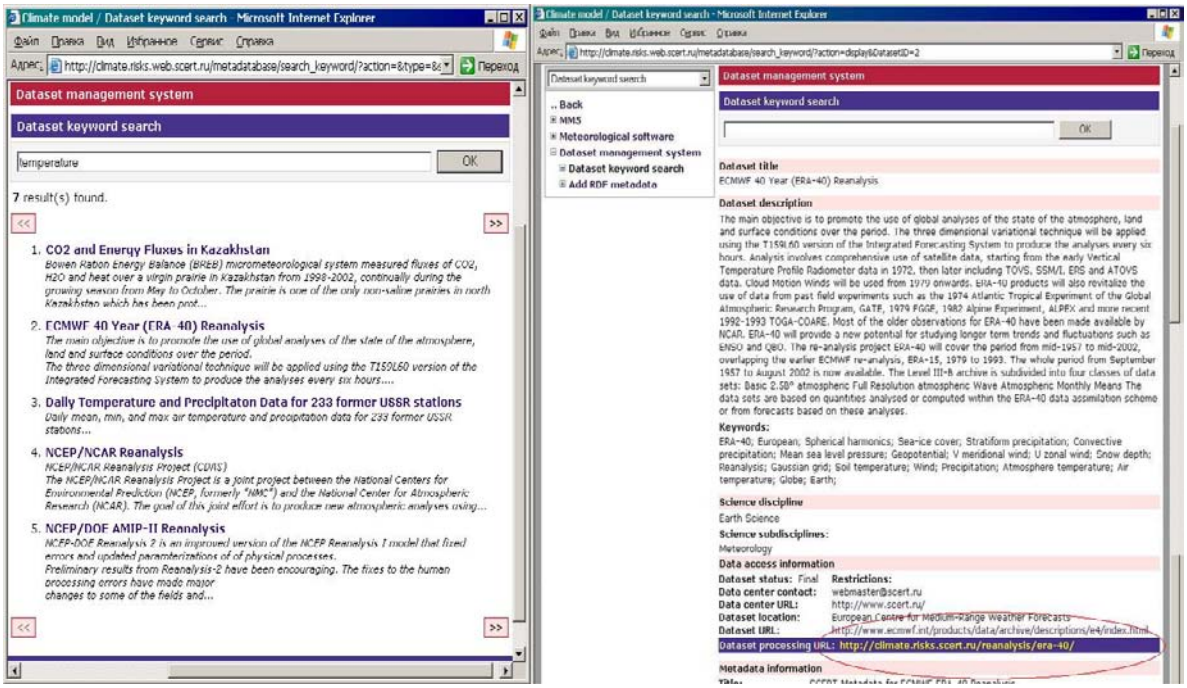

Рис. 3. Список найденных наборов данных и метаданные для реанализа ERA-40

Результат его работы в виде вычисленной поверхности индекса изменения климата представлен на рис. 4. Аналогичным образом на правой панели рис. 4 представлены результаты обработки данных о полях загрязнения воздуха в г. Томск, вычисленные с использованием набора данных, полученным в результате выполнения проекта РФФИ № 05-05-98010.

