ISBN 978-5-19601-104-3

Федеральное государственное бюджетное учреждение науки
ИНСТИТУТ КОСМИЧЕСКИХ ИССЛЕДОВАНИЙ
РОССИЙСКОЙ АКАДЕМИИ НАУК
Федеральное государственное бюджетное учреждение науки
ВЫЧИСЛИТЕЛЬНЫЙ ЦЕНТР им. А.А. ДОРОДНИЦЫНА
РОССИЙСКОЙ АКАДЕМИИ НАУК

ИНФРАСТРУКТУРА НАУЧНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ И СИСТЕМ

Сборник избранных научных статей

Под редакцией
доктора техн. наук Е.Б. КУДАШЕВА,
доктора физ.-матем. наук В.А. СЕРЕБРЯКОВА

Том II

ВЦ РАН

Москва
2014

УДК [002:004.9] (063)
ББК [73+32.973.233]я43

Инфраструктура научных информационных ресурсов и систем. Сборник избранных научных статей. Труды Четвертого Всероссийского симпозиума (С.-Петербург, 6–8 октября 2014 г.). Под ред. Е.В. Кудашева, В.А. Серебрякова. В 2-х тт. Т. 2. М.: ВЦ РАН, 2014.

Симпозиум проводится ежегодно с 2011 г. по Плану научных конференций Отделения математических наук РАН. В 2011 г. и 2012 г. Симпозиум проводился в С.-Петербурге при поддержке РФФИ, в 2013 г. – в Абхазии, г. Сухум – при поддержке Академии наук Абхазии. В 2014 г. Симпозиум проводился в С.-Петербурге при поддержке РФФИ на базе Петербургского Отделения Математического института им. В.А. Стеклова РАН – ПОМИ РАН.

Научная программа Симпозиума «**Инфраструктура научных информационных ресурсов и систем»** ориентирована на рассмотрение проблем и перспектив развития информационно-телекоммуникационных систем; методов, технологий и средств применительно к доступу, хранению и интеллектуальному анализу данных в различных областях фундаментальной науки, разработки информационных систем для научных исследований.

Основные цели Четвертого Симпозиума: методы и технологии интеграции электронных коллекций; взаимодействия информационных ресурсов и формирования электронного документного пространства научных исследований и инноваций, развитие электронных библиотек.

The symposium is held annually since 2011 on the Plan of scientific conferences Department of Mathematical Sciences of RAS. In 2011 and 2012 Symposium was held in St. Petersburg and was supported by RFBR, in 2013 – in Abkhazia, Sukhum – with the support of the Academy of Sciences of Abkhazia. In 2014, the Symposium is held in St. Petersburg on the basis of the St. Petersburg Branch of the Mathematical Institute of the Academy of Sciences – PDMI RAS and is supported by RFBR.

The scientific program of the Symposium is oriented to the infrastructure of scientific information resources and systems geared to the problems and prospects of development of information and telecommunication systems; methods, tools and technology with respect to access, storage, and data mining in various fields of basic science, development of information systems for research.

The main objectives of the Fourth Symposium: methods and techniques of integration of digital collections; interaction of information resources and the generation of the electronic document space research and innovation, the development of digital libraries.

Рецензенты: Г.Н. Заварза, К.Б. Теймуразов

Научное издание

ОГЛАВЛЕНИЕ

ИНФОРМАЦИОННОЕ ПРЕДСТАВЛЕНИЕ ЗАГРЯЗНЕНИЯ ГОРОДСКОГО ВОЗДУХА ИСТОЧНИКАМИ АНТРОПОГЕННОЙ И БИОГЕННОЙ ЭМИССИИ

А.А. Барт[a], *А.В. Старченко*[a], *Д.В. Царьков*[b], *А.З. Фазлиев*[c]

[a] Национальный исследовательский Томский государственный университет

[b] School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK

[c] Институт оптики атмосферы им. В.Е. Зуева СО РАН

bart@math.tsu.ru, starch@math.tsu.ru, dmiry.tsarkov@gmail.com, faz@iao.ru

В работе рассмотрена часть городской информационной модели, представляющая состояния атмосферного пограничного слоя над городом. Решения системы уравнений, характеризующих перенос примеси, представлены в форме онтологической базы знаний. Дано описание метрики созданной онтологии.

Ключевые слова: загрязнение воздуха, онтологическое описание загрязнений

In this paper we consider part of the urban information model that represents the state of the atmospheric boundary layer over the city. Description of solutions of a system of equations describing the transport of pollutants is presented in the form of ontological knowledge base. A description of the ontology metric is given.

Keywords: urban air pollution, ontological description of air pollution

Введение

Загрязненность атмосферного воздуха является известным фактом. Поступление загрязнителей в атмосферу прекратить невозможно, особенно загрязнителей биогенного происхождения, но контроль над процессом необходим. Для контроля необходимs

явное понимание процессов, протекающих в атмосфере, и, информация об источниках загрязнения (их форме, расположении и продолжительности действия).

В настоящее время многочисленные математические модели, развитые для применения к городам и относящиеся к качеству городского воздуха, анализу энергетических затрат, дорожному движению, системы автоматизированных датчиков и т.д., применяются при планировании в городском масштабе в городах Европейского Союза [1–6]. Набор этих моделей, относящихся к разным предметным областям и разномасштабным задачам городского планирования, привел к сложным взаимосвязям в 3D моделях города [1, 5], и существенно облегчил процедуру поддержки решений. Основанный на онтологиях подход обеспечил универсальный и надежный способ взаимосвязи и интеграции моделей, характеризующих разные аспекты развития города.

Модели качества воздуха связаны со сложными процессами, для описания которых требуются много параметров, относящихся к источникам загрязнения, превалирующему ветру или геометрии расположения домов и улиц. В работе [1] описана онтология модели качества воздуха, ориентированная на ситуацию в которой городские каньоны задерживают загрязнители. Эта онтология является частью системы онтологий, в число которых входят также онтология CityGML (модель представления 3D моделей городов, основанная на стандарте GML3 OGS), онтология транспортной системы и онтология процесса городского планирования [1,4,5].

В работе [1] отмечено, что при реализации сложных автоматизированных городских систем, включающих в себя математические модели физических и химических процессов и способы принятия решений, в частности, анализ качества воздуха, важным является применение онтологий, содержащих в себе, как правило,

формализованное описание входных и выходных данных приложений, используемых в потоках работ, определяющие функциональные возможности таких систем. При интеграции в городскую модель потоков работ роль онтологий состоит в описании специфики соответствующих задач, тогда как широко используемая концептуальная схема связана только с описание контента базы данных.

1. ИВС «UnIQuE»

Созданная в ТГУ информационно-вычислительная система «UnIQuE» предназначена для вычисления концентраций примесей, загрязняющих воздух в атмосферном пограничном слое города и представления свойств результатов вычисления в форме данных, информации и знаний для их предоставления исследователям и программным агентам.

В основе системы лежит программное обеспечение, выполняющее решение уравнений математической модели переноса примесей с учетом химических реакций [7,8]. При численном решении, изменения концентраций веществ за счет химических реакций, описывается одним из двух механизмов: GRS[9] или DMI [10]. Основным источником антропогенного загрязнения в городах является транспорт. При моделировании выбросов от автотранспорта учитывается изменение интенсивности трафика в течение суток и средние характерные выбросы автотранспортом. Заводы и промышленные объекты также вносят свой вклад в загрязнение воздуха, выбрасывая из труб продукты горения, причем температура выброса отличается от температуры окружающей среды [11]. При моделировании учета выбросов от предприятий учитываются расположение источника, состав выбрасываемого загрязнителя и скорость выброса. Для моделирования по-

ступления изопрена от лесных массивов используется механизм, предложенный в модели MEGAN [12].

В ИВС автоматически вычисляются значения свойств, характеризующих предсказанные данные о состоянии пограничного слоя. Эти свойства описываются на языке OWL 2 DL в рамках семантического подхода. Доменом, или областью применения этих свойств являются уровни пограничного слоя атмосферы. Описание уровней пограничного слоя имеет конечной целью построение фактологической части онтологии. Созданная онтология представляет логическую теорию, описывающую уровни атмосферного пограничного слоя над городом.

Понятийная часть онтологии содержит таксономию классов. В ИВС значимые для анализа ситуаций уровни пограничного слоя отнесены к определенным классам. Примером такого отнесения является ситуация проверки достоверности прогностических концентраций пяти разных примесей с синхронно измеренными концентрациями. В зависимости от того попадают ли сравниваемые величины в интервал значений измеряемой концентрации, определяемого погрешностью измерений.

Машинная систематизация уровней пограничного слоя и их частей по величине концентраций примесей и по степени совпадения измеренных и предсказанных данных проводится с помощью расширенной версии машины вывода FaCT++ [13]. Расширение машины вывода позволяет относить индивиды онтологии к классам, содержащие пары идентичных по типу индивидов, для которых разность значений выделенного свойства меньше заданной величины.

2. Трехслойная архитектура ИВС «UnIQuE»

В рамках подхода Semantic Web (SW) информационные ресурсы включают в себя данные, информацию и знания. Следование складывающейся терминологии SW, приводит к использова-

нию терминов «данные», «связанные данные» и «онтологии». В SW интерпретация терминов «данные», «связанные данные» и «онтологии» тесно связана с семантикой формальных языков (XML, RDF и OWL). Более того, такой интерпретацией обусловлено использование терминов «слой данных и приложений», «информационный слой» и «слой знаний», введенных в e-Science [14] для описания инфраструктуры информационных ресурсов и информационных систем.

Детали представления ИВС «UnIQuE» в трехслойной архитектуре показаны на рис. 1.

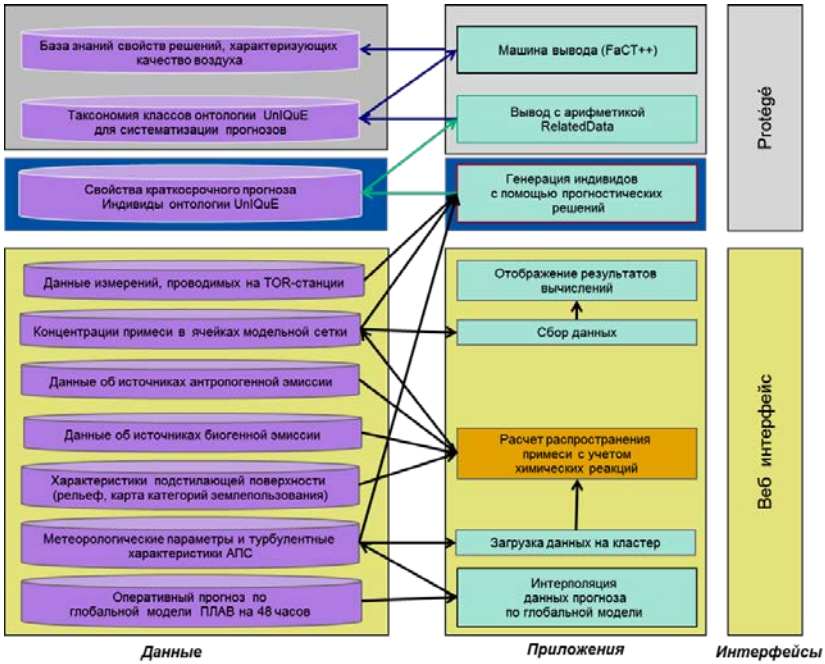

Рис.1. Трехслойная архитектура ИВС «UnIQuE».

Вычисленные решения по численной модели переноса [7, 8] представляют собой наборы данных, обеспеченные структурными метаданными. В число этих метаданных входят время, на которое вычислено решение, номер уровня в атмосферном пограничном слое, химическая модель в рамках которой проводились вычисления и т.д. Значимым для анализа прогноза (предсказания) является уровень пограничного слоя атмосферы. В математической модели для обоих механизмов химических трансформаций веществ рассмотрено 28 уровней. Изучение динамики изменений концентраций примеси осуществляется с периодом в один час.

Информационный слой ИВС «UnIQuE» содержит описание свойств решения задачи почасового прогноза качества воздуха, вычисленного в слое данных. В этом слое формируются индивиды, обладающие этими свойства, их значения и индивиды представляются в форме субъектно-предикатных структур [15]. Целью создания такого слоя является формирования фактологической части онтологии «UnIQuE», представляющей собой логическую теорию о свойствах решений задачи прогноза качества воздуха. Понятийная и фактологическая части онтологии образуют онтологическую базу данных, которую можно использовать для принятия решений.

Информационный слой в ИВС «UnIQuE» содержит факты, характеризующие свойства решений задач переноса примесей и химии газофазных реакций, размещенных в слое данных. Ключевым моментом в формировании фактологической части онтологии является создание индивидов. Индивиды представляют собой наборы фактов (высказываний), предназначенных для каталогизации информационных ресурсов, описывающих прогноз качества воздуха над городом и его пригородами. Структура индивидов позволяет проводить систематизацию индивидов по набору их свойств. В ИВС все значения свойств вычисляемых индивидов формируются приложениями автоматически после получения решения задачи прогноза качества воздуха.

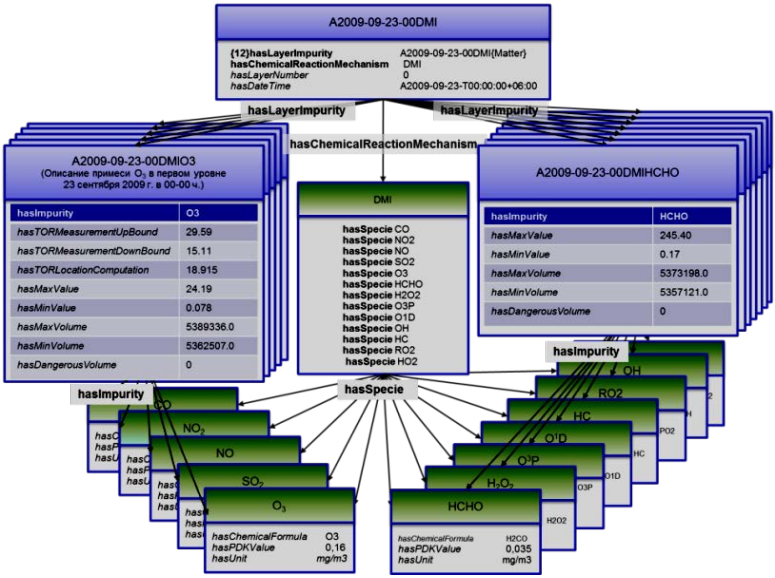

Рис. 2. Субъектно-предикатная структура, характеризующая свойства краткосрочного прогноза качества воздуха (DMI) в первой страте пограничного слоя.

На рис. 2 показана субъектно-предикатная структура, отвечающая индивиду, используемому для описания уровня пограничного слоя атмосферы. Для описания свойств уровня нужны 12 индивидов, каждый из которых описывает свойства молекулы или группы веществ. Название индивида, относящегося к описанию уровня и описанию свойств примеси, формируются динамически. Структуры названия индивидов имеют вид: {A}(YYYY-MM-DD-HH)-(DMI) и {A}(YYYY-MM-DD-HH)-(DMI)(Примесь), где {A} кодирует номер слоя, а YYYY-MM-DD-

НН характеризует дату и час предсказания. Параметр «Примесь» принимает значения соответствующие веществам, используемым в химических моделях DMI и GRS. Особо стоит выделить группу веществ (O_3, SO_2, NO, NO_2, CO) концентрации которых измеряются на TOR станции ИОА СО РАН. Эти измерения связаны с самым нижним уровнем атмосферного пограничного слоя. Для индивида, описывающего свойства нижнего уровня, используется пара свойств

(hasTORMeasurementUpBound,

hasTORMeasurementDownBound, hasTORLocationComputation)

значениями которых 12лассются нижнее и верхнее значения концентрации измеренной величины, а также значение вычисленной концентрации соответствующей примеси. Свойства, выделенные наклонным шрифтом, относятся к конкретным свойствам (Datatype Property), а свойства, выделенные утолщенным шрифтом, соответствуют объектным свойствам (Object Property). Оба этих типа используются в языках OWL DL и OWL 2 DL.

Индивид, характеризующий уровень пограничного слоя, обладает 14 свойствами. Двенадцать свойств (**hasLayerImpurity**) описывают вещества, трансформация которых учитывается в химической модели. Эта модель характеризуется свойством **hasChemicalReactionMechanism** (используемый при моделировании механизм химических реакций), дата предсказания – свойством (hasDateTime) и номер уровня – свойством (hasLayerNumber).

Индивид, описывающий примеси в заданном уровне, обладает семью свойствами: hasMaxValue, hasMinValue, hasMaxVolume, hasMinVolume, hasDangerousVolume и **hasImpurity**. Пара свойств hasMaxValue, hasMinValue характеризует минимальное и максимальное значение концентрации примеси в уровне пограничного слоя, а пара hasMaxVolume, hasMinVolume – значение

объема воздуха, содержащего минимальное и максимальное значение концентрации примеси. Значением свойства hasDangerousVolume определяется суммарный объем воздуха в уровне пограничного слоя, в котором концентрация примеси превышает ПДК. Значение объектного свойства **hasImpurity** представляет индивид, описывающий примесь.

Описанная фактологическая часть онтологии применима в тех ситуациях, для анализа которых существенны минимальные и максимальные значения концентраций примесей, важна стратификация уровней по значениям концентраций примесей и достоверность прогностических концентраций.

Основным ресурсом слоя знаний являются таксономии онтологии «UnIQuE». Они ориентированы на использование в рамках агентных технологий, прежде всего в тех случаях, когда возникают ситуации, требующие автоматического принятия решения при выборе одного ресурса из нескольких ему подобных.

Основным назначением слоя знаний в ИВС «UnIQuE» является систематизация свойств решений задач краткосрочного прогноза качества воздуха, в частности, предоставление пользователю возможности семантического поиска достоверного прогноза качества воздуха с высокой степенью доверия к нему.

В качестве примера автоматического отнесения рассмотрено формирование классов, содержащих уровни с недостоверными значениями концентраций примесей. Критерием достоверности расчета является выполнение неравенства:

$$\{A_{pred}{}^{S} : A_{meas}{}^{S} - D^{S} < A_{pred}{}^{S}\} \cap \{A_{pred}{}^{S} : A_{pred}{}^{S} < A_{meas}{}^{S} + D^{S}\}$$

в котором $A_{meas}{}^{S}$ и $A_{pred}{}^{S}$ значения измеренной и предсказанной концентраций вещества S, а D^{S} погрешность измерения концентрации вещества S.

Слой знаний таксономии классов (T-box) и свойств (R-box). Цель слоя знаний состоит в обеспечении 1Зласс1Ззации намере-

ний исследователя при выборе одного из двух прогнозов, предоставляемых ИВС, другими словами, созданная база знаний является основой для формирования экспертной системы и системы принятия решений.

Созданные в информационном слое классы и свойства в слое знаний кодируются в соответствии с правилами языка спецификации онтологий OWL DL2. Особенностью является использование в работе машины вывода арифметических операций. Примером, демонстрирующим эту технологию, является формирование классов, содержащих индивиды, характеризующие расчетные данные, согласующиеся (**ConsistentMeasured_and_ComputedData**) и не согласующиеся (**InconsistentMeasured_and_ComputedData**) с данными измерений, проведенных на TOR-станции ИОА СО РАН.

Для наполнения онтологии фактами создано прикладное программное обеспечение, состоящие из трех программных модулей, выполняемых последовательно друг за другом. Первый программный модуль, написанный на языке Fortran, осуществляет чтение результатов численных расчетов концентраций компонент примеси и метеорологических характеристик из двоичного файла прямого доступа и вычисляет максимальные, минимальные и превышающие ПДК значения и объемы воздуха в уровне, занимаемые интересующими значениями. Для приземного уровня атмосферного пограничного слоя дополнительно рассчитывается значения концентраций озона, монооксида и диоксида азота, диоксида серы и угарного газа для точки, координаты которой соответствуют координатам TOR-станции ИАО СО РАН. Эти данные служат для сравнения прогноза с данными измерений. Рассчитанные значения и объемы используются во втором приложении, написанном на языке PHP с использованием технологии DOM. Приложение строит индивиды для онтологии на основе синтаксиса rdf и описывает свойства этих индивидов посредст-

вом тегов. Результатом работы приложения является owl-файл. Третье приложение используется только в случае, когда расчет проводится по исторической дате (не прогноз) и имеются данные измерений. Приложение работает с онтологией и использует библиотеку арифметических операций машины вывода Fact++. В процессе выполнения создаются новые классы, содержащие индивиды, в которых вычисленные значения попадают или не попадают в интервал погрешности измерений. В настоящее время библиотека арифметических операций позволяет сравнивать datatype свойства в рамках одного индивида, поэтому вычисление объема воздуха, в котором наблюдается превышение ПДК, проводится на этапе предварительной подготовки данных. После завершения работы машины вывода пользователю становится доступен owl-файл, характеризующий свойства решения.

3. Заключение

В работе описана трехслойная архитектура информационно-вычислительной системы «UnIQuE». Особенностью созданной системы является представление результатов расчетов в новой форме – онтологической базы знаний, что в свою очередь позволяет использовать машину вывода для получения новых знаний о предметной области на основе численных результатов моделирования.

Литература

1. C.Metral, G.Falquet, A.F.Cutting-Decelle, Towards semantically enriched 3D city models:An ontology-based approach, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.151.46
2. V. Di Lecce, C. Pasquale, and V. Piuri, A basic ontology for multy agent system communication in an environmental monitoring system, CIMSA 1004 – International Conf. on computational intelligence for

measurement systems and applications, Boston, MA, USA, 14-16 July 2004, pp.45-50
3. C.Netral, G.Falquet, M.Vonlanthen, An ontology-based model for urban planning communication, Studies in computational intelligence, v.61, pp.61-72, 2007.
4. C.Metral, R.Billen, A.-F. Cutting-Decelle, M.Van Ruymbeke, Ontology-based models for improving the interoperability of 3D urban information, Journal of information technology in construction, v.15, pp.169-184, 2010.
5. C.Mental, G.Falquet, K.Karatzas, Ontologies for the integration of air quality models and 3D city models, arXive: 1201.6511 [CsAI].
6. M.M.Oprea, AIR_POLLUTION_Onto: an ontology for air pollution analysis and control, in: Artificial intelligence applications and innovations III, Ed.: Iliadis I., Vlahavas I., Bramer M., Boston, Springer, pp.135-143.
7. Барт А.А., Беликов Д.А., Старченко А.В. Математическая модель для прогноза качества воздуха в городе с использованием суперкомпьютеров // Вестник Томского государственного университета. Математика и механика. – 2011. – № 3. – С. 15-24.
8. Барт А.А., Старченко А.В., Фазлиев А.З. Информационно-вычислительная система для кратко-срочного прогноза качества воздуха над территорией г. Томска // Оптика атмосферы и океана. – 2012. – Т. 25, № 7. – С. 594-601.
9. Hurley, P. J. The Air Pollution Model (TAPM) Version 2 / P. J. Hurley // CSIRO Atmospheric Research Technical Paper. – 2002. – No. 55. – P. 37.
10. Stockwell, W.R. Comment on «Simulation of a reacting pollutant puff using an adaptive grid algorithm» by R. K. Srivastava et al. / W. R. Stockwell, W. S. Goliff //J. Geophys. Res. – 2002 – Vol. 107. – pp. 4643-4650.

11. Берлянд, М. Е. О расчете загрязнений атмосферы выбросами их дымовых труб электростанций / М. Е. Берлянд, Е. Л. Генрихович, Р. И. Оникул // Труды Геолого-географического общества. – 1964. – вып. 158. – С. 3-21.
12. Guennther, A. B. The Model of Emissions of Gases and Aerosols from Nature version 2.1 (MEGAN2.1): an extended and updated framework for modeling biogenic emissions / A. B. Guennther, X. Jiang, C. L. Heald, T. Sakulyanontvittaya, T. Duhl, L. K. Emmons, X. Wang // Geosci. Model Dev. –2010. –Vol. 5. – pp. 1471-1492.
13. Tsarkov D., Horrocks I. FaCT++ Description Logic Reasoner: System Description // Int. Joint Conf. on Automated Reasoning (IJCAR 2006). 2006. Vol. 4130. pp. 292-297.
14. Berners-Lee, T. The Semantic Web / T. Berners-Lee, J. Hendler, O. Lassila // Scientific American. – 2001.
15. Зиновьев А.А. Основы логической теории научных знаний. – М: Наука, 1967. 260 с.

ОНТОЛОГИЧЕСКОЕ ОПИСАНИЕ СОСТОЯНИЙ И ПЕРЕХОДОВ В КОЛИЧЕСТВЕННОЙ СПЕКТРОСКОПИИ

С. Воронина[a], *А. Привезенцев*[a], *Д.В. Царьков*[b], *А.З. Фазлиев*[a]

[a]Институт оптики атмосферы СО РАН,
пл. Академика Зуева 1, Томск 634055, Россия

[b] School of Computer Science, The University of Manchester, Oxford Road, Manchester, M13 9PL, UK

vss@iao.ru, remake@iao.ru, dmitry.tsarkov@gmail.com, faz@iao.ru

В докладе развито онтологическое описание состояний и переходов молекул в количественной спектроскопии. При таком описании каждом состоянию и переходу с определенными квантовыми числами, соответствуют все опубликованные значения физических величин, относящимися к шести спектроскопическим задачам. Проведено сравнение созданной онтологии с онтологией информационных спектральных ресурсов.

Ключевые слова: состояния и переходы в молекулярной спектроскопии, онтологическое описание молекул

An ontological description of molecular states and transitions used in quantitative spectroscopy is developed. Each state or transition is identified by quantum numbers are characterized by a complete set of published values of physical quantities related to six spectroscopic tasks. A comparison of the ontology with ontology of information spectroscopic resources is given.

Keywords: states and transitions in quantitative spectroscopy, ontological description of molecules

1. Введение

Традиционная процедура получения данных о спектральных характеристиках молекул, применяемая исследователями из прикладных по отношению к количественной спектроскопии предметных областей, связана с использованием публикаций, и, в последнее время, с информационными системами, опирающимися на базы данных (см., например, [1-3]). Для большинства исследователей критерием применимости этих данных является правдоподобность получаемых ими результатов. Такое отношение к данным приводит к тому, что вопросы достоверности и доверия к данным остаются на втором плане, т.к. исследователей некоторых предметных областей устраивает даже часть их, связанная с сильными переходами. Прямая оценка доверия экспертным данным для исследователей является трудоемкой задачей в силу того факта, что большинству из них не доступен полный набор опубликованных и согласованных спектральных данных.

Первое системное изучение полного набора опубликованных данных было сделано в цикле работ (см., например, [4-7] для молекулы воды с целью получения эталонных уровней энергии, с точностью характерной для измерений). Формальное информационное описание полного набора данных было проведено в диссертации [8]. Результатом исследования явилось онтологическое описание информационных ресурсов для шести задач спектроскопии. Позже такое описание было дано для молекулы CO_2 (см., например, [9,10]). Эти работы позволили получать ответы на группы вопросов относящихся к качеству источников спектральных данных, извлеченных из публикаций, по молекулам воды и диоксида углерода.

Практика показала, что использование онтологии информационных ресурсов количественной спектроскопии [8-10] позволило найти ответы на вопросы о том, какие физические величины на-

ходятся в публикациях, насколько качество значений этих величин, но оно не дало ответов на вопросы о конкретных характеристиках состояний и переходов и их согласованности в разных публикациях, и, в частности, в базе данных, входящей в ИС W@DIS.

В данной работе рассмотрены индивиды трех типов, предназначенные для описания состояний и переходов в количественной спектроскопии. Результаты анализа качества переходов размещены в ИС W@DIS в онтологии состояний и переходов.

Краткое описание коллекции состояний и переходов молекул

Коллекция состояний и переходов молекул появилась как следствие участия А.Ф. в работе группы данных по молекуле воды. Основные усилия в области систематизации данных были направлены на сбор и анализ качества спектральных данных изолированной молекулы. Параллельно этим работам в ИОА СО РАН велись работы по систематизации спектральных данных молекул в газовой фазе и расширению числа молекул, для которых были собраны практически все опубликованные спектральные данные.

В качестве примера в Таб. 1 показана статистика опубликованных переходов для молекул воды и диоксида углерода. С целью единого подхода к систематизации, мы связали характеристики переходов и состояний с решениями шести задач спектроскопии (Т1-Т3, Т5-Т7). Использованные в Таб.1 в качестве маркеров задачи Т2 и Т6, соответствуют прямой и обратной задаче нахождения характеристик переходов изолированной молекулы или рассчитанных и измеренным характеристикам переходов, соответственно. К

числу основных характеристик перехода изолированной молекулы отнесены вакуумные волновые числа и коэффициент Эйнштейна. Разбиение T6 на T6(a) и T6(b) используется для выделение общего числа измеренных переходов и числа уникальных переходов, содержащихся в коллекции по определенной молекуле.

Таблица 1. Число вычисленных (T2) и измеренных (T6) переходов молекул воды и диоксида углерода в ИС W@DIS.

Molecule	T2	T6(a)	T6(b)	Molecule	T2	T6(a)	T6(b)
CO_2	4873448	39485	25704	H_2O	3153126	179014	64884

Из табл. 1 следует, что число рассчитанных переходов намного превышает число измеренных переходов. В этой работе для формирования онтологии состояний и переходов использованы только часть рассчитанных переходов. В эту часть входят переходы идентичные имеющимся измеренным переходам.

2. Информационная модель количественной спектроскопии

В данной работе описание предметной области **Quantitative Spectroscopy** основано на использовании упрощенных моделей предметных областей **Molecular Spectroscopy, Thermodynamical Conditions, Bibliography и Mathematical Relations**. Фактологическая часть (A-box) предметной области **Quantitative Spectroscopy** представлена наборами описаний конкретных переходов и состояний (далее, описание конкретного описания состояния или перехода будем называть индивидом состояния или перехода) типа **Molecule_State_QNS**, **Molecule_Transition_QNT** и **Molecule_LineProfile_QNLP**. Эти индивиды описывают части публи-

каций, содержащие решения шести задач спектроскопии о переходах и состояниях изолированных молекул и молекул в газовой фазе. Упрощенная структура индивидов, позволяющая выделить, как общие для них черты, так и особенности, показана на рис.1 на примере описания, относящегося к молекуле воды. В структуру индивидов входят квантовые числа (QN_NM), характеристики, описывающие тип задачи (например, T7), метод решения задачи и тип источника данных (например, первичный, т.е. источник содержащий решение одной задачи с указанием деталей расчета или измерений), значения физических величин (уровень энергии, волновые числа, коэффициент Эйнштейна и т.д.), публикация из которой извлечены данные и бинарные отношения между идентичными физическими величинами и т.д.. В индивид **Molecule_LineProfile_QNLP** дополнительно входят физические величины, описывающие форму контура спектральной линии и термодинамические условия, соответствующие условиям расчета или измерения.

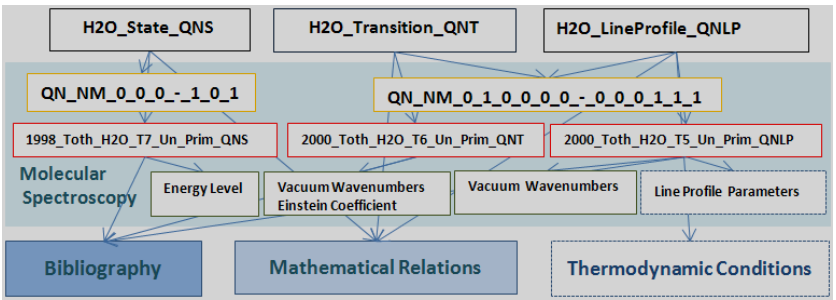

Рис. 1. Структура индивидов предметной области **Количественная спектроскопия.**

Здесь **1999_Toth** идентификатор публикации, из которой извлечены данные, **H2O** –и молекула состояние которой описывается, **Un –** сокращение, соответствующее «метод неизвестен», **Prim –** сокращение, соответствующее типу источника данных, в данном случае «Primary» и **QNS –** квантовые числа состояния (**T** – переход, **LP** – профиль линии).

В индивидах типа **Molecule_LineProfile_QNLP** используются более 20 типов контуров спектральных линий, но в каждом конкретном индивиде такого типа содержатся факты о физических величинах, относящихся к одному типу контура. Часть опубликованных значений физических величин не включена в индивиды. В число таких физических величин входят те, которые относятся к переходам и состояниям, квантовые числа которых не удовлетворяют ограничениям на состояния или правилам отбора, или квантовые числа которых имеют дубликаты в рамках публикации или они не идентифицированы полностью.

Описание публикации дано стандартным способом. Часть библиографических свойств публикации опущена, в числе таких свойств, например, место работы авторов, подробная дата публикации и т.д.. К числу математических отношений, используемых в описании индивидов, относятся максимальные разности физических величин и среднеквадратичные значения. Эти отношения используются как ко всем идентичным физическим величинам, так и к паре групп. Группы формируются из рассчитанных, измеренных или эталонных физических величин.

3. Особенности описания состояний и переходов

Квантовые числа состояний и переходов представлены в индивидах в разных видах: индивидуально и в группах (например, для представления нормальных мод в виде колебательных полос и вращательных квантовых чисел состояний и переходов). Физи-

ческими величинами изолированной молекулы, характеризующими состояние, является уровень энергии, а характеризующими переход – вакуумные волновые числа и коэффициенты Эйнштейна. Спектральные величины изолированной молекулы показаны в явном виде на Рис.1 (**Energy Level, Vacuum Wavenumbers, Einstein Coefficient**), а спектральные величины, зависящие от термодинамических условий, связаны с индивидами группы **Line Profile Patrameters**. В ИС W@DIS используется чуть более десятка контуров, по которым у нас имеются данные измерений. Так, например, для молекулы диоксида углерода наряду с лорентцевским контуром в ИС есть данные по контурам Раутиана, Розенкранца, Гэлатри, Фойгта и т.д.

Индивиды состояний и переходов изменяются во времени. Причиной изменения являются новые значения измеренных или рассчитанных физических величин. Неизменной частью индивидов состояния и перехода являются только индивиды, соответствующие описанию квантовых чисел.

4. Детальная структура описания состояний и переходов

В этом параграфе описывается структура индивидов, относящихся к состояниям и переходам молекул. В качестве языка представления состояний и переходов используется язык OWL DL 2. При описании используются два вида свойств: объектные и конкретные (datatype). На Рис.2-4 эти свойства обозначаются следующим образом: объектные - прямым шрифтом, а конкретные – наклонным шрифтом. Прямоугольниками обозначены индивиды. Первая строка индивида содержит имя индивида. Остальные строки перечисляют свойства, которыми обладает данный индивид. Индивиды сгруппированы по предметным областям. Например, на Рис.2 таких предметных областей две. Все индивиды **Molecule_State_QNS, Molecule_Transition_QNT** и **Mole-**

cule_LineProfile_QNLP обладают свойствами, характеризующими квантовые числа состояния или перехода. Оба этих свойства являются функциональными и обратно функциональными. Это означает, что значения этих свойств уникальны и по ним можно однозначно определить соответствующее состояние или переход.

На рис.2 дана подробная структура библиографической части описания и бинарных отношений между идентичными физическими величинами. На рис. 3-4 эти части представлены в редуцированном виде.

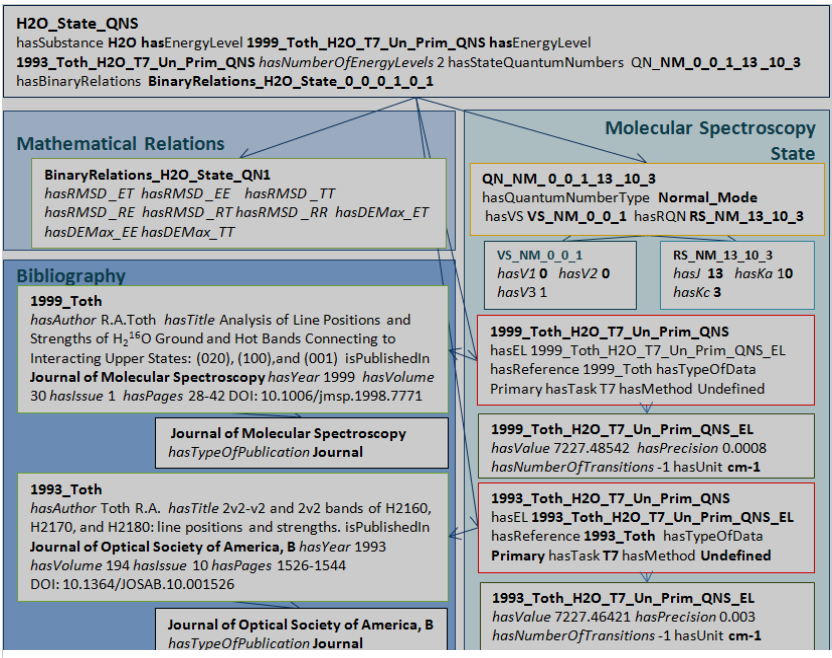

Рис. 2. Структура индивида состояния молекулы.

Здесь *hasRMSD_ET* - среднеквадратическое значение (RMSD) между «измеренными» (E) и расчетными (T) уровнями энергии,

hasRMSD _RE - среднеквадратическое значение (RMSD) между эталонными (R) и «измеренными» (E) уровнями энергии, *hasDEMax_ET* - максимальное значение разности пар идентичных уровней энергии одно из которых измерено (E), а другое рассчитано (T) b QNS = 0_0_1_-_13_10_3.

На рис.3 в структуре имен индивидов использованы следующие обозначения QNT = 0_1_0_0_0_0_-_0_0_0_1_1_1, T6_Un_Prim – индивид связан с задачей T6, эта задача решена неизвестным методом (Undefined), Primary – источник данных связанный с этим переходом является первичным, VWN-волновые числа, EC– коэффициент Эйнштейна.

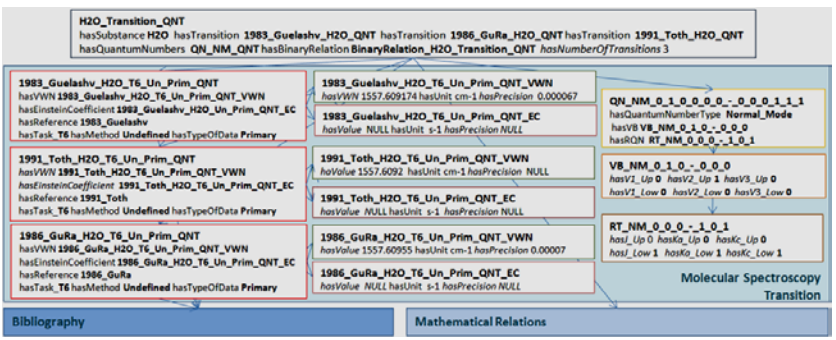

Рис. 3. Структура индивида перехода изолированной молекулы.

На рис.4 NM – представление квантовых чисел (Normal modes), QNLP = 0_1_0_0_0_0_-_0_0_0_1_1_1_1_Lorentz_N2_296-1 – квантовые числа перехода, тип контура, уширяющее вещество, значение температуры (K) и давления (атм), TDShift – температурная зависимость сдвига, TDHW – температурная зависимость столкновительной полуширины.

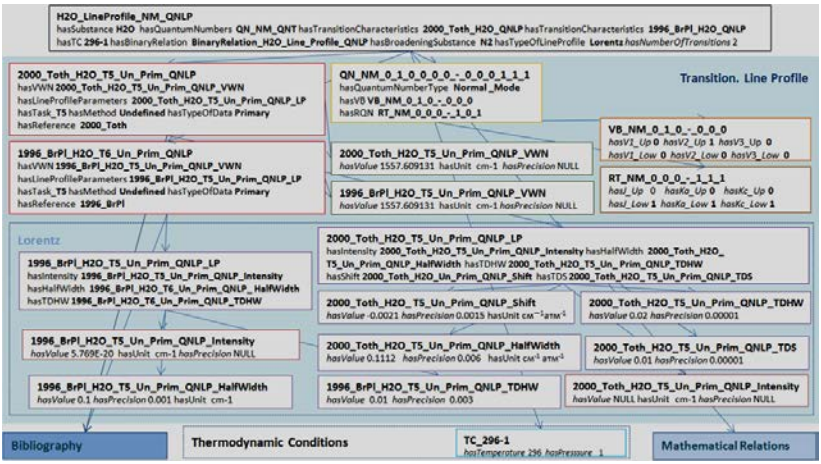

Рис. 4. Структура индивида перехода молекулы в газе.

5. Сравнение онтологии информационных ресурсов и переходов по молекуле $D_2{}^{17}O$

Ниже проведено сравнение онтологии информационных ресурсов и онтологии переходов для молекулы $D_2{}^{17}O$. Сравнение онтологий проведено по нескольким метрикам. Онтология информационных ресурсов содержит 8162 логические аксиомы, 103 класса, 40 объектных и 154 конкретных (datatype) свойств и 1315 индивидов. Выразительность дескриптивной логики ALCHON(D). Значительное число аксиом (7406) являются аксиомами фактологического типа (A-box типа), т.е. они отнесены к классам и свойствам. Большинство индивидов используются исключительно в A-box, хотя некоторые из индивидов являются номиналами.

Онтология молекулярных переходов содержит 30417 логических аксиом, 36 классов, 19 объектных и 26 конкретных свойств и 6504 индивида. Выразительность дескриптивной логики, ис-

пользуемой в онтологии, ALCRIF(D). Как и в онтологии информационных ресурсов, большая часть аксиом имеет A-box тип. В отличие от онтологии информационных ресурсов, онтология молекулярных переходов не содержит номиналов: все индивиды относятся к A-box. В онтологии молекулярных переходов есть несколько аксиом характеризующих смежные классы.

Существующий на данное время вариант онтологии информационных ресурсов содержит лишние аксиомы, которые могут привести к уменьшению производительности работы машины вывода. Другими словами, машина вывода должна делать дополнительную работу, которая не повлияет на конечный результат. Так, например, описание класса **EinsteinCoefficientDescription** содержит анонимный суперкласс (hasUnit **value** {s-1}), а объектное свойство hasUnit является функциональным (это следует из аксиомы (hasUnit Exactly 1 Thing)). В онтологии у этого класса описаны 56 индивидуальных членов, каждый из которых содержит свой экземпляр аксиомы (hasUnit **value** {s-1}), выраженный в форме ObjectPropertyAssertion. Эти 56 аксиом не добавляют новой информации в онтологию (в силу функциональности свойства hasUnit), но учитываются при работе машины вывода, ухудшая её производительность. В свою очередь, онтология молекулярных переходов содержит функциональные свойства, ограничивающие необходимость использования аксиом вида (hasUnit exactly one Thing), и, таким образом, сохраняет семантику моделируемых сущностей. На первый взгляд, такое моделирование более эффективно при выводе. Но требуются дополнительные эксперименты для проверки этого предположения.

Заметим, что размер A-box больше у онтологии переходов, при сравнении с онтологией информационных ресурсов. Это связано с тем, что последняя онтология моделирует метаданные (в частности, информационные источники) для решений спектро-

скопических задач (T1-T7), тогда как онтология молекулярных переходов моделирует сами решения этих задач. Так как размер этих онтологий для молекулы $D_2{}^{17}O$ (и онтологий других молекул) растет, то критически важным является представление этих онтологий таким образом, чтобы они допускали логический вывод и ответы на запросы. Вычислительно, логический вывод в сложных онтологиях требует больших затрат: вычислительная сложность задачи проверки совместности онтологии или включения классов для данных онтологий составляет 2NEXPTIME. Это значит, что для онтологии размера n проверка её совместности может потребовать в худшем случае 2^m операций, где m=2^n. Эта сложность может быть уменьшена путём неиспользования ряда конструкторов в онтологии. Мы планируем исследовать различные подходы к моделированию с целью увеличения скорости вывода при сохранении необходимой точности моделирования.