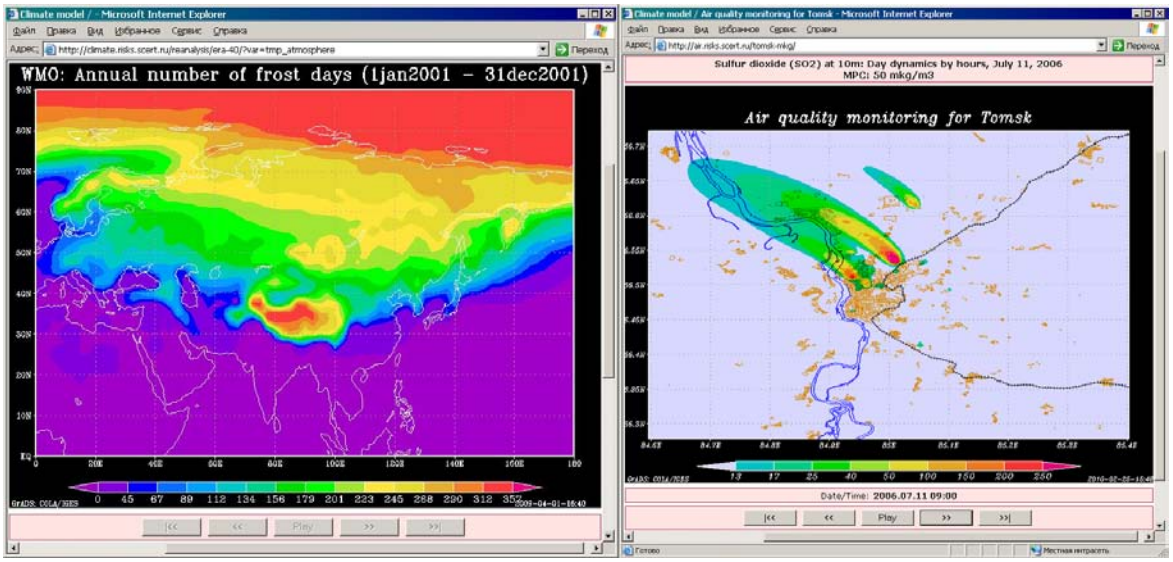

Рис. 4. Число дней с заморозками за 2001 г. по данным реанализа ECMWF ERA-40. Концентрация диоксида серы над г. Томск, 11.07.2006

Заключение

Предложенная архитектура модели системы позволяет использовать преимущества технологий Semantic Web при реализации схем метаданных и алгоритмов информационного поиска. Описанная система представляет собой необходимый элемент распределенной информационно-вычислительной среды для поддержки мультидисциплинарных исследований Сибири.

Список литературы

1. *Tschirner S., Zipf A.* Finding geodata that otherwise would have been forgotten- GeoXchange: a SDI-based portal for sharing free geodata // In Proceedings of the 2005 Workshop on Geographic information Retrieval (Bremen, Germany, November 04, 2005). ACM, New York, NY, 2005. P. 39–44.

2. *Mccartney P., Jones M.* Using XML-encoded Metadata as a Basis for Advanced Information Systems for Ecological Research // In Proceedings of the 6th World Multi-Conference on Systemics, Cybernetics, and Informatics (SCI 2002), Orlando, Florida, 2002.

3. *Wiegand N.* Searching for geospatial government-produced data // In Proceedings of the 2005 National Conference on Digital Government Research (Atlanta, Georgia, May 15–18, 2005). ACM International Conference Proceeding Series. 2005. Vol. 89. Digital Government Society of North America. P. 269–270.

4. *Max J. Egenhofer.* Toward the semantic geospatial web // In GIS ’02: Proceedings of the 10th ACM international symposium on Advances in geographic information systems. N. Y., USA: ACM Press, 2002. P. 1–4.

5. *Nambiar U., Ludaescher B., Lin K., Baru C.* The GEON portal: accelerating knowledge discovery in the geosciences // In Proceedings of the 8th Annual ACM international Workshop on Web information and Data Management (Arlington, Virginia, USA, November 10, 2006). WIDM '06. ACM, N. Y., 2006. P. 83–90.

6. *Decker S., Melnik S., Harmelen F. V., Fensel D., Klein M. C. A., Broekstra J., Erdmann M., Horrocks I.* The Semantic Web: The Roles of XML and RDF // In Proceedings of IEEE Internet Computing. 2000. P. 63–74.

7. *Guha R., McCool R., Miller E.* Semantic search // In Proceedings of the 12th international Conference on World Wide Web (Budapest, Hungary, May 20–24, 2003). WWW '03. ACM, N. Y., 2003. P. 700–709.

8. *Bhagdev R., Chapman S., Ciravegna F., Lanfranchi V., Petrelli D.* Hybrid search: Effectively combining keywords and semantic searches // The Semantic Web: Research and Applications. ESWC 2008, Tenerife, Spain, June 1–5. Springer, 2008. P. 554–568.

9. *Heflin J., Hendler J.* Semantic Interoperability on the Web // In Proceedings of Extreme Markup Languages 2000. Graphic Communications Association. 2000. P. 111–120.

10. *Титов А. Г.* RDF-схема для метаданных по метеорологии и климату // Измерения, моделирование и информационные системы для изучения окружающей среды / Под общ. ред. проф. Е. П. Гордова. Томск: Изд-во Томск. ЦНТИ, 2006. С. 58–61.

11. *Wood D., Gearon P., Adams T.* Kowari: A platform for semantic web storage and analysis // In XTech. 2005, May 2005.

12. *Harris S., Gibbins N.* 3store: efficient bulk RDF storage // In Proceedings of the First International Workshop on Practical and Scalable Semantic Systems (PSSS'03). 2003. P. 1–20; URL: http://www.eprints.aktors.org/archive/00000273/.