6. Заключение

Доклад посвящен расширению автоматического описания информационных ресурсов в количественной спектроскопии. Расширением является детальное описание состояний и переходов молекул, выполненное с помощью языка OWL 2 DL. В докладе приведены описания трех основных групп индивидов, характеризующих уровни энергии и переходы в изолированной молекуле и молекулы в газовой фазе. Каждый индивид содержит полный набор значений физических величин, характеризующий состояние или переход, определяемый заданным набором квантовых чисел. Описаны классы онтологии и проведено сравнение онтологии информационных ресурсов и переходов для молекулы $D_2{}^{17}O$.

Литература

1. Rothman, L.S., Gordon, I.E., Babikov, et al., The HITRAN 2012 Molecular Spectroscopic Database, Journal of Quantitative Spectroscopy and Radiative Transfer, 130, 4-50 (2013)
2. Jacquinet-Husson, N., Crepeau, L., Armante, R., et al., The 2009 edition of the GEISA spectroscopic database, J. of Quant. Spectros. and Rad. Transfer, 112(15), 2395-2445 (2011)
3. Dubernet, M.L., Boudon, V., Culhane, J.L., et al., Virtual atomic and molecular data centre, Journal of Quantitative Spectroscopy and Radiative Transfer, 111(15), 2151-2159 (2010)
4. Tennyson, Jonathan, Bernath, et al., IUPAC Critical Evaluation of the Rotational-Vibrational Spectra of Water Vapor. Part I. Energy Levels and Transition Wavenumbers for $H_2{}^{17}O$ and $H_2{}^{18}O$, J. of Quant. Spectros. and Rad. Transfer, 110(9), 573-596 (2009)
5. Tennyson, J. , Bernath, P. F., Brown, et al., IUPACcritical evaluation of the rotational–vibrational spectra of water vapor. Part II: Energy levels and transition wavenumbers for $HD^{16}O$, $HD^{17}O$, and $HD^{18}O$, J. of Quant. Spectros. and Rad. Transfer, 111(15), 2160-2184 (2010)
6. Tennyson, Jonathan, Bernath, Peter F., Brown, Linda R., et al., IUPAC critical evaluation of the rotational–vibrational spectra of water vapor, Part III: Energy levels and transition wavenumbers for $H_2{}^{16}O$, J. of Quant. Spectros. and Rad. Transfer, 117, 29–58 (2013)
7. Tennyson, Jonathan, Bernath, Peter F., Brown, Linda R., et al., IUPAC critical evaluation of the rotational–vibrational spectra of water vapor. Part IV. Energy levels and transition wavenumbers for $D_2{}^{16}O$, $D_2{}^{17}O$, and $D_2{}^{18}O$, J. of Quant. Spectros. and Rad. Transfer, 142, 93–108 (2014)
8. Privezentsev, A.I., Construction of ontological knowledge bases and software for description of molecular spectroscopy information resources, PhD Thesis, Tomsk, (2009)

9. Lavrentiev, N. A., Privezentsev, A.I., Fazliev A.Z., Computed knowledge base for describing information resources in molecular spectroscopy. 2. Data model of quantitative spectroscopy, Russian Digital Libraries Journal, 14(2), (2011)
10. Privezentsev, A.I, Tsarkov, D.V., Fazliev, A.Z., Computed knowledge base for describing information resources of molecular spectroscopy. 3. Basic and applied ontologies**,** Russian Digital Libraries Journal, 15(2), (2012)

ОБ ИДЕНТИФИКАЦИИ И КРОСС-ИДЕНТИФИКАЦИИ НЕБЕСНЫХ ОБЪЕКТОВ

О.Ю. Малков, О.Б. Длужневская, П.В. Кайгородов, Д.А. Ковалева, Н.А. Скворцов

Институт астрономии РАН, Москва

malkov@inasan.ru

Проблема обозначения небесных объектов появилась в астрономии давно и окончательного решения не нашла до сих пор. Параллельное существование и интенсивное использование десятков систем обозначений приводит также к необходимости постоянно решать проблемы кросс-идентификации. Особенно остро эти вопросы стоят для двойных и кратных систем - объектов, выглядящих, обозначаемых и каталогизируемых различными группами исследователей по-разному. В работе кратко обозреваются существующие методики обозначения одиночных и кратных астрономических объектов (с акцентом на применяемую в Базе данных двойных звезд систему BSDB), а также принципы кросс-идентификации.

Ключевые слова: кросс-идентификация, базы данных, двойные звезды.

Proper designation of astronomical objects is a chronic problem in astronomy, which solution is far from certain. Astronomers use a dozen of quite different designation schemes, and it causes perpetual problems of cross-identification. It is especially topical for binary and multiple systems, as these objects are observed, designated and catalogued differently by various researchers. We shortly review existing schemes of designation of single and multiple astronomical objects, describe BSDB system, implemented in Binary star database (BDB) and discuss problems and solutions of cross-identification.

Keywords: cross-identification, databases, binary stars

1. Идентификация небесных объектов

Астрономические объекты стали каталогизироваться еще во II в.н.э. Астрономические каталоги, созданные до начала XVII в., насчитывали до полутора тысяч объектов, а затем, с изобретением телескопа, число объектов стало стремительно расти. Каталоги 70-х годов прошлого века, когда стали образовываться центры астрономических данных и создаваться астрономические базы данных, насчитывали до 2 млн. объектов, а современные каталоги включают млрд. объектов, и в ближайшем будущем это число увеличится на 1-2 порядка.

Необходимо заметить, что постоянно растет не только число объектов, подлежащих каталогизации, но и количество каталогизируемых параметров. Если первые каталоги включали 3-4 параметра на объект (две координаты, визуальная оценка блеска и грубый классификатор уровня "звезда-туманность"), то современные каталоги характеризуют объект десятками и сотнями параметров, не считая различных функций (распределение энергии в спектре, функция изменения блеска со временем и пр.).

При наименовании небесных объектов (необходимом для каталогизации) астрономы не следовали какому-то одному правилу, точнее, применяли различные подходы. Хронологически первыми нужно считать методы, опирающиеся на систему созвездий - поименованных участков небесной сферы, содержащих группы звезд. Были введены в обращение, в частности, система Байера, именующая звезды в созвездии греческими буквами в порядке убывания яркости; система Флемстида, просто нумерующая звезды в созвездии с увеличением их прямого восхождения (т.е., для северного полушария Земли, справа налево); система Аргеландера-Хартвига (для переменных звезд, в порядке их открытия), использующая сложную комбинацию двухбуквенных обозначений

и, в случае их нехватки, порядковых номеров. Нехватка букв в греческом алфавите (особенно для протяженных созвездий) стала очевидной довольно быстро, а система Флемстида теряла свою стройность, если в созвездии обнаруживали новую, более слабую звезду, расположенную на небесной сфере среди уже перенумерованных.

С введением систем небесных координат исследователи получили возможность создавать имена (идентификаторы), базирующиеся на значениях координат. Идентификаторы компилировались из округленных (или, иногда, сокращенных) значений координат, при этом для объектов Галактики применялась экваториальная, а для внегалактических объектов - галактическая система координат. Такие системы уже не были связаны с границами созвездий, но тоже оказались не совсем устойчивыми: из-за прецессии и нутации земной оси, а также из-за собственных движений координаты небесных тел изменяются (у некоторых - весьма значительно). Таким образом, координатно-ориентированные имена звезд оказывались, вообще говоря, разными в разных системах, если введение этих систем в строй отстояло друг от друга на более чем 10 лет (это число зависит, конечно, от позиционной точности каталога).

Кроме того, каталоги, базирующиеся на таких системах и традиционно отсортированные по значению прямого восхождения (аналог земной долготы), оказались весьма неудобны для планирования наблюдений, так как в них могли соседствовать объекты с очень разными значениями склонений (аналог земной широты). С увеличением числа каталогизируемых небесных объектов этот недостаток становился все более досадным, и авторы каталогов, содержащих более 100 тыс. объектов в конце концов перешли на зонную систему идентификации. Небесная сфера "нарезалась" на

зоны (пояса) по склонению (а позже, для каталогов с числом объектов, превышающим 10 млн., и эти пояса пришлось "порезать" на более мелкие зоны), и объекты в каталоге сортировались по зонам, а внутри зоны - по прямому восхождению.

Наконец, в "тематических" каталогах, содержащих объекты (или параметры объектов) определенного типа, принято просто нумеровать их в порядке открытия. Помимо упомянутых выше переменных звезд можно привести в пример каталог спектроскопических двойных систем и (с некоторыми оговорками) ставший в последнее время популярным, в связи с открытием экзопланет, каталог близких звезд Глизе (V/70A, указана нумерация каталога в Базе данных каталогов VizieR [1]).

Таким образом, небесные объекты оказались поименованы и включены в различные каталоги в соответствии с их:

- положением в созвездии,
- координатами,
- координатами в зонах небесной сферы,
- блеском (разным в различных фотометрических системах),
- очередностью открытия
- или вообще без всякой системы, как, например, объекты в широко известном каталоге туманностей Мессье [2].

Нужно отметить, что практически все эти (даже самые архаичные) системы наименований сохранились, успешно применяются по сей день и продолжают создавать проблемы для успешной кросс-идентификации небесных объектов.

Так, самые яркие (и следовательно, наиболее хорошо изученные и включенные во многие каталоги) звезды имеют 4-5 десятков общеупотребительных наименований.

Данная работа посвящена идентификации звезд, однако практически все вышесказанное справедливо для туманностей и скоплений нашей Галактики и для более удаленных объектов (галактик и квазаров). В статье рассматриваются, с одной стороны, схемы идентификации одиночных и кратных звёзд как системы их обозначения в каталогах, а с другой стороны, - методы кросс-идентификации звёзд, использующие как их идентификаторы, так и наблюдательные и астрофизические параметры для отождествления звёзд в разных каталогах.

2. Кросс-идентификация небесных объектов

Разнообразие методов создания астрономических каталогов поставило астрономов перед задачей выработки схем кросс-идентификации содержащихся в них объектов. Действительно, точная позиционная информация, содержащаяся в каталоге А, будучи проанализированной совместно с высокоточной фотометрией из каталога Б и данными радионаблюдений из каталога В, позволяет получить более полную картину образования, строения и эволюции Галактики, чем данные из каждого из этих каталогов по отдельности. При этом три упомянутых каталога создавались разными коллективами по разным методикам, вообще говоря, в разное время и использовали различные системы идентификации. Базы астрономических данных, объединяющие неоднородную информацию, требуют в первую очередь решения проблемы кросс-идентификации объектов.

Проблема кросс-идентификации (КИ) астрономических объектов состоит в отождествлении одних и тех же объектов среди неоднородных данных из разных каталогов. Она является частным случаем рассматриваемого в информатике направления разрешения сущностей (entity resolution). Комплексный подход к разре-

шению сущностей обычно представляется в виде определённой последовательности действий, включающей [4,5]:

- связывание элементов схем данных разных источников, соответствующих по смыслу, предварительную очистку и приведение к единообразному представлению неоднородных данных из связанных атрибутов;
- индексирование данных с целью снижения попарного перебора сравниваемых кортежей из разных источников;
- применение методов сравнения данных различных типов, включая строки, числовые типы, даты, пространственные координаты, множества, записи в целом;
- выделение набора данных, однозначно определяющих уникальные объекты, а при невозможности его выделения оценка близости данных из разных источников по определённым критериям и принятие решения об отождествлении описанных ими объектов.

Естественной основой разрешения объектов применительно к решению задачи кросс-идентификации небесных объектов являются их пространственные координаты. Данные атрибутов, содержащие координаты объектов, приводятся к единообразному представлению с учётом используемых форматов координат и разности эпох наблюдения. Близость координат с определённым допущением решает проблему попарного сравнения кортежей в каталогах. Координатное совмещение данных различных каталогов позволяет решать до 80 процентов проблем, связанных с КИ.

Однако чисто координатного подхода оказывается недостаточно, если речь идет о плотных звездных полях (скоплениях или кратных системах, см. ниже), о быстро движущихся и/или переменных объектах, о данных с различающимся угловым разрешением, а соответственно, различной точностью координат и т.п. В

таких случаях почти всегда для КИ удается использовать атрибуты каталогов, содержащие фотометрическую информацию. При приведении фотометрических данных к единообразному представлению приходится принимать во внимание тот факт, что блески (или цвета, т.е., разницы блесков) объектов в различных фотометрических системах, вообще говоря, различны и, хотя и подчиняются неким корреляционным соотношениям, соотношения эти опять-таки различны для объектов различной природы (которая, как правило, при КИ остается неизвестной).

Полезным дополнительным параметром является также классификатор объекта, однако он присутствует в каталогах далеко не всегда и представляет, как правило, достаточно грубую оценку природы объекта (точечный-протяженный). Естественно, используется и вся другая информация (например, спектральный тип объектов), если она присутствует в обоих кросс-идентифицируемых каталогах.

Таким образом, критерии отождествления и оценки близости объектов при кросс-идентификации разрабатываются с учётом позиционных и фотометрических параметров, рассчитанных астрофизических величин, параметров фотометрических систем и углового разрешения оборудования, используемого для обзоров.

Одной из особенностей данных астрономических каталогов является то, что некоторые из них уже включают данные об именах описываемых объектов в соответствии с определёнными системами идентификации, принятыми в других каталогах. Приведённые в каталогах имена идентифицируют соответствующие объекты в других каталогах, и таким образом, являются результатом уже решённой при их составлении задачи отождествления между конкретными парами каталогов.

Однако кросс-идентификация для разных пар каталогов может производиться различными методами, различные виды иденти-

фикаторов имеют разный смысл. Использование разных критериев отождествления объектов, различных систем идентификации и способов связывания имён в каталогах может само по себе рождать конфликты идентификации. В том числе, точность используемых методов отождествления невозможно выяснить из самих идентификаторов. Поэтому для проверки корректности существующих связей идентификаторов объектов и для разрешения конфликтов между идентификаторами при отождествлении объектов более, чем в двух каталогах, приходится заново прибегать к решению задачи КИ на основе наблюдательных и астрофизических параметров.

С учетом изложенных выше соображений астрономам удалось решить (и удается решать, с появлением новых каталогов и обзоров) большинство проблем КИ, что находит свой результат в публикуемых таблицах КИ и создаваемых базах астрономических данных различных типов. Приведённые выше критерии, однако, не во всех случаях достаточны для решения задачи кросс-идентификации. Нередко возникает необходимость в более специфических методах. Для некоторых типов объектов задача кросс-идентификации далека от окончательного разрешения, и, в первую очередь, это относится к двойным и кратным звездам.

3. Особенности идентификации и кросс-идентификации кратных звезд

Двойные и кратные звезды весьма многочисленны и не исключено, что их доля среди звезд Галактики (если включать в их число и планетные системы) весьма близка к 100 процентам. Столь высокая кратность объясняется особенностями звездообразования, в частности, необходимостью для вращающегося и сжимающегося протозвездного газо-пылевого облака избавиться от осевого момента инерции, что проще всего осуществить за счет

фрагментации на компоненты и/или образования планетной системы.

Далеко не все двойные звезды наблюдаются именно как двойные. Для этого паре нужно либо находиться достаточно близко к наблюдателю и быть достаточно широкой (тогда компоненты будут наблюдаться по отдельности), либо демонстрировать доплеровское смещение линий в спектре и/или переменность блеска из-за орбитального движения компонентов, либо проявлять себя как источник рентгеновского излучения (возникающего из-за аккреции вещества на один из компонентов) и т.п. Эти и другие типы двойных регистрируются с помощью различных методик различными коллективами и им, естественно, присваиваются обозначения в рамках различных схем идентификации. В результате к "традиционным" проблемам схем идентификации прибавляется несколько новых, характерных именно для двойных (кратных) систем.

Прежде всего, требуется разработать методику обозначений для идентификации компонентов кратной системы. Для двойных звезд эту проблему решали традиционно, добавляя к идентификатору системы в качестве суффиксов буквы A и B. Но уже с тройными системами поступали по-разному. В тесных системах, когда оказывалось, что компонент A представляет собой на самом деле двойную звезду, новые два компонента получали обозначения Aa и Ab. Этот принцип, помимо прочего, отражал и тот факт, что кратные системы (кроме самых широких) должны быть иерархическими, иначе они будут динамически нестабильными и просуществуют недолго. Исследователи же широких систем, где компоненты, как правило, наблюдаются по отдельности, а уровни иерархии не очевидны, обозначали вновь открытый компонент буквой C. Аналогично эти принципы распространялись на системы более высокой кратности.

Эти схемы, естественно, не идеальны. Помимо того, что появляются трудно форматируемые обозначения типа Aa1, а в неиерархических системах особенно высокой кратности (которые некоторые исследователи, впрочем, предпочитают называть скорее скоплениями) не хватает букв латинского алфавита. Открытие компонента на промежуточном иерархическом уровне является более редким событием, но также нарушает описанные выше принципы наименования объектов.

Еще одной, дополнительной трудностью, присущей даже двойным системам является порядок присвоения букв A (главному) и B (вторичному) компонентам, точнее, неоднозначность ответа на вопрос, какой компонент в паре является главным. Для исследователей визуально-двойных звезд это - более яркий компонент (оставим в стороне вопрос о порядке присвоения букв в парах с компонентами одинаковой яркости, а также то обстоятельство, что в разных фильтрах относительная яркость компонентов может быть разной, а в некоторых случаях самым ярким агентом в системе является даже не звезда, а аккреционный диск вокруг одной из звезд), для исследователей переменных звезд - более горячий. При моделировании тесных двойных систем принято считать главным компонентом более массивную на сегодняшний день звезду, а с точки зрения звездной эволюции главный компонент - изначально более массивная звезда (из-за переноса массы в системе в процессе эволюции это могут быть разные компоненты). С точки зрения кинематики двойной системы главный компонент - меньший по массе. И существуют, наконец, задачи, для которых удобно считать главным компонентом больший по размерам.

Все эти обстоятельства приводят к тому, что компоненты (и сами системы) двойных и кратных звезд получают в различных каталогах весьма различные обозначения (присваиваемые в соот-

ветствии с различными схемами идентификации), и задача КИ, более-менее решенная для одиночных объектов, становится гораздо более сложной для двойных и кратных систем.

4. Система идентификации BSDB

При создании Базы данных двойных звезд (BDB, [3]), включающей в себя сведения о двойных и кратных системах всех наблюдательных типов, авторами разработана схема обозначений BSDB, призванная разрешить существующие проблемы идентификации и кросс-идентификации двойных систем. BSDB должна была удовлетворять (и удовлетворяет) следующим критериям:

- ни один объект не должен носить более одного идентификатора;
- ни один идентификатор не должен быть присвоен более, чем одному объекту;
- открытие новых компонентов в системе не должно нарушать принципы присвоения идентификаторов;
- система должна быть несложной, близкой к традиционным и интуитивно понятной исследователям двойных.

При присвоении идентификатора по системе BSDB мы выделяем три категории объектов: система, пара, компонент. Это подход следует считать пионерским, и диктуется он тем обстоятельством, что каждая из трех категорий характеризуется своим набором наблюдательных данных. Компонент характеризуется массой, радиусом, температурой, светимостью, и т.п. (то есть, тем набором астрофизических параметров, которым характеризуется, например, одиночная звезда). Пара - это два объекта (каждый из которых, кстати, тоже может оказаться парой), связанных гравитационно. Эта категория характеризуется такими параметрами как относительное положение членов пары на небесной сфере

(для визуальных двойных), орбитальными параметрами (период обращения, эксцентриситет орбиты и пр. - для орбитальных и части спектроскопических двойных), интегральным блеском (для фотометрически неразрешенных) и спектром (для спектроскопически неразрешенных). Здесь следует заметить, что наблюдатели имеют дело преимущественно именно с парами, и именно информация о парах, как правило, и включается в каталоги. Наконец, такая категория как система характеризуется общими параметрами: возраст, металличность, расстояние, кинематика в Галактике и пр. Некоторые параметры могут приписываться различным категориям: так координаты характеризуют каждый из компонентов в случае разрешенной двойной и пару - в случае неразрешенной.

Идентификатор BSDB состоит из цифровой части, компилируемой из значений небесных координат и предваряемой символом 'J' (означающим, что координаты относятся к эпохе 2000.0 года); индикатора "система-пара-компонент" ("s", "p", "c"), отделяемого двоеточием; и буквенным обозначением, в общих чертах напоминающим знакомые исследователям двойных звезд схемы обозначений. Так, обозначения объектов некой тройной системы будут выглядеть следующим образом:

J000144.48+590527.1:s
J000144.48+590527.1:pAa-Ab
J000144.48+590527.1:cAa
J000144.48+590527.1:cAb
J000144.48+590527.1:pA-B
J000144.48+590527.1:cB

Координатная часть обозначения BSDB внутри системы не меняется, несмотря на то, что координаты компонентов, вообще говоря, могут различаться.

Принципы создания идентификатора BSDB удовлетворяют правилам, утвержденным Международным астрономическим союзом.

5. Решение проблем кросс-идентификации двойных и кратных систем

Схема обозначений BSDB должна быть всеобъемлющей, поэтому необходимо позаботиться о том, чтобы кратные системы всех наблюдательных типов могли получить в ней соответствующие обозначения. Для этого коллективом ведется работа по созданию общего каталога идентификаций двойных звезд (предварительное название - Identification List of Binaries, ILB), который должен включать BSDB обозначения для всех каталогизированных в настоящее время двойных систем, а также предоставить такую возможность и для будущих списков/каталогов/обзоров двойных.

К каталогу ILB постепенно подключаются каталоги двойных систем, начиная с самых широких и, одновременно, самых представительных. Каждому объекту, встречающемуся впервые, приписывается уникальное обозначение BSDB. Объекты, уже имеющиеся в ранее исследованных каталогах, дописываются в соответствующие строки ILB. Новые объекты, входящие в уже существующие в ILB звездные системы, приводят к корректировкам соответствующих разделов каталога. При этом приходится решать проблемы КИ, которые возникают даже в том случае, когда кратная система принадлежит только одному наблюдательному типу и, следовательно, ее составляющие поименованы хоть, возможно, и по-разному, но, по крайней мере, в соответствии с одной и той же схемой идентификации. Задача усложняется, когда в системе присутствуют объекты, проявляющие свою двойственность по-разному (т.е., принадлежащие различным наблюдатель-

ным типам, изучаемые различными группами исследователей и, в результате, имеющие весьма разные обозначения). Более того, объект, представляющийся одиночным с точки зрения одной методики наблюдений, может оказаться двойным или кратным с точки зрения другой; это является следствием разницы в позиционной и фотометрической точности используемых каталогов (методик наблюдения), сказывается на приписываемых идентификаторах и усложняет проблему КИ.

Для решения проблем КИ привлекается вся имеющаяся в каталогах информация, в первую очередь - позиционная и фотометрическая, а также уже содержащаяся в некоторых каталогах кросс-идентификация. В процессе КИ, однако, нами обнаруживаются ошибки как в оригинальных каталогах, так и в базах данных общего назначения, о чем мы сообщаем их авторам.

Около 90 процентов всех проблем КИ удается разрешить автоматически, и это делает проблему создания унифицированных методов КИ в принципе решаемой. Для оставшихся 10 процентов все же требуется ручной подход - опять-таки, в сотрудничестве с авторами оригинальных каталогов.

Каталог ILB будет постоянно пополняться, станет основой для базы данных BDB, а также может служить для других приложений. Методика КИ компонентов, пар и систем двойных и кратных звезд также должна считаться оригинальной (как и система обозначений BSDB); она постоянно модифицируется и станет полезной для будущих астрономических обзоров.

6. Заключение

Работа посвящена одной из проблем астрономии: обозначению двойных и кратных объектов, а также кросс-идентификации этих объектов. Дан обзор систем и стандартов идентификации астрономических объектов, описаны сложности и особенности

для кратных объектов. Обсуждаются также существующие методы и средства кросс-идентификации объектов. Описана методика обозначений, применяемая в Базе данных двойных звезд БДБ, а также иллюстрируются примеры разрешения конфликтов идентификации.

Благодарности. Работа выполнена при поддержке грантов РФФИ 12-02-31904, 12-07-00528, и Программы Президиума РАН Поддержка ведущих научных школ (грант НШ-3620.2014.2).

Литература

1. VizieR database http://vizier.u-strasbg.fr/viz-bin/VizieR.
2. O'Meara S.J. The Messier objects. - Cambridge University Press, 1998. – P. 3. – 304 p.
3. Malkov O.Yu., Kaygorodov P.V., Kovaleva D.A., Oblak E., Debray E. 2014, Astronomical and Astrophysical Transactions, 28, 235 .
4. Christen P. Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection. - Springer, 2012.
5. Christen P. A survey of indexing techniques for scalable record linkage and deduplication. // IEEE Transactions on Knowledge and Data Engineering. - 2011

МЕТОДЫ УПРАВЛЕНИЯ ДАННЫМИ, ИХ ОРГАНИЗАЦИИ И АНАЛИЗА В АСТРОФИЗИЧЕСКИХ ИССЛЕДОВАНИЯХ

О.П. Желенкова, В.В. Витковский
САО РАН, Университет ИТМО
zhe@sao.ru, vvv@sao.ru

Проведен обзор текущего состояния разработок IVOA, а также применяемых стандартов и технологий в коммуникационной инфраструктуре виртуальной обсерватории. Рассмотрены этапы жизненного цикла астрономических данных, их обеспеченность стандартами в контексте постоянно растущего объема данных.
Ключевые слова: цифровые коллекции, информационные инфраструры, виртуальная обсерватория.

The current state of the IVOA activity, and applied standards and technologies of a communication infrastructure of the Virtual Observatory was reviewed. Also we considered the lifecycle of astronomical data and their provision of standards in the context of an ever-growing amount of data.
Keywords: digital collections, information infrastructure, virtual observatory.

Введение

В астрономии с начала 20-го века мы наблюдаем третью «революцию» в экспериментальных данных. Это касается представления информации, методов накопления, объемов данных, а также обработки и анализа наблюдений, начиная с аналоговых светоприемников – фотоэмульсий, затем - цифровых светоприемников, а в настоящее время – гигабайтные объемы информации, получаемые в одном наблюдении, являются нормальным явлением. Поменялись способы и технологические средства, применяемые при хранении данных – от стеклотек до информационных систем,

поддерживаемых системами управления базами данных. Изменились и способы обмена информацией – от обмена копиями наборов данных, до эффективного и простого доступа к информации посредством веб-сервисов. Результат последнего - это создание виртуальной обсерватории (ВО). Традиционный подход, при котором исследователь анализирует данные с помощью процедур обработки только на своем рабочем компьютере, уже ушел в прошлое. Однако и сейчас нужно прилагать значительные усилия для организации, сортировки и обработки данных. Анализ больших объемов данных остается сложной, трудоемкой и ресурсоемкой задачей. И возможным решением этих задач является систематическое использование новых технологий интернета, систем управления базами данных и др. Системы управления базами данных имеют механизмы и обеспечивают средства для организации работы с большими наборами данных, эффективного поиска, сортировки, поддержки целостности данных и пр.[1].

Ожидаемые объемы цифровой информации в астрономии заставляют обратить внимание и на жизненный цикл данных, который можно представить в виде следующих этапов: получение и накопление данных, хранение, обмен и доступ к данным, обработка, анализ данных. Стандартизация этих этапов могла бы существенно повысить эффективность исследований, принимая во внимание присутствующий контекст Больших Данных. Далее мы проведем обзор имеющихся в астрономии стандартов и применяемых технологий для каждого из этапов и начнем с текущего статуса виртуальной обсерватории (ВО).

Текущее состояние разработок IVOA

Международный альянс IVOA (International Virtual Observatory Allience, http://www.ivoa.net) существует уже более 10 лет. Его основу теперь составляют следующие группы: координационная группа TCG, 6 рабочих групп WD (Applications, Data

Access Layer, Data Model, Grid and Web Services, Registry, Semantics), 5 групп по интересам IG (Data Curation&Preservation, Education, Knowledge Discovery in Databases, Theory, Time Domain), 3 комитета (Exec, Standards and Processes, Science Priorities), одна рабочая группа находится в неактивном состоянии (VOTable), WG VOQueryLanguage слилась с WG DAL, WG VOEvent преобразована в IG Time Domain.

Основа архитектуры ВО была заложена в первых разработках. Она опирается на общепринятые стандарты – HTTP и XML, а также SOAP/WSDL или REST для описания веб-сервисов, стандарты семантической сети - RDF и OWL. Инфраструктура ВО - это ресурсы и сервисы. К ВО-ресурсу можно отнести – веб-страницы, единицы хранения, интерактивные приложения и т.д., то есть то, к чему можно обратиться по URL-адресу. Ресурс можно использовать в ВО-инфраструктуре, если он описан стандартными метаданными, и кроме того, содержимое ресурса должно соответствовать стандартной модели данных. К ВО-сервису относятся такие программные приложения, которые что-то делают для пользователя, используя связи между двумя компьютерами. Для того, чтобы сервис использовался в ВО, он должен работать по ВО-протоколу, а его метаданные, описывающие, что это за сервис, как он работает и что производит, должны быть опубликованы. Философия архитектуры ВО состоит в том, чтобы обеспечить разработчиков только стандартными интерфейсами, не диктуя, как и что делать внутри приложения. За годы существования структура ВО уточнялась, модернизировалась, добавлялись новые стандарты. На текущий момент разработано 42 стандарта со статусом «рекомендация» и 69, еще не доведенных до этого уровня [2]. Далее проведем обзор составных частей ВО и поддерживающих их стандартов.

Доступ к данным. Следующие протоколы IVOA определяют базовые методы доступа для различных типов данных: SIA (Simple Image Access) - для архивов изображений, SSA (Simple Spectral Access) - для спектров, SCS (Simple Cone Search) - для каталогов и несколько других протоколов. Эти протоколы называются «простыми», поскольку только несколько параметров необходимо для организации запроса данных (обычно координаты области поиска). Для более гибкого доступа разработан протокол TAP (Table Access Protocol). Он предоставляет более универсальный способ обращения к базам данных обзоров в виде SQL-подобных запросов. Кроме поиска изображений внутри заданной области, он позволяет организовать поиск по дате наблюдений, фильтру, имени исследователя, и т.д. Несмотря на то, что SQL является стандартом для запросов к базам данных, на практике всегда имеются разные варианты реализации этого стандарта. По этой причине и из-за специфики предметной области нужна дополнительная стандартизация, что и сделано при разработке TAP и ADQL (Astronomical Data Query Language) протоколов. ADQL – еще один способ реализации запросов в ВО-среде, где SQL-запрос можно выразить простой текстовой строкой или в XML-формате, а затем принимающий сервис преобразовывает его в то, что соответствует сервисам базы данных.

Сервисы данных ВО обычно возвращают результаты в формате данных для таблиц, известном как стандартный ВО-формат VOTable. При поиске в каталогах в таком виде возвращается окончательный результат. При поиске же изображений мы получаем ВО-таблицу, включающую URL-адреса изображений, соответствующих критерию поиска, т.е. сами изображения не передаются при выполнении запроса, а дальнейшая организация их передачи определяется приложением.

Ресурсы и регистры. Два ключевых стандарта определяют ВО-ресурс - идентификатор ресурса RI (Resource Identifier), однозначно специфицирующий расположение ресурса, и метаданные ресурса RM (Resource Metadata), описывающие ресурс. Информация о ресурсах собирается в регистрах или реестрах ресурсов. Не существует единого централизованного реестра. В принципе, любой желающий может создать реестр согласно стандарту VOR (VO Resource), который определяет метаданные ресурса в записи реестра. Общая идеология сохраняется и в серии стандартов RE (Registry Extension), где определяется дополнительная информация, необходимая для описания разных типов ресурсов, например - VODS (VO Data Service), ARE (Application Reg Ext) и т.д. Стандарт RI (Registry Interface) описывает метод, с помощью которого приложения должны общаться с регистрами.

В ВО используется два типа регистров – поисковые регистры, к которым обращаются приложения, и регистры для публикации, куда помещается информация о ресурсах. Один реестр может собирать информацию из другого реестра. Поисковые реестры обычно поддерживают актуальным состояние информации о ресурсах, в то время как реестры для публикации не обязательно должны быть полными. Если надо знать все реестры, которые существуют в ВО, то IVOA поддерживает реестр реестров.

Модели данных и семантика. Сервис, соответствующий протоколу SCS, выполняет поиск строк каталога, попадающих в заданный диапазон координат, причем возвращаемые данные, как правило, содержат еще другие столбцы таблицы. Чтобы разобраться в физическом смысле полученной информации, пользователю нужно было бы обратиться к документации соответствующих каталогов. Но если возвращаемые данные представлены в стандартном формате и используются стандартные способы задания того, какие величины в колонках, то тогда возможно ав-

томатизировать весь процесс распознавания табличных данных. Для установления соответствия между названием колонки и физическим смыслом, имеющихся в ней данных, поддерживается словарь, известный как UCD (Universal Content Descriptor). Так столбец с именем "Alpha" может быть чем угодно, но если он имеет соответствующий UCD - "POS.EQ.RA", то посредством UCD-словаря можно определить, что это - прямое восхождение в экваториальной системе координат. Однако, может быть несколько столбцов в таблице с одинаковым UCD - например, RA источника и RA кадра детектора, где источник обнаружен. В этом случае, чтобы определить, что есть что, нужна полная модель данных. Разработано несколько моделей данных, таких как PhotDM - для данных фотометрии, ObsCoreDM - для описывания коллекции наблюдательных данных, VOEvent - для транзиентных событий и т.д., но разработка полной модели данных не завершена. Модели данных используют стандарт Utypes для определения элементов модели, стандарт STC (Space Time Coordinates) - для описания координат пространства и времени, Units - для единиц измерения физических величин. Более общий по сравнению с UCD стандарт Vocabulary, базирующийся на W3C RDF и SKOS, разработан для гибкого определения словарей, чтобы предоставить возможность исследователям определять и поддерживать свои собственные словари, которыми могут воспользоваться и другие.

Распределенная вычислительная инфраструктура. ВО является распределенной системой, архитектура которой опирается на сервисы, способные взаимодействовать друг с другом. Для этого, кроме стандартов, описывающих астрономические ресурсы и протоколы доступа, нужны стандарты, определяющие связи между сервисами. Так VOSI (VO Support Interfaces) определяет порядок взаимодействия любого веб-сервиса в среде ВО. Этот

стандарт выполнен на основе WSDL или WADL. Связанный с VOSI стандарт WSBP (WebServicesBasicProfile) является своего рода путеводителем для создания веб-сервисов. Точно так же UWS (UniversalWorkerService) является шаблоном для настройки асинхронных сервисов с сохранением состояния, где можно указать вид работы в терминах JDL (Job Description Language), вернуть статус выполнения задания при опросе и так далее. Это особенно полезно при запуске удаленных приложений, которые могут управляться набором заданных параметров без необходимости постоянного взаимодействия с ними.

Наличие разделяемой области памяти, которая используется для временного хранения данных пользователя или результатов обмена между взаимодействующими сервисами, является еще одним аспектом, присущим распределенной инфраструктуре. Так же как с базами данных и ADQL, подход ВО состоит в том, чтобы только обеспечить интерфейс к удаленной системе хранения, не привязываясь к конкретной технологии. Такой интерфейс обеспечивается стандартом VOSpace. Участвующая в инфраструктуре система хранения должна иметь уникальный адрес следующего вида vos://aaa.bbb.ccc/xxx, обеспечить методы добавления и удаления объектов, манипулирования метаданными объекта, а также предоставить URI, с помощью которых контент объектов будет доступен.

Совместные исследования и аутентификация. На начальных этапах ВО полагалось, что используются только общедоступные данные. Однако в астрономии есть данные с ограниченным доступом, причем надо реализовывать различные сценарии доступа. Это требует определения идентичности, методов аутентификации запроса и способов авторизации. ВО опирается на использование промышленных стандартов, таких как шифрование открытого ключа, формат отличительного имени DN (Distin-

guished Name), сертификаты X509, TLS (Transport Layer Security) и т.д. Стандарт SSOAM (SingleSignOn: Authentication Mechanisms) устанавливает утвержденные способы использования промышленных стандартов в ВО-сервисах. Пользователю удобно авторизоваться один раз в сессию, а учетные данные должны переадресовываться в соответствующие сервисы. Стандарт CDP (Credential Delegation Protocol) определяет, как это делать безопасным способом.

Приложения. Пользователи могут одновременно работать с несколькими различными видами данных такими как изображения, спектры, каталоги, временные ряды и т.д. Вместо тяжеловесных универсальных приложений, которые делают все, более эффективно работать с небольшими приложениями, специализированными под конкретный вид обработки. Однако желательно иметь возможность обмениваться данными и информацией между этими приложениями. Например, приложение TOPCAT позволяет работать с табличными данными. У этого приложения есть возможность отправить таблицу в приложение для визуализации, такое как ALADIN, чтобы наложить табличные данные поверх изображения. Это осуществимо с помощью протокола SAMP (Simple Application Messaging Protocol). Любое SAMP-совместимое приложение регистрирует свое присутствие в концентраторе и может направлять сообщения другому работающему приложению, а также стартовать такой концентратор, если нет ни одного работающего. Для SAMP-протокола определен стандартизированный список сообщений, можно передавать FITS-изображения, VO-таблицы, изображения в формате JPEG, и т.д.

Отметим, что на текущий момент инфраструктура ВО не занимается и не предназначена для производства контента ресурсов. Разработка фундаментальных блоков ВО привела к тому, что она является эффективной и широко используемой системой. Из-

за этого неминуемо возникают новые цели - реализовать в рамках единой инфраструктуры производство контента ресурсов, включив в нее наблюдательные средства. При этом контент, получаемый на телескопах, конечно же, должен быть подготовлен так, чтобы соответствовать стандартным требованиям. Имеются промежуточные решения, опыт разработок которых будет в дальнейшем использован. Так, например, NOAO DMS (National Optical Astronomy Observatory Data Management System) включает в себя все компоненты для обеспечения потока данных информации от телескопов до пользователя ВО. Пока это рассматривается в качестве долговременной перспективы, а основная деятельность IVOA направлена на разработку стандартов для сервисов поиска, доступа, анализа и визуализации данных, которые частично поддерживают потребности жизненного цикла астрономических данных.

Получение и накопление данных

Астрономические данные получаются наземными и космическими телескопами, которые являются дорогостоящими экспериментальными комплексами. Для эффективного использования этих инструментов во многих ведущих обсерваториях мира реализованы системы потокового получения данных (системы класса end-to-end), в которых используется понятие наблюдательного цикла. Он рассматривается как единый технологический процесс, состоящий из взаимосвязанных этапов, куда входит подача заявок на наблюдательное время, составление расписания, подготовка к наблюдениям, сам процесс наблюдений, архивирование необработанных «сырых» данных, подготовка и проверка калибровочного материала, проверка правильности заполнения заголовков файлов, обработка, архивирование научных данных. В качестве примера можно привести DFS (Data Flow System), раз-

работанную в ESO (European Southern Observatory, http://www.eso.org). Она является набором программных средств и процессов для сквозного управления потоком данных, включающем как планирование, так и проведение астрономических наблюдений, а также сохранения наблюдательных данных в архиве. На самом нижнем своем уровне DFS опирается на технологию систем управления реляционными базами данных. Базы данных используются в качестве постоянного хранилища для большинства объектов и информации, проходящих сквозь систему. На верхних своих уровнях система использует комбинацию веб-ориентированных средств и клиент-серверных приложений на Java и C++. В частности, надежная и быстрая репликация баз данных на трансконтинентальных расстояниях (телескопы ESO находятся в Чили) имеет решающее значение при своевременной реакции на быстроменяющиеся погодные условия для оперативного решения возникающих проблем. Аналогичные системы есть и в других наземных обсерваториях. В САО РАН также ведутся разработки по реализации отдельных компонентов наблюдательного цикла для оптических телескопов обсерватории [3,4].

Из общепринятых стандартов и используемых средств для наблюдательного цикла можно назвать IAU стандарт хранения данных FITS, стандарты IVOA, а также RTML (Remote Telescope Markup Language) на базе XML [5,6], но они не обеспечивают полностью потребность этого этапа в стандартах.

Хранение данных

Из-за того, что астрономические явления часто носят переменный характер на разных временных интервалах, то долговременное хранение наблюдений обычно входит в компетенцию обсерваторий. Цифровые носители, используемые для хранения, требуют специального оборудования для раскодирования их со-

держимого. При считывании, записи и хранении информации могут возникнуть ошибки, что может привести к потере данных. Конечно, надежность устройств растет, и возникающие ошибки можно контролировать программно. Но компьютерное оборудование меняется с такой скоростью, что время жизни устройства считывания оказывается меньше, чем время физического разрушения цифрового носителя информации. Это требует эпизодического переписывания данных на новые носители. Поэтому при долговременном хранении данных в архивах необходимо пересматривать технологию хранения каждые несколько лет и раз в 3-5 лет переносить содержимое с устаревших на новые носители.

Системы хранения данных и технологии, связанные с ними, используются в астрономических центрах данных и обсерваториях, где объем информации велик. Копии каталогов и обзоров часто хранятся в нескольких центрах данных, что обеспечивает дополнительную сохранность и доступность.

Астрономические данные хранятся в формате FITS. Это – самодокументируемый формат, позволяющий записывать в файл, кроме цифровых данных наблюдений, еще и параметры, описывающие и идентифицирующие наблюдение. При всех своих положительных качествах и популярности в астрономическом сообществе FITS-стандарт имеет недостатки. Они связаны с тем, что стандарт был разработан в начале 80-х прошлого века для обмена данными между различными компьютерными платформами. По этой причине единственными обязательными в стандарте FITS являются ключевые слова, определяющие представление данных. Стандартизованы ключевые слова для описания координатной системы изображений. Остальные параметры не являются обязательными, включая и информацию о наблюдении. В FITS-стандарте не определены правила генерации новых ключевых слов. За время использования формата в разных обсерва-

ториях появились свои клоны ключевых слов. В результате один и тот же параметр наблюдения может по-разному называться. Формат представления данных VOTable, используемый сервисами ВО, гораздо лучше организован для семантического описания данных. Но, к сожалению, не всегда возможно отображение FITS-файлов в формат VOTable.

С развитием технологий меняются требования к форматам, к степени обработки данных и т.п. Данные, хранящиеся в архивах, по этим причинам начинают устаревать, что требует специальных усилий по реорганизации и реинженерингу данных. Так произошло с появлением ВО, где, кроме требований к сервисам по совместимости со стандартами ВО, есть требования к качеству контента, а именно, в инфраструктуре используются калиброванные или “science-ready” данные.

Обмен и доступ к данным

Основными форматами при обмене и доступе к данным остается FITS-формат, а также VOTable формат, еще используются графические форматы - jpeg, png. Механизмы и методы ВО, в общем-то, полностью решили проблему обмена и доступа к данным для любого исследователя.