13. *Berners-Lee T.* Why RDF model is different from the XML model. September 1998; URL: http://www.w3.org/DesignIssues/RDF-XML.html

14. *Celino I., Della Valle E., Cerizza D., Turati A.* Squiggle: a Semantic Search Engine for indexing and retrieval of multimedia content // In Proceedings of the 1st International Workshop on Semantic-Enhanced Multimedia Presentation Systems (SEMPS-2006). Athens, Greece, December 6, 2006.

15. *Davies J., Weeks R., Krohn U.* QuizRDF: Search technology for the SemanticWeb // In Proceedings of the 37th Annual Hawaii International Conference on System Sciences (HICSS'04). BTexact Technologies, IEEE Computer Society, 2004. P. 112–119.

16. *Victoria S. Uren, Yuangui Lei, Enrico Motta*: SemSearch: Refining Semantic Search / The Semantic Web: Research and Applications / ed. Sean Bechhofer et al. 5th European Semantic Web Conference, ESWC 2008, Tenerife, Canary Islands, Spain, June 1–5, 2008. Proceedings. Vol. 5021 of Lecture Notes in Computer Science. Springer, 2008. P. 874–878.

17. *Li Ding, Tim Finin, Anupam Joshi, Rong Pan, Scott R. Cost, Yun Peng, Pavan Reddivari, Vishal Doshi*, and *Joel Sachs*. Swoogle: a search and metadata engine for the semantic web // In CIKM '04: Proceedings of the thirteenth ACM conference on Information and knowledge management, N. Y., USA: ACM Press, 2004. P. 652–659.

18. *Hildebrand M., Ossenbruggen J. van, Hardman L.* An analysis of search-based user interaction on the semantic web. Technical Report INS-E0706, Centrum voor Wiskunde en Informatica, Amsterdam, The Netherlands. May 2007.

19. *Page L., Brin S., Motwani R., Winograd T.* The pagerank citation ranking: Bringing order to the web // Technical report, Stanford Digital Library Technologies Project, 1998; URL: http://ilpubs.stanford.edu:8090/422/1/1999-66.pdf.

20. *Гордов Е. П., Окладников И. Г., Титов А. Г.* Разработка элементов информационно-вычислительной системы на основе веб-технологий для исследования региональных природно-климатических процессов // Вычислительные технологии. 2007. Т. 12, спец. вып. 3. С. 20–28.

21. *Broekstra J., Kampman A., Harmelen F. van*. Sesame: An Architecture for Storing and Querying RDF Data and Schema Information // Semantics for the WWW. MIT Press, 2001.

22. *Svab O., Svatek V., Kavalec M., Labsky M.* Querying the RDF: Small Case Study in the Bicycle Sale Domain // In Proceedings of the Dateso 2004 Annual International Workshop on Databases, Texts, Specifications and Objects. Desna, Czech Republic, April 14–16, 2004. Vol. 98. P. 84–95.

23. *Gordov E. P., Lykosov V. N., Fazliev A. Z.* Web-portal on environmental sciences «ATMOS» // Adv. Geosci. 2006. Vol. 8. P. 33–38; URL: www.adv-geosci.net/8/33/2006/.

24. *Barbera M., Di Donato F., Morbidoni C., Tummarello G.* Hyperjournal software, php scripting and semantic web technologies for the open access // In Proceedings ESWC: European Semantic Web Conference (2nd. 2005. Heraklion). Heraklion (Greece), 2005.

25. *Окладников И. Г., Титов А. Г., Мельникова В. Н., Шульгина Т. М.* Веб-система для обработки и визуализации метеорологических и климатических данных // Вычислительные технологии. 2008. Т. 13, спецвып. № 3. С. 64–69.

Материал поступил в редколлегию 06.02.2010

A. G. Titov, E. P. Gordov, I. G. Okladnikov

APPLICATION OF SEMANTIC WEB TECHNOLOGIES IN THE INFORMATION-COMPUTATIONAL SYSTEM FOR ENVIRONMENTAL DATA ANALYSIS

The working model of the software system for storage, semantically-enabled search and retrieval along with processing and visualization of environmental datasets containing results of meteorological and air pollution observations and mathematical climate modeling is presented. Specially designed metadata standard for machine-readable description of datasets in the form of RDF Schema is introduced. At present the system is available as a Web server and is implementing dataset management functionality including SeRQL-based semantic search as well as statistical analysis and visualization of selected data archives. The proposed system represents a step in the process of development of collaborative information-computational environment to support multidisciplinary investigations of Earth regional environment.

Keywords: Information systems, metadata, Semantic Web, meteorology, climate change.