Публикация данных современных обзоров в настоящее время чаще выполняется серией релизов. Это добавляет авторам компилятивных каталогов дополнительный объем работ по актуализации данных и, как следствие, сравнение и проверку новых данных с имеющимися результатами. Чтобы найти и получить данные пользователь сам инициирует взаимодействие с ВО. Всякий раз, когда пользователь хочет узнать об обновлениях, ему надо повторить первоначальный запрос, сравнить полученный результат с имеющимся и скопировать, если это требуется, данные. При постоянно растущих объемах данных, появлении новых релизов

обзоров и каталогов требуется другой подхода для отслеживания новой информации о небесных объектах, интересных пользователю. Веб-приложение для поддержки данных пользователя VOdka (VO Data Keeping-up Agent) [7] было единственной разработкой для решения этой задачи. Агент ретранслирует запросы пользователей в инфраструктуру ВО и рассылает уведомления об обновлениях. При выбранном пользователем темпе опроса агент асинхронно посылает один и тот же запрос, сформулированный пользователем, и фиксирует результаты, отражающие временной срез информации, а также выполняет сравнение этих срезов и оповещает пользователя по электронной почте. У пользователя есть возможность просматривать результаты запросов, сохраненные в snapshot-файлах, журналы сравнения этих файлов, копировать снимки и новые появившиеся данные, а также инкрементальные файлы, включающие старые, новые и пропущенные данные.

Обработка данных

В астрономических исследованиях крайне важной является обработка наблюдательных данных, и системы, предназначенных для решения этой задач, существовали и до появления программных средств ВО. Чаще всего используют несколько универсальных систем редукции астрономических данных, таких как MIDAS (www.eso.org/sci/software/esomidas), IRAF(iraf.noao.edu), AIPS(www.aips.nrao.edu), IDL(www.ittvis.com). Они включают большой набор команд и специализированных пакетов, а также встроенный командный язык, который позволяет работать с системами в интерактивном и пакетном режимах, писать программы, приспосабливающие возможности систем для обработки наблюдательных данных пользователя. В последнее время для этих систем разработаны программные интерфейсы к Python, Perl и Java в

дополнение к уже существующим API для C и FORTRAN. С помощью этих интерфейсов можно интегрировать возможности систем обработки астрономических данных и программных средств ВО.

Объемы данных от нового поколения инструментов таких, как ALMA (Atacama Large Millimeter/submillimeter Array, http://www.almaobservatory.org), LSST (Large Synoptic Survey Telescope, http://www.lsst.org), Pan-Starrs (Panoramic Survey Telescope & Rapid Response System, http://pan-starrs.ifa.hawaii.edu), LOFAR (Low-Frequency Array for radio astronomy, http://www.lofar.org), SKA (Square Kilometer Array, http://www.skatelescope.org) ожидаются терабайтные, при обработке которых потоки работ будут играть определяющую роль в получении научных результатов. Потоковая парадигма для работы с распределенными данными состоит в повторном использовании простых сервисов для реализации более сложных алгоритмов. В противоположность бизнес-процессам потоки работ для научных целей в большей степени ориентированы на потоки данных. Компоненты потока задач изолированы друг от друга посредством протоколов, а именно определены правила для запуска, а также структура входных и выходных данных. Поток задач запоминается в виде xml-файла, который можно потом повторно использовать и редактировать его для изменения параметров.

Такой подход используется в проекте AstroGrid (http://www.astrogrid.org). Архитектура компонентов и сервисов AstroGrid носит название Common Execution Architecture (CEA). В центре AstroGrid-архитектуры – потоки задач (Workflow). Имеются графические средства для контроля и выполнения задач, - можно собрать последовательность и параметры задач будут передаваться автоматически через еще один компонент системы – регистр (Registry). И последняя компонента MySpace –

виртуальная область хранения, в котором запоминаются результаты вычислений, временные файлы, новый вид информации - файлы запросов и файлы потоков работ, что позволяет пользователю настраивать и перезапускать задачи. Кроме AstroGrid есть еще проекты и разработки для поддержания потоков работ, например, система управления потоками задач Taverna (http://www.taverna.org.uk), которая включает Taverna Workbench Astronomy для создания и выполнения астрономических потоков задач с использованием сервисов ВО и других REST-сервисов; Kepler (http://kepler-project.org) – типичная научно-ориентированная система для потоков задач, стремящаяся к универсальности; MyExperiment (http://www.myexperiment.org) – сайт социальной сети для обмена и разделения потоков задач.

В последние годы, учитывая как популярность потоков работ, так и для удовлетворения новых потребностей, появился новый тип – «адаптивные потоки работ» или WDO (Workflow-Driven Ontologies). Основной их характеристикой является возможность менять структуру потока работ во время исполнения. Отметим, что ВО-сервисы могут быть использованы в качестве компонентов для ориентированных потоков задач, так как их выполнение независимо от платформы, и они обеспечивают воспроизводимость результатов.

Анализ данных

Анализ данных в астрономии связан с извлечением полезной информации из наборов данных, а именно: с обнаружением зависимостей, ассоциаций, изменений, аномалий и статистически значимых структур и событий в данных. Это близко по смыслу к термину «data mining» или интеллектуальный анализ данных. Интеллектуальный анализ данных называют четвертой парадигмой науки. К первым трем парадигмам относятся опыт, теорию,

объясняющую результаты опыта, и численное моделирование. Четвертая парадигма чем-то похожа на третью, за исключением того, что вместо сложных моделей, ее методы имеют дело с экспоненциально растущими и сложными наборами данных, содержащих много полезной информации.

Сложность анализа в астрономии проистекает не только из-за количества, разнородности данных, но также из-за высокой внутренней размерности, которая делает невозможным их описание или визуализацию любым обычным способом. Применение методов когнитивной визуализации существенно расширяет возможность оценки, отбора и объединения данных при анализе больших цифровых коллекций. Тот факт, что процедура динамической визуализации не опирается на априорные сведения о природе объектов, а значит и не привносит в проекции искажающих влияний той или иной модели, дает возможность полной мере использовать возможности интуиции, имеется в виду, прежде всего профессиональный опыт, эмпирические знания, на которых основывается профессиональная интуиция.

В стандартных системах обработки астрономических данных есть обширный набор средств для применения разных методов статистического анализа данных, классификации объектов, аппроксимации выявленных закономерностей и пр.

Поскольку данные в астрономии не связаны по смысловому содержимому, а также не обеспечивается передача знаний о небесном объекте, полученных другими исследователями, то невозможно выполнить запрос типа «найти все источники в каталогах, которые являются квазарами» и т.п. Это существенно ограничивает пользователю эффективную работу с информацией. В последнее время появилось несколько проектов, направленных на развитие сервисов, поддерживающих смысловую связность информации. Далее для примера приводятся несколько проектов.

Цель проекта AstroDAbis [8] - создание независимого механизма публикации пояснений (комментариев, аннотаций). Пояснения могут создаваться пользователем для одиночного объекта («объект X есть квазар») или для нескольких объектов («объект с номером 123 в каталоге A есть то же самое, что объект с номером 456 в каталоге B»). Как полагают авторы AstroDAbis, этим решаются проблемы передачи знаний, создания компилятивных каталогов и реализации их связи с родительскими каталогами. Авторы статей, где представлена информация, полученная на основе анализа каталогов, с помощью аннотаций могут передать знания о небесном объекте в форме, которая может быть использована в последующих запросах к каталогу. Когда возникает потребность объединить два каталога и создать компилятивный каталог, например, слияние оптических данных SDSS и инфракрасных данных тех же источников из UKIDSS, то такая связь позволит обойтись без повторной кросс-идентификации ресурсов. С помощью аннотации новые каталоги, полученные таким образом, можно сделать доступными для программного обеспечения, и связи между каталогами будут однозначно зафиксированы. Аналогичные разработки не являются новыми в науке. Например, имеется аннотирование данных в генетике - Distributed Annotation System, http: //www.biodas.org), или в Интернете – RDF (Resource Description Framework) и LOD (Linking Open Data)). AstroDAbis также имеет LOD-интерфейс, который обеспечивает создание URI для аннотируемых объектов каталогов, что подготавливает платформу для экспериментов с Semantic Web в астрономии.

Другой проект ADSASS (The ADS All-Sky Survey) [9] направлены на превращение системы NASA ADS (Astrophysics Data System), широко известной среди астрономов своей полнотой в качестве полнотекстового библиографического ресурса, в карту неба. Система ADS не является источником наблюдательных данных, но является неявным хранилищем ценных астрономических данных из публикаций в форме изображений, таблиц и ссылок на небесные объекты. Необходимо сделать эту ценную информацию

доступной для запросов и просмотра. В результате выполнения проекта астрономы получат карту неба, которая по нажатию клавиши будет активировать ссылки на статьи, показывая, какая часть неба в них описывается, а также слой исторических данных на базе хранилища astroreference и изображений, извлеченных из статей, которые можно использовать для анализа. Система ADSASS будет опираться на постоянно обновляемую базу данных тэгов, которая предназначена как для обнаружения новой информации о небесных объектах по любой тематике, так и для поиска событий переменного характера по данным исторического слоя.

Литература

1. Szalay A. and Gray J., 2020 Computing: Science in an exponential world, http://www.nature.com/news/2006/060320/full/440413a.html.
2. International Virtual Observatory Alliance. Documents & Standards. http://www.ivoa.net/documents/.
3. Желенкова О.П., Черненков В.Н., Шергин В.С., Пляскина Т.А., Витковский В.В. Программные системы и информационные ресурсы для обеспечения астрофизических исследований. Инфраструктура спутниковых геоинформационных ресурсов и их интеграция. Сб. научных статей под ред. М.А. Попова и Е.Б. Кудашева, Киев, Карбон-Сервис, 2013. -192 стр. С. 167-173.
4. О.П. Желенкова, В.Н. Черненков, Т.А. Пляскина, В.С. Шергин, В.В. Витковский. Интеграция программных систем поддержки астрофизических наблюдений. XIV Всероссийская конференция RCDL-2012. Труды конференции, с. 229-235, изд. УГП им.Айламазяна РАН, ISBN 978-5-901795-30-9.
5. C. Pennypacker, M. Boer, R. Denny, F. V. Hessman, J. Aymon, N. Duric, S. Gordon, D. Barnaby, G. Spear and V. Hoette. RTML - a

standard for use of remote telescopes. A&A Vol. 395, N 2, 727 – 731 (2002).
6. A. Klotz. Protocols for Robotic Telescope Networks. Advances in Astronomy, Article ID 496765, 8 p. (2010).
7. Laurino O., Smareglia R. VOdka: A Data Keeping-Up Agent for the Virtual Observatory // Astronomical Data Analysis Software and Systems XX. – Boston:ASP, 2011, 442, 571-574.
8. Gray N., Mann R.G., Morris D., Holliman M., Noddle K. AstroDAbis: Annotations and Cross-Matches for Remote Catalogues // e-print, 2011, arXiv:1111.6116, 1-4.
9. Pepe A., Goodman A., Muench A. The ADS All-Sky Survey // e-print, 2011, arXiv: 1111.3983, 1-4.

WPS-СЕРВИСЫ ПРОСТРАНСТВЕННОГО АНАЛИЗА СОСТОЯНИЯ ОКРУЖАЮЩЕЙ СРЕДЫ И ПРИРОДНЫХ РЕСУРСОВ

Р.К. Фёдоров, А.С. Шумилов

Институт динамики систем и теории управления СО РАН,
Иркутск, Россия
fedorov@icc.ru

В статье описывается разработка среды распределенных вычислений на основе стандарта Web Processing Service, который унифицирует использование Web-сервисов, предоставляющих услуги пространственной обработки растровых и векторных данных.

Ключевые слова: OGC, WPS, JavaScript, V8, Calipso

The article describes the development of the WPS-based (Web Processing Service) distributed process environment, which simplifies the use of different web-services, offering processing and analysis of the geospatial data.

Keywords: OGC, WPS, JavaScript, V8, Calipso

7. Введение

Решение задач анализа состояния окружающей среды и природных ресурсов, как правило, требует применения нескольких программных систем (ПС), разрабатываемых специалистами различных предметных областей. Например, моделирование загрязнения атмосферного воздуха населенных пунктов используются методы инвентаризации источников загрязнений (печное отопление, автотранспорт и т.д.) и моделирования распространения загрязнений. Эти методы являются достаточно сложными и часто реализовываются разными коллективами. Взаимодействие коллективов осуществляется редко в частности из-за сложности использования и интеграции пакетов программ. В итоге при решении сложных задач решается одна или несколько подзадач на

должном уровне в зависимости от ресурсов коллектива, а остальные подзадачи решаются упрощенным способом.

На сегодняшний день активно развивается взаимодействие между программными системами (интероперабельность) через Интернет, используя стандарты Open Geospatial Consortium (OGC) [1]. Одним из перспективных стандартов OGC является Web Processing Service (WPS) [2], который унифицирует использование сервисов обработки пространственных данных (ПД) через Интернет. Например, это могут быть сервисы, реализующие пространственную обработку растровых и векторных данных, геомоделирование, методы статистики. Достоинствами данного стандарта являются простота, возможность предоставления метаданных, поддержка длительного выполнения сервисов и т.д. В области обработки ПД он позволяет объединить и интегрировать программные системы, созданные разными разработчиками, для решения сложных задач, требующих объединения специалистов в различных предметных областях. В статье рассматривается подход, позволяющий объединить распределенные данные и WPS-сервисы для решения задач.

8. Распределенная вычислительная среда

Для эффективного взаимодействия WPS-сервисов и данных требуется организация распределенной вычислительной среды. Вычислительная среда должна предоставлять возможность хранения данных и выполнения алгоритмов (программ) над этими данными. Соответственно распределенная вычислительная среда должна пользователю и WPS-сервисам предоставлять доступ и возможность размещения в Интернет исходных данных, промежуточных данных и результатов вычислений (обработки) и выполнение распределенных WPS-сервисов.

В ИДСТУ СО РАН разрабатывается геопортал, который реализует основные функции среды. Геопортал предоставляет пользователям функции хранения данных в следующем виде:

1) файлов в представляемой пользователю директории файловой системы геопортала на базе системы хранения данных (СХД) SAN ReadyStorage 3994;

2) реляционных таблиц в СУБД PostgreSQL.

Геопортал предоставляет файловый менеджер для работы через Интернет с файловой системой и подсистему для создания, редактирования, отображения реляционных данных.

Для организации выполнения WPS-сервисов необходимо учитывать их распределение в сети Интернет. Соответственно необходимо обеспечить инвентаризацию и поиск WPS-сервисов. Кроме того требуется более детальная спецификация входных и выходных параметров, чем это определено стандартом WPS. Детальная спецификация позволит упросить пользовательский интерфейс и проводить верификацию входных данных еще до запуска WPS-сервиса. Инвентаризацию и поиск WPS-сервисов реализует разработанный в рамках проекта каталог. Каталог WPS-сервисов разработан в виде модуля системы управления контентом Calipso. Регистрация WPS-сервиса в каталоге происходит в несколько этапов. На первом этапе пользователь вводит: название сервиса, его описание, данные для обращения к WPS-службе. На втором этапе каталог запрашивает, в соответствии со стандартом WPS, по введенному адресу метаданные и отображает список имеющихся WPS-сервисов. После выбора нужного сервиса, выполняется запрос на получение метаданных о параметрах WPS-сервиса. На последнем этапе пользователь дополняет информацию о параметрах: используемый для ввода элемент управления и его свойства, пользовательское название параметра, поясняющий текст. Данная информация применяется для генерации поль-

зовательского интерфейса, верификации параметров и запуска WPS-сервиса. Поиск в каталоге может производиться по названию и описанию работы WPS-сервиса.

Для запуска пользователем WPS-сервиса генерируется форма ввода параметров. Для каждого параметра используется элемент управления, указанный при регистрации WPS-сервиса. Разработаны следующие элементы управления: edit – для ввода строковых значений; number – для ввода числовых значений; checkbox – для ввода булевых значений; rectangle – для указания экстента (прямоугольной области на карте) и т.д. При запуске WPS-сервисов необходимо упростить указание данных из СХД и СУБД в качестве значений входных и выходных параметров.Поэтому разработаны элементы управления, которые работают с данными пользователя в СХД и СУБД: file – для выбора файла из СХД; file_save – для сохранения файла в СХД; select_table - для выбора таблицы из СУБД; select_table_attr - для выбора атрибута таблицы и т.д. При наличии на форме элементов управления file, file_save, rectangle создается карта, на которой отображаются данные, связанные с этими элементами. Набор элементов управления является расширяемым.

После ввода пользователем в форме Web-браузера данных среда должна обеспечить запуск WPS-сервисов и передачу, получение этих данных в соответствии со стандартом WPS и политиками регламентации доступа. В форме WPS-сервиса формируется список значений параметров и передается запрос подсистеме управления WPS-сервисами на выполнение. Подсистема управления сервисами производит обработку параметров для их передачи в соответствии со стандартом WPS, регистрирует экземпляр выполняемого WPS-сервиса, запускает WPS-сервис, получает и сохраняет результаты работы. В зависимости от типа параметров обрабатываются значения следующим образом:

1. Файл. Если передаваемый параметр является файлом, расположенным в директории пользователя в СХД, то необходима передача данного файла WPS-сервису. В соответствие со стандартом WPS-сервису передаются URL адрес файла, используя который, WPS-сервис должен по протоколу HTTP скачать файл. Файл должен находиться в открытом доступе. Так как данные пользователей могут содержать информацию ограниченного доступа, то любой файл, хранимый в СХД, по умолчанию не может быть свободно доступным. Таким образом, необходимо одновременно защитить и предоставить WPS-сервисам доступ к хранимым в СХД файлам. Механизм доступа и контроля обращений к файлам построен следующим образом – каждому экземпляру выполнения WPS-сервиса присваивается уникальный идентификатор. Если выполняемый WPS-сервис должен получить в качестве параметров файлы, то для каждого из файлов создается уникальная ссылка, привязываемая именно к выполняемому экземпляру. Таким образом, WPS-сервис может свободно скачать требуемый файл из СХД по сгенерированной ссылке определенное число раз, причем все ссылки, созданные для определенного экземпляра, уничтожаются как только метод завершит свою работу. Если файл является результатом работы WPS-сервиса, то подсистема выполнения сценариев загружает в СХД в соответствующую пользовательскую директорию по протоколу HTTP. URL–адреса файлов берутся из метаинформации в ExecuteResponse документе, который формируется по окончании работы WPS-сервиса.

2. Экстент (прямоугольная область на карте). На клиентской части (браузере) получение экстента производится в формате WKT [1]. Для передачи WPS-сервису такого типа данных производится преобразование в соответствии со стандартом WPS.

3. Таблица базы данных PostgreSQL. При передаче в качестве параметра ссылки на таблицу формируется строка соединения

драйвера GDAL для непосредственного соединения WPS-сервиса с базой данных. В текущей версии таблицы PostgreSQL доступны только для локальных WPS-сервисов, но в дальнейшем будет реализация для регламентированного доступа к базе данных извне.

Интеграция WPS-сервисов осуществляется в виде сценариев, определяющих последовательность применения, передаваемые параметры и т.д.. Для разработки сценариев WPS-сервисов предлагается использовать язык JavaScript, где обращение к WPS-сервисам производится с помощью специальных функций. Использование языка JavaScript обладает рядом преимуществ, в частности JavaScript является полноценным языком программирования с поддержкой асинхронного выполнения и сохранением контекста, что очень важно, учитывая возможную длительность выполнения WPS-сервисов. Для интерпретации сценариев на языке JavaScript и непосредственного обращения к WPS-сервисам разработан специальный модуль, написанный на C++ с использованием JavaScript интерпретатора Google V8.

9. Разработанные WPS-сервисы

В рамках геопортала разработан ряд сервисов. Перечислим некоторые из них:

Сервис расчета плотности точечных объектов в ячейках регулярной сетки. На входе сервиса слой векторных объектов. На выходе количество объектов, находящихся в ячейках регулярной сетки. Пользователь может задать размер ячейки и область обработки. Сервис производит подсчет количество объектов в ячейках, а если задан атрибут семантики слоя входных данных, то производится суммирование значенийуказанного атрибута. Данный сервис используется, например, для расчета выбросов от точечных объектов.

Сервис расчета плотности линейных объектов в ячейках регулярной сетки. На входе сервиса слой линейных векторных объектов. На выходе общая длина участков линейных объектов, находящихся в ячейках регулярной сетки. Пользователь может задать размер ячейки и область обработки. Сервис для каждого линейного объекта производит трассировку и суммирование длин участков линейных объектов в ячейках, а если задан атрибут семантики слоя входных данных, то производится умножения длины участка объекта в ячейке на значение атрибута, а затем общее суммирование. Данный сервис используется для расчета выбросов дорожной сети.

Сервис изменения значений атрибутов.Данный сервис необходим для массового задания значений атрибута объектов, находящихся внутри полигонов. На входе сервиса название атрибута источника, слой источник полигональных объектов с желаемыми значениями атрибута и результирующий векторный файл (точечный или линейный), в котором значения атрибута меняются. Например, данный сервис можно использовать для задания характеристик печного отопления в определенных районах населенного пункта.

Сервис интерполяции точечных данных на ячейках регулярной сетки методом естественных соседей. На входе сервиса слой точечных векторных объектоввиде векторных файлов (поддерживаемых библиотекой GDAL) или таблицы PostgreSQL. На выходе интерполируемые значения в ячейках регулярной сетки, в формате GeoTIFF.Метод естественных соседей используется из библиотекиCGAL.

Сервисы построения карт рельефа, уклонов, экспозиции на основе данных радарной топографической съемки (SRTM) на

указанную территорию.Пользователь указывает экстент желаемой территории иразмер ячейки. Сервис автоматически определяет требуемые файлы данных SRTM и выполняет обработку.Результатработы сервисовсохраняется в виде набора файлов в формате GeoTIFF.

Сервис расчета нормализованного разностного вегетационного индекса (Normalized Difference Vegetation Index – NDVI) на основе мультиспектральных космических снимков. Производит расчет на основе данных пользователей, которые должны находиться в системе хранения данных геопортала. Результат сохраняется в формате GeoTIFF.

Сервисы алгебры GRID, которые позволяют производить умножение на число, сложение, вычитание GRID данных. С помощью данных сервисов,например, можно объединить набор GRID файлов.

Сервисы, обеспечивающие пространственный статистический анализ территорий. Вычислительное ядро сервиса составляют математические методы, реализованные в известной библиотеке AlgLib Free edition. В созданном сервисе реализованы возможности анализа пространственных данных через нахождение коэффициентов корреляции и регрессии. На вход для анализа поступают два файла формата GeoTIFF, описывающих различные характеристики одной территории. Для описания серии файлов используется специализированный файл в формате MTIFF. В этом случае серия GeoTIFF файлов позволяет описать динамику изменения некоторой характеристики в течение времени. Информация из GeoTIFF файлов экстрагируется и представляется в виде одномерных массивов данных с плавающей точкой. Между этими данными проводится анализ и поиск зависимостей.

Заключение

Комбинация WPS-сервисов и подсистем геопортала СХД, СУБД PostgreSQL позволяет упростить обработку пространственных данных в Интернет. Для передачи данных WPS-сервисам достаточно разместить их в геопортале с помощью подсистем ввода/редактирования данных, файлового менеджера и FTPS-сервера. Результаты работы WPS-сервисов размещаются также в геопортале. В дальнейшем каталог WPS-сервисов может выступать как общедоступная база сервисов, решающих задачи разных предметных областей. Возможность интеграции сервисов посредством специальных Javascript функций, делает геопортал удобной средой разработки новых методов поддержки междисциплинарных научных исследований.

Литература

1. OGC 05-007r7, OpenGIS® Web Processing Service / редактор: Peter Schut [Open Geospatial Consortium, Inc., 2007]. URL: http://www.opengeospatial.org/standards/wps (датаобращения: 21.10.2012).
2. OpenGIS Web Processing Service (WPS) Implementation Specification, v1.0.0. Release date: June 08, 2007. – URL: http://www.opengeospatial.org/standards/wps [15 февраля 2012]

ТЕСТИРОВАНИЕ ПЕРЕДАЧИ БОЛЬШИХ ДАННЫХ В ВИРТУАЛЬНОЙ СРЕДЕ И ЧЕРЕЗ СЕТЬ ИНТЕРНЕТ

С.Э. Хоружников, В.А., Грудинин, А.Е. Шевель, В.Б. Титов, О.Л. Садов, Е.И. Корытько, А.Е. Шкребец, О.И. Лазо, А.А. Орешкин, А.Б. Каирканов

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
Санкт-Петербургский государственный университет (СПбГУ)
sdn.ifmo.ru

Передача больших данных играет важную роль для развития науки и бизнеса. Некоторые научные эксперименты генерируют десятки петабайт данных в год. Существует несколько различных инструментов-утилит для передачи больших объемов данных. Как правило, такие инструменты пишутся самими исследователями под конкретные задачи.

Тестируются различные утилиты: bbcp, bbftp, FDT, gridFTP и др. Исследуется возможность использования параллельных каналов связи и управления ими в реальном времени. Для тестирования используются сервера с ОС Scientific Linux 6.5, виртуальная среда под управлением Openstack Icehouse, симуляция сети Интернет. Приведены результаты первых тестов, графики и выводы.

Ключевые слова: данные, передача, Линукс, сеть.

The transfer of Big Data over computer network is important and unavoidable operation in the past, now and in any feasible future. Some science experiments generates tens of petabytes per year. There are a number of methods to transfer the data over computer global network (Internet) with a range of tools. In this paper the transfer of one piece of Big Data from one point in the Internet to another point in Internet in general over long range distance.

Several free of charge systems to transfer the Big Data are analyzed here: bbcp, bbftp, FDT, gridFTP, etc. In testing are used servers under Scientific Linux 6.5, Openstack Icehouse and simulation on Internet. Here are results of first tests.

Keywords: data, Linux, transfer, network.

Введение

Проблема больших данных 1 известна уже несколько лет. Со временем меняется объем и характер данных. Возникают вопросы: как хранить, передавать и обрабатывать большие данные. Рассмотрим проблему передачи больших данных по глобальным компьютерным сетям.

Источники больших данных

Известен целый ряд так называемых «генераторов» больших данных 2 - 10. В основном это наука – исследования, эксперименты, новые масштабные проекты.

В наших исследованиях большие данные – это 100 Тб и более. Возможно, в дальнейшем эта цифра увеличится. Время передачи по глобальным компьютерным сетям (Интернет) зависит от реальной пропускной способности канала и объема передаваемой информации. Если у нас канал 1 Гбит/с, мы получаем примерно 100 МБ/с, т.е. передача 100 ТБ займет 100 000 с = 277,8 часа = 11,6 дня. В это время параметры сети могут меняться. Например, процент потерянных пакетов, реальная пропускная способность, доступность канала и другие параметры могут меняться непредсказуемо. Также важно учесть большое количество сетевых настроек ядра линукса. Наиболее важные из них - TCP Window size, MTU и др.

Инструменты/утилиты для передачи больших данных

Идеи сравнения систем передачи данных

- Многопотоковый режим передачи данных – возможность использования нескольких TCP потоков параллельно.
- Многоканальный режим передачи данных – возможность использования более чем одного канала параллельно.

- Возможность настройки низкоуровневых параметров, например, TCP Window size.
- Возможность возобновления передачи данных после ошибки

Передача данных состоит из нескольких шагов: чтение данных с диска, передача по сети, запись полученных данных на удаленном компьютере. Нас интересует именно передача данных по сети.

Низкоуровневые системы передачи данных

- UDT 11 – библиотека, которая использует для передачи udp протокол, а не tcp. В некоторых случаях помогает увеличить использование канала и уменьшить время передачи.
- RDMA 12
- MP TCP 13 – протокол, позволяющий использовать несколько каналов параллельно для одной передачи.
- Openssh 14
- BBCP 15
- BBFTP 16
- Xdd 17 – утилита, разработанная для оптимизации передачи данннных и I/O процессов для систем хранения
- FDT 18 – Java-утилита для многопотоковой передачи
- gridFTP

Среднеуровневые системы передачи данных

- FTS3 19 – продвинутый инструмент для передачи больших объемов информации.

- SHIFT 20 – интересная разработка для передачи данных в LAN и WAN.

Высокоуровневые системы передачи данных

PhEDEx – Physics Experiment Data Export используется и разработано для экспериментов CERN 7. Эксперименты генерируют большое количество данных (130 ПБ в 2013г.), которые требуется обрабатывать в кластерах, разнесенных территориально (около 10 в различных странах и континентах). Так как между сайтами как правило несколько каналов передачи, PhEDEx позволяет использовать альтернативный маршрут, если текущий не доступен.

Испытательный стенд

На данный момент стенд состоит из двух серверов HP DL380p Gen8 E5-2609, Intel(R) Xeon(R) CPU E5-2640 @2.50GHz, 64 GB с ОС Scientific Linux 6.5. Для тестирования каждой утилиты используется по две виртуальные машины. Одна как передатчик, другая – приемник. Другими словами, для тестирования в виртуальной среде используется 10 виртуальных машин под управлением Openstack Icehouse, также используется PerfSonar для мониторинга каналов связи.

Для изучения особенностей передачи данных различного типа создается директория с файлами случайной длины. Общий объем данных, средний размер файлов и др. параметры задаются при создании тестовой директории. Содержимое файлов подобрано так, чтобы избежать возможного влияния сжатия данных во время передачи.

На начальном этапе планируется провести сравнительное тестирование всех вышеупомянутых систем передачи данных в локальной сети, чтобы убедиться, что все работает правильно. Планируется записывать и сохранять все параметры (/proc) и логи во

время тестирования. Это позволит в дальнейшем разобраться, что повлияло на результаты. Все эти действия автоматизированы и включены в скрипты для тестирования. Скрипты и описания их работы доступны на Гитхаб https://github.com/itmo-infocom/BigData.

Рис. 1. Средняя скорость передачи 100 ГБ.

Тестирование проводилось на развернутых тестовых виртуальных машинах host-10-10-20-21 (BigData-20G-1) и host-10-10-20-23 (BigData-20G-2) с установленной системой НауЛинукс 6.5.

Операция передачи данных выполнена 20 раз. На основе полученных данных составлен график (Рисунок 1): номер исследования по горизонтали и средняя скорость передачи данных по вертикали. Среднее арифметическое скорости передачи 15 измерений (измерения № 1-13, 15, 17) составляет 41301.6 КБ/с.

С целью проверки, как это повлияет на скорость передачи данных, проведены несколько исследований – операция передачи

данных осуществлялась параллельно с операцией создания ещё одного тестового каталога. В ходе исследований № 14 и 16 операция создания тестового каталога параллельно с передачей была запущена на машине host-10-10-20-21. В ходе исследования № 18 операция создания тестового каталога параллельно с передачей была запущена на машине-приемнике host-10-10-20-23. В этих исследованиях (№ 14, 16, 18) процесс создания каталога, запущенный на виртуальных машинах параллельно с операцией передачи данных, снизил среднюю скорость до 29408 КБ/с, т.е. на 28.8 %. В ходе исследований № 19 и 20 операция создания тестового каталога параллельно с передачей была запущена на железной машине, на которой развернуты виртуальные машины. Здесь скорость передачи снизилась до 40720.5 КБ/с, т.е. на 1.4 %.

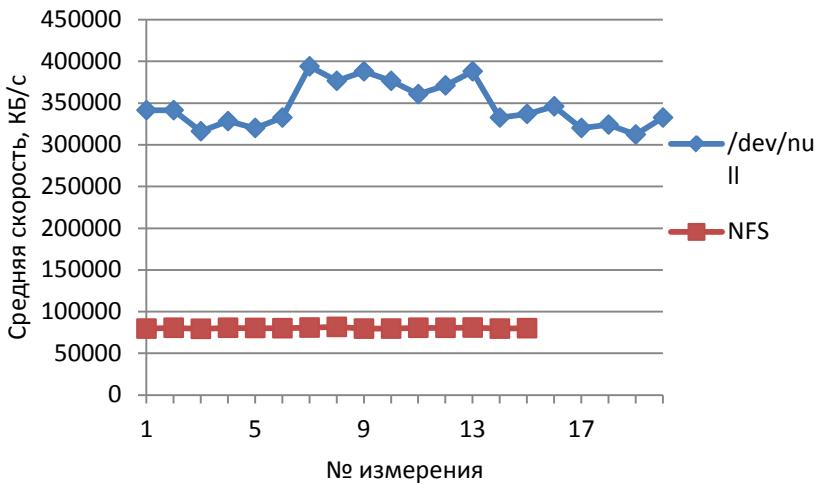

Рис. 2. Средняя скорость передачи 25 ГБ.

В ходе исследований возникло предположение, что фактически измеряется быстродействие дисковой памяти, а не скорость передачи по виртуальной линии связи между двумя виртуальными машинами. Один из вариантов исключения влияния быстродействия дисковой памяти - это писать в /dev/null. Серия тестов (рисунок 2) показывает скорость передачи по линии. Один и тот же массив данных 25 ГБ передавался на дисковую память, подключенную по NFS, а затем в /dev/null. При монтировании исходных файлов через NFS из основной памяти и записи переданных (принятых) данных в дисковый том, который монтирован по iSCSI, средняя скорость передачи составляет примерно 80 МБ/с. Аналогичный тест с устройством /dev/null в качестве устройства записи показывает скорость 300-400 МБ/с. Скорость передачи данных значительно выше, т.к. теперь измеряется именно скорость передачи по виртуальному каналу связи, а не скорость записи на диск или чтения с диска.

Список источников

1. Big Data – http://en.wikipedia.org/wiki/Big_data.
2. Information Revolution: Big Data Has Arrived at an Almost Unimaginable Scale // http://www.wired.com/magazine/2013/04/bigdata/.
3. Square Kilometer Array - http://skatelescope.org/.
4. Large Synoptic Survey Telescope - http://www.lsst.org/lsst/.
5. Facility for Antiproton and Ion Research –http://www.fair-center.eu/.
6. International Thermonuclear Experimental Reactor – http://www.iter.org/.
7. CERN - http://www.cern.ch/.
8. Lucinda Borovick Richard L. Villars // White paper. The Critical Role of the Network in Big Data Applications //

http://www.cisco.com/en/US/solutions/collateral/ns340/ns517/ns224/ns944/critical_big_data_applications.pdf [last read 1.12.2013]

9. The Center for Large-scale Data Systems Research at the San Diego Supercomputer Center // http://clds.sdsc.edu/ [last read 1.12.2013]
10. William E. Johnston, Eli Dart, Michael Ernst, Brian Tierney // Enabling high throughput in widely distributed data management and analysis systems: Lessons from the LHC // https://tnc2013.terena.org/getfile/402 *(text) and* https://tnc2013.terena.org/getfile/716 *(presentation)*
11. UDT: Breaking the Data Transfer Bottleneck - http://udt.sourceforge.net/.
12. Brian Tierney, Ezra Kissel, Martin Swany, Eric Pouyoul //Efficient Data Transfer Protocol for BigData - www.es.net/assets/pubs_presos/eScience-networks.pdf // Lawrence Berkeley National Laboratory, Berkeley, CA 94270 // School of Informatics and Computing, Indiana University, Bloomington, IN 47405
13. MutiPath TCP – Linux Kernel Implementation – http://mptcp.info.ucl.ac.be/, http://multipath-tcp.org/
14. OpenSSH http://openssh.org/
15. BBCP – utility to trasnfer the data over network - http://www.slac.stanford.edu/~abh/bbcp/.
16. BBFTP - Utility for bulk data transfer - http://doc.in2p3.fr/bbftp/.
17. Stephen W. Hodson, Stephen W. Poole, Thomas M. Ruwart, Bradley W. Settlemyer // Moving Large Data Sets Over High-Performance Long Distance Networks // Oak Ridge National Laboratory, One Bethel Valley Road, P.O. Box 2008 Oak Ridge, 37831-6164 // http://info.ornl.gov/sites/publications/files/Pub28508.pdf [1.12.2013]
18. Fast Data Transfer - http://monalisa.cern.ch/FDT/.
19. File Transfer Service – FTS3 – http://www.eu-emi.eu/products/-/asset_publisher/1gkD/content/fts3; https://svnweb.cern.ch/trac/fts3
20. Data Transfer Tools http://fasterdata.es.net/data-transfer-tools/

СЕМАНТИЧЕСКАЯ ИНТЕГРАЦИЯ БИБЛИОТЕЧНЫХ ДАННЫХ

В.А.Серебряков[a], *К.Б.Теймуразов*[a], *О.Н.Шорин*[b]

[a] Вычислительный центр им.А.А.Дородницына РАН, Москва
[b] Российская национальная библиотека, Санкт-Петербург

В Российской государственной библиотеке и Российской национальной библиотеке начат совместный проект, целью которого является публикация библиотечных данных библиотек, входящих в состав Национальной электронной библиотеки, в соответствии с принципами Linked Open Data. Реализация данного проекта позволит получить доступ к библиографической информации, хранящейся в ряде крупнейших библиотек России, в виде, пригодном для машинной обработки. Набор данных состоит из нескольких десятков миллионов записей. В процессе семантической интеграции предстоит решить ряд актуальных задач: разработка онтологии предметной области, конвертация библиотечных данных из различных MARC-форматов в RDF, публикация данных и предоставление SPARQL точек доступа к ним. Наличие открытого доступа к одному из самых крупных в мире массиву библиографической информации с возможностью обнаружения семантически связанных данныхбудет являться одной из составляющих развития как культуры в целом, так и отдельных направлений книжной отрасли в частности.

Ключевые слова: Открытые связанные данные, семантическая паутина, связанные данные, библиографическая запись.

The Russian State Library and Russian National Library launched a joint project, the aim of which is to publish the library in accordance with the principles of Linked Open Data. This project will provide access to bibliographic information stored in a number of the largest libraries of Russia, in a form suitable for machine processing. The data set consists of several tens of millions of records. A number of pressing problems will be solved in the process of semantic integration: the development of the domain ontology, the conversion of library data from various MARC-formats to RDF, the publication of data and the provision of SPARQL access points to them.

Keywords: Linked Open Data, semantic Web, bibliographic record.

В большинстве своем, информация, представленная на сайтах, предназначается для людей, поскольку основу интернета составляет гипертекст. Это означает, что основный смысл, значение скрывается в самом тексте, что значительно усложняет процесс извлечения этой сути, пригодного для автоматизированной обработки. В 2006 году Тимом Бернерсом-Ли была предложена надстройка над существовавшим интернетом, которая позволила бы автоматизированным системам извлекать информацию, анализировать её, устанавливать взаимосвязи и генерировать новую информацию. Такой подход он назвал «семантической паутиной».

Тим Бернерс-Ли предложил использовать термин «связанные данные» для реализации семантической паутины. Основное отличие семантической паутины заключается именно в термине «данные», которое ставилось в противовес существовавшему на тот момент, пусть и «гипер-», но всё же «тексту». В нашей жизни мы оперируем множеством данных: информация о стоимости продуктов в магазине, расписание авиарейсов, информация об авторстве литературного произведения.

Анализируя данные, человек может принять взвешенное решение. Например, имея данные о наличии и стоимости книги в разных книжных магазинах, а также информации о месторасположении и часах работы этих магазинов, человек способен сделать выбор и купить необходимую ему книгу по оптимальной цене в близлежащем работающем магазине. К сожалению, автоматизировать этот процесс в терминах гипертекста чрезвычайно сложно [1].

Для оперирования данными необходимо было решить несколько ключевых вопросов [2]:

- Каким образом обеспечить доступ к данным, чтобы их можно было повторно использовать?

- Как должно происходить обнаружение данных, связанных с уже имеющимися данными?
- Как приложения должны интегрировать разнородные данные, полученные из большого числа заранее неопределенных источников?

Также как World Wide Webизменил способы работы с текстом, с документами, необходимо было придумать механизмы поиска, доступа, интеграции и использования данных.

Тим Бернерс-Ли сформулировал четыре основных принципа связанных данных [3]:

1. Применение универсальных идентификаторовURIв качестве имен сущностей.
2. ПрименениеHTTPURIдля реализации возможности обращения по именам, чтобы они могли быть найдены как людьми, так и программными системами.
3. Предоставление полезной информации о сущности при обращении по URI, используя стандартизованные форматы.
4. Включение ссылок на другие связанные URI для облегчения поиска.

Для реализации этих принципов было предложено использовать модель преставления данных RDF (Resource Description Framework), которая пригодна для машинной обработки. Структурно выражения в RDFпредставляют собой триплеты. Каждый триплет состоит из субъекта, предиката и объекта. Выражение RDF-триплета означает, что отношение, указанное предикатом, связывает предметы, обозначенные как субъект и объект [4]. Например, предикат «является автором» может связывать субъект «Достоевский» и объект «Преступление и наказание». Основная

идея RDFсостоит в том, чтобы показать взаимосвязь одних данных с другими.

RDF не является форматом, а представляет собой абстрактную модель для описания взаимоотношений между данными в виде триплетов. Для сериализации RDF-триплетов существует несколько способов. Наиболее распространенным способом является представление в виде XML -RDF/XML. Синтаксис RDF/XMLстандартизован консорциумом W3C и широко используется для публикации связанных данных в интернете.

Для встраивания RDF-триплетов непосредственно в HTML-документы используют формат сериализации RDFa. Изначально RDFинформацию указывали в виде комментариев в HTML-документах, однако впоследствии триплеты стали органично встраивать в объектную модель документа (Document Object Model, DOM).

Существует способ сериализации RDF, ориентированный на создание и чтение триплетов человеком –Turtle. N-Triplesявляется подмножеством Turtle, в котором отсутствует возможность использования пространства имен (namespaces) и других методов сокращенияразмера файла, например, компактные URI (CURIE) или вложенные конструкции. В связи с этим файл, написанный с использованием N-Triples, получается гораздо больше, чем с использованием Turtleи даже RDF/XML. Но у N-Triplesесть одно неоспоримое преимущество: благодаря отсутствию механизмов сокращения размера файла каждая строка содержит в себе исчерпывающий объем информации, поэтому файл N-Triplesможет быть считан и разобран построчно.

Множество современных языков программирования поддерживают нотацию JSON, поэтому неудивительно, что существует способ сериализации RDF/JSON.

RDF представляет собой абстрактную модель представления данных с помощью триплетов и никоим образом не затрагивает семантики описываемых данных. Для выражения семантики используются словари, таксономии и онтологии, которые задаются с использованием языков RDFS (RDF Vocabulary Description Language), SKOS (Simple Knowledge Organization System) и OWL (Web Ontology Language) соответственно [5].

SKOSпредставляет собой словарь иерархически организованных терминов, а RDFS и OWL являются словарями для описания концептуальных свойств в терминах классов, свойств, экземпляров классов и операций. Например, формальная семантика OWLописывает, как получать логические следствия, т.е. факты, которые не представлены в онтологии буквально, но следуют из ее семантики.

С использованием принципов, предложенных Тимом Бернерсом-Ли, в интернете реализуется проект открытых связанных данных (Linked Open Data), целью которого является интеграция данных, информации и знаний посредством глобальных идентификаторов ресурсовURIи моделью данных RDF.

Библиотеки нашей страны хранят у себя множество различных данных: информация о записавшихся в библиотеку читателях, имеющихся в наличии книг, отсканированных образах различных изданий. Среди множества хранящихся в библиотеках данных особое значение имеет библиографическая информация, выраженная в виде библиографических записей, создаваемых непосредственно в библиотеках в процессе каталогизации книг.

«Библиографическая запись - элемент библиографической информации, фиксирующий в документальной форме сведения о документе – объекте записи, позволяющие его идентифицировать, раскрыть его состав и содержание в целях библиографического поиска. В состав библиографической записи входит биб-

лиографическое описание, дополняемое, по мере необходимости, заголовком, терминами индексирования (классификационными индексами и предметными рубриками), аннотацией (рефератом), шифром хранения документа, дополнительными точками доступа, сведениями о связи с другими библиографическими записями и другой дополнительной информацией о документе, обеспечивающей доступ к нему, датой завершения обработки документа, сведениями служебного характера» [6].

С точки зрения связанных данных библиографические записи представляют огромный интерес, поскольку хранящаяся в них информация взаимосвязана: авторы связаны со своими произведениями, сериальные издания связаны друг с другом через общую часть, издательства имеют непосредственное отношение к изданным у них книгам и т.д.

В мировом сообществе реализуется ряд проектов, направленных на публикацию библиографической информации в Linked Open Data. В частности, одним из первых проектов этом направлении являлась инициатива Библиотеки Конгресса (Library of Congress), в рамках которой было опубликовано более 260 тысяч авторитетных записей. Стоит отметить также проект создания Виртуального Международного Авторитетного Файла (Virtual International Authority File), в котором участвуют более 35 национальных библиотек мира [7]. Целью этого проекта является сопоставление одних и тех же авторитетных записей из разных библиотек мира.

Проект The Open Library можно смело назвать наиболее амбициозным, поскольку его конечной целью является создание отдельной веб-страницы для каждой выпущенной книги. На данный момент на сайте представлена информация о 20 миллионах книг и 6 миллионах авторов.

В Министерстве культуры Российской Федерации предпринимаются попытки, направленные на реализацию нового этапа развития Национальной электронной библиотеки (НЭБ). Основной целью этого этапа является обеспечение свободного, равного и всеобщего доступа граждан нашей страны к документной информации историко-культурного, научного и образовательного назначения через сеть Интернет, предоставляемой на основе единой общенациональной системы создания и эффективного использования цифровых библиотечно-информационных ресурсов и сервисов [8].

Достижение поставленной цели будет осуществляться путем решения ряда задач:

1. Формирование распределенного фонда, в состав которого будут входить актуальные научные и образовательные материалы, востребованные жителями страны произведения, социально значимая информация.
2. Обеспечение доступа к распределенному цифровому фонду путем создания единой точки доступа, предоставляющей развитый набор сервисов по поиску материалов в распределенном массиве информации.
3. Решение нормативно-правовых аспектов деятельности Национальной электронной библиотеки, в частности унификация содержания государственных заданий для различного вида библиотек с возможностью внесения изменений в перечни оказываемых электронных услуг.

В процессе реализации нового этапа развития НЭБ из библиотек различной ведомственной подчиненности будет аккумулирована уникальная по своей полноте библиографическая информация. Публикация собранных данных в семантически связанном виде выведет НЭБ в ряды лидеров проектов в мировом библио-

течном сообществе, как по объемам опубликованных данных, так и по количеству источников, участвующих в интеграции.

В Российской государственной библиотеке и Российской национальной библиотеке был реализован совместный проект, целью которого являлосьсоздание программной системы, позволяющей осуществить публикацию данных библиотек, входящих в состав НЭБ, в соответствии с принципами Linked Open Data. Архитектура данной программной системы предусматривает инфраструктурные особенности функционирования НЭБ, в частности децентрализацию процессов формирования, хранения фондов и вариативность технологических решений, используемых в отдельно взятых библиотеках – участниках НЭБ.

В рамках реализации программной системыбыл решен ряд принципиальных задач.

1. Разработка онтологии предметной области на базе существующих решений. При создании онтологии предметной области максимально использовались термины из уже существующих и широко используемых словарей [9]. Такой подход значительно снизил вероятность того, что для существующих программных систем могла потребоваться дополнительная конвертация данных или даже изменение приложения.

Были изучены проекты Библиотеки конгресса США, прежде всего стандарт METSпредставления описательных, административных и структурных метаданных цифровых библиотек. Также был исследован проект Europeana, который в качестве метаданных использует стандарт Dublin Core [10]. Немаловажным было изучение опыта проекта Delos, который выпустил документ Digital Library Reference Model. Также был учтен стандарт Publishing Requirements for Industry Standard Metadata (PRISM), разработанный издательствами для обмена метаданными о публикациях.

2. Осуществление интеграции с автоматизированными библиотечными информационными системами (АБИС) и конвертация библиографических записей в унифицированный формат. Для автоматизации процесса комплектования, каталогизации, книговыдачи, межбиблиотечного обмена большинство библиотек используют АБИС. Как следствие, все библиографические описания, имеющиеся в библиотеке, хранятся в АБИС. Библиотеки, являющиеся участниками и партнерами Национальной электронной библиотеки, используют 4 основных АБИС: Aleph, Ирбис, MarcSQL, Opac-Global.

Перечисленные АБИС имеют широкие возможности по интеграции с внешними системами, используя различные протоколы. Для каждой АБИС были изучены различные возможности подключения, позволяющие экспортировать библиографические описания. В частности, для Aleph модуль интеграции был реализован с помощью использования протокола OAI-PMH [11], а для системы Ирбис запрос библиографического описания осуществлялся по протоколу Search/Retrieve via URL версии 1.1. Интеграция с MarcSQL осуществляется путем прямого доступа к базе данных АБИС. Из Opac-Global описания экспортируются путем прямых HTTP-запросов к служебным адресам системы.

Записи, получаемые из АБИС библиотек, представлены в MARC-форматах. В России используются 2 различных формата хранения библиографических описаний: MARC21 и RUSMARC (диалект формата UNIMARC). Оба этих формата являются бинарными. MARC21 – это международный формат, разработанный Библиотекой Конгресса. Для этого формата существует множество утилит, позволяющих конвертировать файлы в MARC21/XML. В конечном итоге записи, полученные в формате MARC21, конвертируются в формат MODS [12]. В силу малой распространенности формат RUSMARC не имеет утилит по конвертации из бинарного вида в представление, основанное на использовании XML. В связи с этим был разработан конвертор, который преобразует записи из формата RUSMARC в MODS.Для экспортиро-

ванных записей было создано единое хранилище, которое ежедневно пополняется новыми и отредактированными описаниями в формате MODS из АБИС библиотек.

3. Осуществление взаимного обогащения данных из различных библиотек. В случае появления в хранилище нескольких библиографических записей на одну и ту же книгу или авторитетных записей на одного и того же автора из различных библиотек, отличающихся друг от друга по степени детализации, раскрытия информации, наличием точек доступа, ссылок, автоматически создается или уточняетсяобъединенная запись, максимально полно раскрывающая первоисточники.

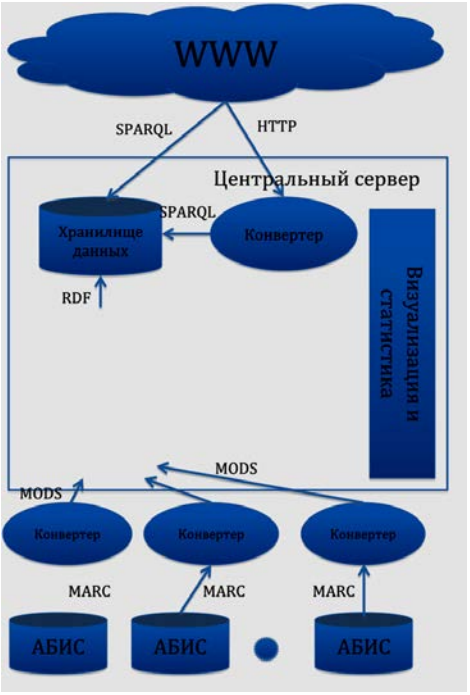

Рис. 1. Архитектура системы интеграции.

4. Конвертация и решение вопроса о хранении данных. К обогащенным записям, расположенным в едином хранилище, применяется преобразование, которое трансформирует записи из формата MODS в формат RDF, в соответствии с предметной онтологией. Для обеспечения точки доступа к RDFданным с помощью языка запросов SPARQL (SPARQL Protocol and RDF Query Language) был разработан механизм хранения RDF-триплетов. Для этого были проанализированы несколько подходов: автоматическая конвертация MODS в RDF-триплеты «на лету» для каждого запроса, хранение заранее сконвертированных данных в реляционной базе данных, хранение данных в специализированном хранилище триплетов. Каждый из этих подходов имеет свои преимущества и недостатки [13].

Например, автоматическая конвертация данных по каждому запросу не приводит к дублированию данных, но требует реализации сложной логики и обладает низкой производительностью. Хранение же триплетов, в свою очередь, приводит к дублированию данных, что требует дополнительного физического пространства и механизмов модификации триплетов в случае изменения библиографических записей. Поскольку объем информации, хранимой в библиографических записях, является несущественным, а изменение информации происходит только в одном направлении, что устраняет вероятность возникновения коллизий, было принято решение о хранении данных в специализированном хранилище триплетов.

5. Выбор данных для связывания и публикация записей в Linked Open Data. По правилам публикации данных в LOD новые сущности должны ссылаться на уже опубликованные наборы. Для этого были исследованы уже опубликованные массивы данных на предмет возможности использования их в качестве субъектов в RDF-триплетах [14].Использование специализиро-

ванного хранилища триплетов позволило автоматически создать SPARQLточку доступа к данным и обертки вокруг неё в виде обычного веб-сервера.

6. Реализация модуля визуализации полученного результата. Для отладки всего процесса публикации обогащенных записей в Linked Open Data и верификации результата был создан веб-сайт, на котором визуально отображены исходные записи, полученные из различных источников, обогащенная запись, результаты, опубликованные в LOD. Интерфейс позволяет строить отчеты на основе статистической информации и экспортировать их в различных форматах: xls, csv, xml, html.

Положительный эффект от публикации библиотечных данных в семантически связанном виде, пригодном для машинного использования, трудно переоценить. В процессе реализации этого проекта был решен ряд принципиальных задач, связанных с разнородностью используемых российскими библиотеками программных систем, форматов представления данных, протоколов взаимодействия. Для достижения поставленной целибыл использован опыт передовых библиотек мира, адаптированный к специфике каталогизации литературы в России. В результате была создана модульная система, способная с использованием минимальных усилий подключать новые библиотеки в качестве источников библиографических данных.

Список литературы

1. Berners-Lee T., James H., Lassila O. The Semantic Web. Scientific American Magazine, March 26, 2008.
2. Heath T., Bizer C. *Linked Data: Evolving the Web into a Global Data Space*. Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool.

3. Berners-Lee T. Linked Data – Design Issues. http://www.w3.org/DesignIssues/LinkedData.html
4. Спецификация языка RDF. http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
5. Спецификация языка OWL. http://www.w3.org/TR/2012/REC-owl2-syntax-20121211/
6. Российский коммуникативный формат представления библиографических записей в машиночитаемой форме. http://www.rusmarc.ru/rusmarc/format.html
7. VIAF project description. http://www.oclc.org/viaf.en.html
8. Заседание коллегии Министерства культуры РФ от 23.04.2014г. http://mkrf.ru/m/494838/
9. Hannemann J., Kett J. Linked Data for Libraries. 76TH IFLA GENERAL CONFERENCE, 2010.
10. Haslhofer B., Isaac A. data.europeana.eu. The Europeana Linked Open Data Pilot. Proc. Int'l Conf. on Dublin Core and Metadata Applications, 2011.
11. The Open Archives Initiative Protocol for Metadata Harvesting. http://www.openarchives.org/OAI/openarchivesprotocol.html
12. Metadata Object Description Schema. http://www.loc.gov/standards/mods/mods-overview.html
13. Böhme C. Towards an Infrastructure for the Synchronisation of Library Metadata. Semantic Web in Libraries, 2012.
14. Volz J., Bizer C., Gaedke M., Kobilarov G. Discovering and maintaining links on the web of data. In Proceedings of the International Semantic Web Conference, pages 650–665, 2009.

ГЕОИНФОРМАЦИОННЫЕ ВЕБ-СИСТЕМЫ ДЛЯ ЗАДАЧ ИНФОРМАЦИОННОГО ОБЕСПЕЧЕНИЯ РЕГИОНАЛЬНОГО УПРАВЛЕНИЯ

О.Э. Якубайлик

Институт вычислительного моделирования СО РАН, Красноярск

В работе рассматриваются проблемы и опыт решения задач информационного обеспечения регионального управления, связанных с использованием прикладных картографических веб-систем. Обсуждаются особенности организации и структуры разработанного программного обеспечения, практический опыт его внедрения в министерствах и ведомствах администрации Красноярского края.

Ключевые слова: веб-картография, Интернет-ГИС, банк пространственных данных, геопортал, геоинформационная веб-система, картографический веб-сервис, системы поддержки принятия решений.

The paper discusses problems and experience in design and implementation of web mapping software for information support of regional governance. The architecture and structure of the developed software is discussed along with the practical experience of its implementation in the ministries and departments of the administration of Krasnoyarsk Krai.

Keywords: web mapping, web GIS, spatial geodata, geoportal, web GIS system, web map service, decision support software system.

Интенсивное развитие технологий геоинформационных систем и глобальной сети Интернет, наблюдаемое с конца 90-х годов, привело к формированию новой парадигмы программного обеспечения для представления, обработки и анализа геопространственных данных. Речь идет о геоинформационных системах для Интернет (веб-ГИС), которые, в отличие от традиционных ГИС, могут оперировать с размещаемыми и распределенны-

ми в сети Интернет громадными массивами геопространственной информации и предусматривать удаленную обработку данных на высокопроизводительных компьютерах; они изначально являются многопользовательскими, обладают схожей с обычными настольными ГИС базовой функциональностью.

Основой современных веб-ГИС стали новые веб-технологии, такие как механизмы частичного асинхронного обновления веб-страницы (AJAX), HTML5 для построения веб-приложений, геопространственные веб-сервисы консорциума OGC, и проч. За последние годы были созданы, стали популярными многочисленные инструментальные программные средства – целый ряд успешных программных библиотек и продуктов серверного и клиентского уровня для создания интерактивных картографических веб-систем – OpenLayers и Leaflet, Sencha и GeoExt, GDAL/OGR и FDO, Jx и Fusion, MapServer и GeoServer,MapGuideOpenSource,PostgreSQL иPostGIS,Proj4js и GEOS,GeoNetwork и deegree, и многие другие. Помимо свободного программного обеспечения значительный вклад также был сделан лидерами рынка коммерческих ГИС – новые возможности для построения веб-приложений в линейке программ ESRI (ArcGIS), поддержка пространственных типов данных в Microsoft SQL Server, и т.д. Другая составная часть успеха – формирование и поддержка стандартов отрасли на обмен информацией: геопространственные веб-сервисы WMS, WFS, WCS, и проч.; развитие средств для совместного ввода и обработки картографических данных, например – OpenStreetMap. В совокупности с популярными сегодня решениями «Веб 2.0» типа кооперативного сбора данных о дорожных пробках или публикации геопривязанных фотоснимков в социальных сетях, они обеспечили новое качество массовых информационных услуг. Наконец, нельзя не отметить еще одну составляющую успешного развития веб-ГИС – наблю-

дается заметный прогресс в доступности данных дистанционного зондирования (ДДЗ).

Появление общедоступных веб-ресурсов с ДДЗ высокого разрешения и мультимасштабных карт-схем городов и территорий, содержащих разнообразные сведения о различных объектах («точки на карте»), а также соответствующие программные интерфейсы для их использования в собственных разработках (API Карт Google, Яндекс, 2GIG, и проч.) спровоцировали взрывной рост картографических веб-приложений различной направленности – интерактивные карты погоды, городские бизнес-справочники и схемы маршрутов транспорта, туристические атласы, интерактивные социально-экономические, общественно-политические и другие тематические карты и схемы. В результате можно говорить о том, что в информатике возникло новое направление исследований, связанных с построением архитектуры многозвенных геопространственных информационных систем, новый класс программного обеспечения – картографические веб-приложения, или – геоинформационные Интернет-системы. Отличия веб-ГИС от традиционных ГИС представляются достаточно значительными – ряд авторов, называя новое направление «неогеографией», даже считает, что это уже не ГИС; эта тема активно дискуссируется на различных форумах в Интернет [3].

Для рассматриваемой задачи создания программного обеспечения картографических веб-приложений – региональных геоинформационных веб-систем – чаще всего используют подход, в рамках которого принятие решения о выборе системной архитектуры – один из этапов разработки. Это связано с тем, что данная предметная область – относительно новая, в ней сейчас нет «однозначных» лидеров в области базового программного обеспечения указанного типа – наоборот, существует ряд конкурирующих

технологических решений, которые выглядят привлекательными, достаточными для обеспечения реализуемых задач.

Представляется логичным рассматривать проблему выбора технологической платформы для реализации систем указанного типа, т.к. вряд ли найдется одна универсальная программа, удовлетворяющая весь спектр возможных потребностей. На первый взгляд, анализ рынка программного обеспечения подсказывает, что сначала нужно сделать выбор одного из двух альтернативных вариантов – коммерческое программное обеспечение типа семейства приложений ESRI ArcGIS или свободное и бесплатное ПО (FOSS – free&opensourcesoftware) ГИС – «настольные» ГИС, инструментальные средства для веб-картографии, геопространственные библиотеки для чтения/записи и обработки пространственных данных, и т.д. Каждый из этих вариантов характеризуется функциональной полнотой, при этом имеет свои плюсы и минусы. Коммерческие системы требуют вложений на старте, но многие функции можно сразу использовать, а благодаря технической поддержке сроки внедрения минимальны. Открытые/свободные ГИС на практике сложнее начать использовать, но по эффективности и производительности они не уступают коммерческим, и при наличии квалифицированных специалистов всегда можно расширить их функционал [2].

Оставляя за скобками финансовый, философский и конъюнктурный аспекты выбора, хотелось бы отметить, что сегодня на практике чаще всего нет противопоставления двух рассматриваемых подходов. И причина в том, что сейчас коммерческие и свободные ГИС хорошо дополняют друг друга – благодаря совместимости форматов данных, стандартам информационного обмена, основанным на веб-сервисах, и т.д. Можно, например, выполнять анализ пространственных данных в ESRI ArcGIS, конвертировать их в MapInfo для передачи заказчику, и при этом ис-

пользовать свободное ПО Mapserver для представления на веб-страницах, а каталог пространственных метаданных формировать средствами GeoNetworkOpenSource. При этом для хранения пространственных данных использовать открытую СУБД PostgreSQL с модулем расширения PostGIS, что для подавляющего большинства задач практически не уступает по производительности и функциональным возможностям лидеру коммерческих СУБД Oracle с расширением для работы с пространственными данными OracleSpatial.

Наиболее популярная в настоящее время концепция построения картографического веб-приложения предполагает создание набора взаимосвязанных веб-сервисами компонент, выполнение которых осуществляется одновременно на компьютере-сервере и компьютере-клиенте (многозвенная архитектура), а также формирование набора пространственных данных, как правило – в формате популярных ГИС или с использованием специализированной геопространственной СУБД. При этом первоначальная подготовка геоданных для веб-приложения осуществляется, как правило, за рамками рассматриваемой веб-системы, для этого сегодня обычно используют стандартные настольные ГИС (MapInfo, ArcGIS, QGIS, и проч.) [3].

Отличительными характеристиками современных геоинформационных веб-систем стали следующие их особенности:

- интеграция картографического веб-приложения с системой управления веб-контентом, ее средствами управления доступом пользователей, администрирования и настройки интерфейса, формирования информационных блоков веб-портала;
- совершенствование пользовательского интерфейса: создание элементов управления картой и геоданными в стиле

традиционных настольных ГИС – плавающие панели с инструментами-кнопками, интерактивные древовидные раскрывающиеся меню со списками слоев карты, контекстная настройка свойств отображения данных, и т.д.;

- расширенная поддержка информационного обмена геоданными между элементами картографического веб-приложения и сторонними системами на основе открытых технологических стандартов – веб-сервисов OGC;
- оформление наборов используемых геопространственных данных в виде каталогов с соответствующими метаданными, создание самостоятельных программных средств для навигации и поиска геоинформации в этих каталогах (геопорталы).
- Опираясь на опыт выполненных разработок, попробуем сформулировать программно-технологические особенности веб-систем рассматриваемого класса с функциональной точки зрения –
- Система должна состоять из клиентской и серверной частей, реализуя тем самым технологию «клиент – сервер». Применение в основе серверной части приложения шаблона проектирования MVC (модель – представление – контроллер) предоставляет широкие возможности для решения поставленных задач. Использование данной архитектуры предполагает разделение данных приложения, пользовательского интерфейса и управляющей логики на три отдельных компонента. Модификация каждого компонента может осуществляться независимо. Помимо стандартных элементов MVC, ключевыми блоками рас-

сматриваемой системы также являются шаблоны представления страниц и AJAX-обработчики.

- Геопространственные данные системы регистрируются в каталоге ресурсов. Связь веб-приложения с каталогом ресурсов ведётся на основе сервис-ориентированной архитектуры, реализованной при помощи протокола SOAP/XML. Создается набор функций, доступный в виде прикладного программного интерфейса (API), предоставляющий возможности поиска ресурсов, их фильтрации, управления, редактирования, копирования, перемещения и т.д.
- Формируется два набора программных интерфейсов: общий (клиентский) интерфейс для пользовательских приложений и расширенный интерфейс (серверный) для приложений, имеющих возможность управления каталогом ресурсов, его объектами и отношениями между объектами.
- Информационное обеспечение прикладных региональных геоинформационных веб-систем, создаваемое в соответствии с указанными здесь принципами, состоит из нескольких типов геопространственных и прочих данных, соответствующих средств для их обработки –
- табличные данные, которые хранятся в СУБД – они обеспечивают бизнес-логику системы, ее базовое содержание.
- тематические пространственные данные, связанные с непосредственным содержанием задач и функций прикладной системы. Данные этого типа могут создаваться как в рамках интерфейсов пользователя прикладной системы

(на основе веб-браузера или приложения для Windows), так и средствами сторонних приложений – различные программы ГИС/CAD, и т.п.

- картографические подложки – информационные ресурсы вспомогательного характера – карты и мозаики спутниковых снимков, используемые при визуализации тематических пространственных данных системы. Используются как собственные наборы данных, так и сторонние сервисы, предоставляемые компаниями Яндекс, Google, 2ГИС, и др.
- средства для создания и администрирования пространственными метаданными, связанных с ними сервисами визуализации, поиска и фильтрации.
- средства для классификации геопространственных данных – стандартизация системы классификаций обеспечивает возможность множественного использования данных в различных проектах.
- инструменты для управления веб-публикацией сведений о геопространственных данных – это интерфейсы пользователя для формирования тематических разделов геопортала на основе системы управления веб-контентом.
- инструменты интеграции систем, информационного обмена и программного взаимодействия на основе информационно-аналитических и картографических сервисов геопортала.

Практический опыт разработки прикладных систем рассматриваемого класса основан на использовании СУБД PostgreSQL с модулем PostGIS для хранения пространственных данных, языка сценариев PHP для реализации бизнес-логики приложений. Фор-

мирование пользовательских и программных интерфейсов осуществлялось с помощью свободного программного обеспечения – более десятка различных сторонних программных библиотек, а также собственных разработок. Используемые программные компоненты являются лидерами в соответствующих категориях программного обеспечения, обеспечивают широкий спектр возможностей, гибкую адаптацию к требованиям создаваемых систем, возможность настройки и модернизации, оперативного исправления ошибок [1,2].

Открытая архитектура и базовое свободное программное обеспечение позволяет в кратчайшие сроки выполнить реализацию практически любых новых (дополнительных) функций в создаваемой системе, что при использовании дорогостоящего коммерческого программного обеспечения может оказаться в принципе невозможным. Дополнительным бонусом является возможность интеграции с любыми сторонними коммерческими продуктами, если в этом есть необходимость. Например, если у Заказчика уже есть и используется СУБД Oracle – можно настроить подсистему хранения пространственных данных создаваемого комплекса на использование этой СУБД. А если ее нет – установить и использовать свободное программное обеспечение СУБД PostgreSQL [4].

Перечень избранных проектов, выполненных на основе рассматриваемого подхода:

- «Банк пространственных данных Красноярского края» для Министерства информатизации и связи Красноярского края;
- «Автоматизированная система мониторинга муниципальных образований» для Министерства экономики и регионального развития Красноярского края;

- «Геоинформационная система мониторинга состояния окружающей природной среды в зоне действия предприятий нефтегазовой отрасли Красноярского края» для Сибирского федерального университета и Министерства природных ресурсов и лесного комплекса Красноярского края;
- Диспетчерско-навигационная система мониторинга автотранспорта на основе спутниковых данных ГЛОНАСС/GPS «РЕГНАСС» для Мини-стерства транспорта и связи Красноярского края;
- Геоинформационная веб-система «Сеть образовательных услуг Красноярского края» для Министерства образования и науки Красноярского края;
- «Карта здравоохранения Красноярского края» для Министерства здравоохранения Красноярского края;

Модульная архитектура рассматриваемых систем, использование стандартных веб-сервисов для обмена данными между этими модулями, обеспечивает быструю адаптацию имеющегося программного обеспечения под требования заказчика, тиражирование отдельных компонентов, их взаимодополняемость. В свою очередь, регистрация создаваемых ресурсов – пространственных данных в централизованном каталоге геопортала – обеспечивает возможность их одновременного применения в нескольких разработках. Такой сервис-ориентированный подход, основанный на активном внедрении веб-технологий в прикладные информационные системы, все чаще применяется в настоящее время.

Список литературы

1. Якубайлик О.Э., Кадочников А.А., Матвеев А.Г., Пятаев А.С., Токарев А.В. Программно-технологическое обеспечение геопортала ИВМ СО РАН // Информационные системы для научных исследований: Сборник научных статей. Труды XV Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 10-12 октября 2012 г. – СПб., 2012. – С. 143-148.
2. Якубайлик О.Э. Картографические веб-приложения и сервисы Красноярского геоинформационного портала СО РАН. – В кн.: Геоинформационные технологии и математические модели для мониторинга и управления экологическими и социально-экономическими системами: ред. кол.: Ю.И. Шокин [и др.]; под ред. И.Н. Ротановой; Рос.акад. наук, Сиб. отделение, Ин-т водных и экологич. проблем. – Барнаул: Пять плюс, 2011. – С. 94-100.
3. Якубайлик О.Э. Проблемы формирования информационно-вычислительного обеспечения систем экологического мониторинга // Вестник СибГАУ - 2012. - Вып. 3(43). - С. 96-102.
4. Матвеев А.Г., Якубайлик О.Э. Проектирование и разработка программно-технологического обеспечения для геопространственных веб-приложений // Фундаментальные исследования. – 2013. – № 10 (часть 15). – стр. 3358-3362.

ПРОГРАММНО-ТЕХНОЛОГИЧЕСКОЕ ОБЕСПЕЧЕНИЕ ГЕОПРОСТРАНСТВЕННЫХ ВЕБ-ПРИЛОЖЕНИЙ

О. Э. Якубайлик, А.А. Кадочников, А.В. Токарев

Институт вычислительного моделирования СО РАН, Красноярск

Рассматривается авторский опыт разработки программно-технологического обеспечения для создания геопортала и интеграции его компонентов, особенности организации, архитекторы программного обеспечения. Геопортал – это сложная многозвенная веб-система, ориентированная на обработку и представление пространственных данных. Разработанный набор программных библиотек для создания и интеграции отдельных компонентов веб-приложений в единую систему обеспечивает средства сквозной аутентификации и авторизации пользователей, связанные с ними компоненты пользовательского интерфейса, различные прикладные интерфейсы и функции. Рассматриваются такие элементы как подсистемы хранения геоданных, картографической веб-визуализации, поиска по метаданным, администрирования прав доступа, и пр.

Ключевые слова: геопортал, веб-картография, геопространственное веб-приложение, веб-ГИС, картографический веб-сервис, каталог пространственных метаданных, ИПД, инфраструктура пространственных данных.

Discusses the experience in the development of software and technologies for the creation of the geoportal and integration of its components. GIS web application is interpreted as complex multi-tier web system which is focused on spatial data processing and presentation. Special software is developed for the implementation of this GIS applications. This software is a collection of special purpose software libraries, including authentication and authorization of users, interface components, various APIs and functions. It also includes such elements as storage geodatabasesubsystem, web map visualization module, metadatasearch interface, administration tools, and so on.

Keywords: geoportal,web mapping, web GIS, geospatial web application, , web GIS system, web map service, spatial metadata catalog, SDI, spatial data infrasructure.

Одной из ведущих тенденций в развитии геоинформационных систем в последние годы стала их глубокая интеграция с технологиями Интернет. Однако представления о том, какой должна быть эта интеграция постоянно меняются, уточняются. Еще несколько лет назад картографический веб-сайт чаще всего рассматривался как относительно простое средство для визуализации интерактивной карты и/или мозаики спутниковых изображений. Сегодня обычно используют более сложные модели информационных систем. Как правило, говорят о геопорталах и комплексных картографических веб-приложениях различной тематической направленности, инфраструктуре пространственных данных (ИПД), сервисах для создания и обработки геоинформации в распределенной информационно-вычислительной среде. Также все большую популярность приобретают новые технологии и возможности Интернет – от популярных социальных сетей, блогов и вики-документов до сервисов облачных и распределенных вычислений. Все чаще собственные прикладные разработки интегрируются со становящимися все более доступными в Интернет различными картографическими пространственными данными, в том числе – спутниковыми снимками высокого разрешения [1].

Первым картографическим веб-приложением считают созданную в 1993 году программу MapViewer Исследовательского центра Пало-Альто компании Xerox (PARC), которая позволяла пользователям в интерактивном режиме отправлять запросы из браузера к серверу и получать фрагменты карт в формате GIF.

Именно это приложение и его функциональная концепция стало родоначальником большинства более поздних версий картографических веб-систем. В конце 90-х в процесс создания систем указанного класса включились ведущие производители программного обеспечения ГИС – ESRI, Intergraph и другие начинают разработку коммерческих приложений для веб-картографии. Примерно с этого же момента времени начинает свою историю наиболее успешный из некоммерческих проектов с открытым исходным кодом – MapServer, разработка которого началась в Университете штата Миннесота. В последующие годы и в настоящее время идет интенсивное формирование рынка веб-картографии. Возникают многочисленные программные разработки – от простых средств визуализации заранее подготовленных карт в браузере до сложных распределенных систем обработки корпоративной геопространственной информации.

Сформировавшаяся в результате концепция картографического веб-приложения предполагает создание комплекса программ, выполнение которых осуществляется одновременно на компьютере-сервере и компьютере-клиенте (многозвенная архитектура), а также формирование набора пространственных данных, как правило – в формате популярных ГИС или с использованием специализированной геопространственной СУБД (рис. 1). При этом первоначальная подготовка геоданных для веб-приложения осуществляется чаще всего за рамками стандартного веб-браузера – для этого обычно используют профессиональные ГИС (MapInfo, ArcGIS, и проч.).

Интерфейс пользователя рассматриваемых систем обычно создается в виде Интернет-ресурса с двумя видами доступа – пользовательским и административным. Пользовательский интерфейс предназначен для навигации и поиска опубликованных информационных ресурсов. Интерфейс администратора позволяет редактировать различные данные, создавать новые информационные ресурсы и управлять их публикацией, настраивать различные характеристики представления информации [2].

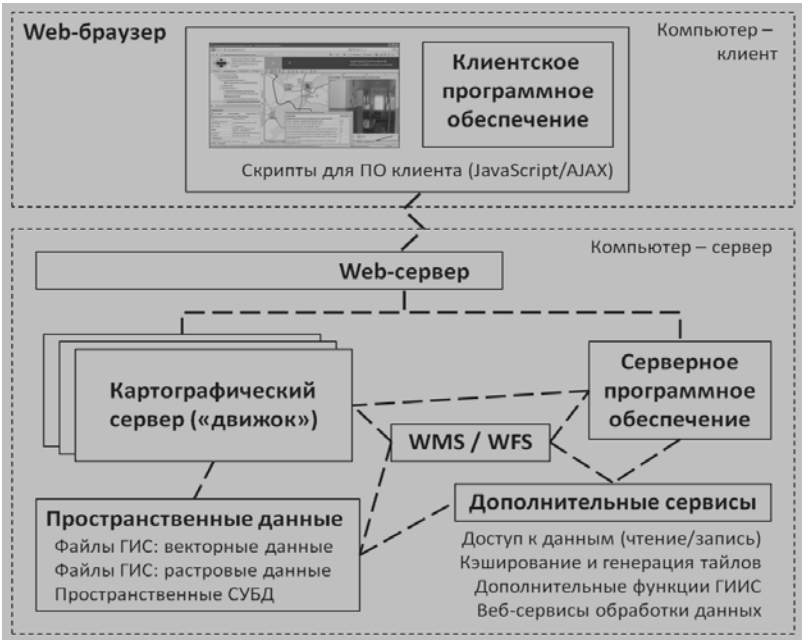

Рис. 1. Основные элементы геоинформационной веб-системы.

В результате комплексного анализа существующих подходов, проблем и решений, собственного опыта прикладных разработок в рассматриваемой предметной области, был сформирован перечень основных технических требований, которые можно предъявить к программному обеспечению геопространственного веб-приложения. Прежде всего, ее целесообразно рассматривать как распределенную информационно-аналитическую систему, основанную на гибридных технологиях – клиент-серверная и многозвенная внутренняя архитектура системы, распределенное хранение и обработка данных, ГИС и веб-технологии, прикладные веб-сервисы и стандарты информационного взаимодействия. Элементами этой системы могут и должны быть как ресурсоемкие прикладные подсистемы, выполняющие значительный объем вычислений, в том числе – с использованием производительных суперкомпьютеров, так и относительно "легкие" приложения дляпростой визуализации данных, которые могут работать и на современных мобильных устройствах (смартфонах и планшетах, не говоря уже о нетбуках и ультрабуках).

Одним из ключевых компонентов в рассматриваемой архитектуре программно-технологического обеспечения геопространственного веб-приложения является геопортал.

В соответствии с общепринятым современным пониманием термина "геопортал", его программное обеспечение как минимум должно обеспечивать решение следующих двух основных задач:

- ведение каталога пространственных метаданных, с набором необходимых операций – ввод, редактирование, удаление метаданных, и т.д.; средства поиска по метаданным – по категориям, пространственному местоположению, ключевым словам, и т.п.
- обеспечение «веб-сервисов» – средств визуализации представленных на портале пространственных данных и

их загрузки на компьютеры пользователей, интерфейсов для обеспечения прямого доступа к данным, их преобразования, и проч.

Геопорталы обычно имеют возможности разграничения прав доступа пользователей, что позволяет создавать ресурсы ограниченного доступа, формировать персональные настройки и ресурсы.

Решение рассмотренных двух задач нередко обеспечивается различными программными продуктами, которые работает совместно, дополняя друг друга. Типичный пример – поддержка каталога пространственных метаданных с помощью программного обеспечения GeoNetworkOpensource, а «веб-сервисов» – средствами комбинированной программной платформы GeoServer, объединяющей в себе сервер приложений и сервер веб-приложений, которая позволяет создавать стандартные картографические веб-сервисы и представлять их в виде интерактивных карт.

Анализ существующих подходов и тенденций, опыт собственных разработок в данной области позволил сформулировать несколько системообразующих «универсальных» компонент для рассматриваемого класса информационных систем [3,4]:

Подсистема ведения архива базовых геопространственных данных

Должна обеспечивать организацию хранения и управления данными, средства для их загрузки и удаления, резервного копирования, и т.п. Должна быть предусмотрена возможность регистрации в архиве внешних баз данных, в т.ч. – пространственных, с организацией прозрачного доступа к ним, через единый программный интерфейс – т.е. пользователь, который получает данные из архива базовых геопространственных данных не обязан

знать, откуда именно берутся геоданные – из размещенного на этом же сервере shp-файла или через подключение к стороннему серверу (ArcGISServer и аналоги).

Система прикладных программных (картографических) веб-сервисов

Предполагается создание набора средств для различного доступа к данным, организации запросов к ним, в том числе – на основе стандартных отраслевых протоколов/интерфейсов типа широко используемых открытых стандартов Консорциума OGC – картографических веб-сервисов WMS, WFS, и т.п. Также должна быть предусмотрена возможность организации ресурсоемких вычислений на стороне сервера – сегодня существует несколько альтернативных решений в данном направлении – WPS (WebProcessingService) Консорциума OGC, и пр.

Подсистема управления пространственными метаданными

Подсистема предназначена для поиска и навигации по имеющимся пространственным данным, в том числе – с помощью пространственных запросов. Подсистема должна предусматривать возможность работы с различными классификаторами данных, быть совместимой с существующими стандартами на метаданные, допускать соответствующий импорт/экспорт. Основа подсистемы – каталог метаданных – должен обеспечивать весь комплекс задач управления/администрирования метаданными. Предполагается глубокая интеграция со всеми прочими рассматриваемыми здесь компонентами прикладной геоинформационной системы.

Веб-приложение – интерфейс пользователя к данным и средствам их обработки (геопортал, картографический веб-портал).

Современное программно-технологическое решение чаще всего строится на основе веб-технологий системы управления контентом (веб-сайт). В рамках централизованного веб-интерфейса должны быть разработаны средства поиска и навигации по имеющимся пространственным и прочим данным, в том числе – с помощью пространственных запросов. Сервисы визуализации информации должны предоставлять пользователю данные в текстовом, картографическом и табличном виде. Подсистема должна предусматривать возможность работы с различными классификаторами, формами представления информации по настаиваемым шаблонам, и т.д. Должна быть обеспечена загрузка и выгрузка данных разного типа. Вся информация должна предоставляться на основе разрешений (разделение прав доступа), желательно наличие возможности персонализации интерфейса пользователя.

Веб-сервисы информационной системы экологического мониторинга можно условно разделить на две категории:

Служебные сервисы (ограниченного доступа)

Модульная архитектура разработки информационной системы экологического мониторинга формирует необходимость строгой формальной спецификации информационного обмена между ее компонентами. В этом контексте необходим набор веб-сервисов как инструмент для приема/передачи данных внутри системы, между ее составными частями. Например, модуль «Веб-портал» с помощью реализованного в нем интерфейса пользователя для поиска данных формирует запрос подсистеме «Каталог пространственных метаданных» для получения информации. Далее – подсистема «Каталог пространственных метаданных» обращается к модулю «Архив геоданных» для получения списка ресурсов определенного типа. Ограничения в доступе к служебным сервисам

связаны с тем, что с их помощью можно много чего сделать, и в том числе – нарушить работоспособность системы в целом (при неумелом использовании). Именно поэтому доступ к ним чаще всего ограничен разработчиками.

Публичные прикладные сервисы

В отличие от упомянутых выше служебных сервисов – эти сервисы являются публично доступными. Они обеспечивают выполнение различных запросов пользователей к данным, хранящимся в системе экологического мониторинга. Технологически они практически не отличаются от служебных, но нарушить работоспособность системы они не могут. Примерами сервисов являются различные операции визуализации тематической мониторинговой информации, которые используются при формировании информационных страниц веб-портала.

Список литературы

1. Матвеев А.Г., Якубайлик О.Э. Разработка веб-приложения для обработки и представления пространственных метаданных геопортала. // Вестник СибГАУ. – 2012. – Вып. 2 (42). – С. 48-54.
2. Попов В.Г., Якубайлик О.Э. Разработка модели геоинформационной аналитической Интернет-системы для задач мониторинга и анализа состояния региона // Горный информационно-аналитический бюллетень. – 2009. – Т. 17. – С. 39-44.
3. Якубайлик О.Э., Попов В.Г. Технологии для геоинформационных Интернет-систем // Вычислительные технологии.–2009. – Т.14, № 6. – С.116-126.
4. Якубайлик О.Э., Гостева А.А., Ерунова М.Г., Кадочников А.А., Матвеев А.Г., Пятаев А.С., Токарев А.В. Разработка средств информационной поддержки наблюдений за состоянием окружающей природной среды // Вестник КемГУ. – 2012. – № 4 (52) Т.2 – С. 135-141.

РАЗВИТИЕ ГЕОИНФОРМАЦИОННЫХ РЕСУРСОВ НА ОСНОВЕ ИНТЕГРАЦИИ И ОБРАБОТКИ ДАННЫХ НАЗЕМНОГО И АЭРОКОСМИЧЕСКОГО ЗОНДИРОВАНИЯ И БАЗ ГЕОДАННЫХ СРЕДСТВАМИ ГЕОПОРТАЛОВ

И.К. Лурье, А.Р. Аляутдинов, Т.Е. Самсонов

Географический факультет МГУ имени М.В.Ломоносова

lurie@mail.ru, alik@geogr.msu.ru, tsamsonov@geogr.msu.ru

Рассматривается проблема интеграции геоинформационных ресурсов и обеспечения свободного доступа к ним для обеспечения исследований окружающей среды. Суть инновационной комплексной разработки состоит в формировании базовых положений и методологии интеграции сервисов данных дистанционного зондирования, карт и баз геоданных средствами геопортала, а также в развитии поисковых и геоинформационно-аналитических инструментов для работы с пространственно-определенными данными из разных источников через геопортал и картографические веб-приложения.

Ключевые слова: пространственно-определенные данные, геоинформационные ресурсы, интеграция, базы геоданных, геопортал, геоинформационно-картографические технологии, аэрокосмическое зондирование, распределенные базы данных

In article the problem of integration of geographic information resources and providing free access to them for environmental research is considered. The essence of innovation complex development is in the construction of the basic provisions of the methodology and service integration of field spectrometry and remote sensing data, maps and geodatabases by facilities of geoportals, as well as in the development of search and geo-analytical tools for processing spatial data from different sources across the geoportal and web-mapping application.

Keywords: geospatial data, geo-information resource, integration, geodatabase, geoportal, geo-mapping technologies, aerospace sensing data, distributed database

Задачи, связанные с формированием и использованием информационных ресурсов, содержащих пространственно определенные данные, являются актуальными в области научных исследований состояния и изменений окружающей среды. Последние годы доступ к космической информации и информационным технологиям ее тематического использования существенно расширился благодаря популярности онлайновых сервисов типа Google Earth или Яндекс Карты. Активно развиваются системы коллективного доступа и обработки данных дистанционного зондирования Земли, методы создания баз пространственно привязанных данных и их накопления, в том числе, с использованием публичных Интернет-ресурсов и средств веб-картографирования.

Наиболее быстрыми темпами идет развитие онлайновых средств доступа к распределенным ресурсам пространственных данных и геоинформационным услугам - специализированных геопорталов, направленных в том числе, на обеспечение научных исследований актуальными материалами дистанционного зондирования Земли и картографической информацией, что дает мощный толчок развитию инновационной и научной деятельности в области исследования окружающей среды.

Такие геопорталы, как правило, представляют собой аппаратно-программные геоинформационные комплексы, совмещающие возможности как прямого (специальные антенны), так и серверного приема космических снимков, техническое и программное обеспечение для обработки снимков, работы с базами геоданных и онлайн-архивами, в которых размещаются как исходные материалы, так и создаваемые по ним карты и иные результаты исследований, а также систему формирования заказов на снимки (съемки) и обучение пользователей.

На кафедре картографии и геоинформатики географического факультета МГУ имени М.В. Ломоносова уже несколько лет ве-

дутся теоретические исследования и разработки геоинформационно-картографических технологий интеграции пространственно определенных данных (карт, космических снимков и баз геоданных) [2]. Они направлены на развитие новых методов современного тематического картографирования и исследования изменений окружающей среды на основе формирования структурированных и стандартизированных географических информационных ресурсов[1].

Междисциплинарные исследования динамики окружающей среды

В рамках проекта РФФИ междисциплинарных фундаментальных исследований разрабатываются новые геоинформационно-картографические технологии изучения динамики окружающей среды, основанные на создании и использовании онлайн-баз данных (спектральных библиотек) наземных гиперспектральных и многоканальных спектрометрических измерений, карт, космических снимков и результатов их дешифрирования, каталогов метаданных на геопортале для выявления взаимосвязей между наземными спектральными характеристиками объектов и их отображением на гиперспектральных и многозональных космических снимках, что способствует повышению эффективности методов дешифрирования и картографирования изменений среды в разных природных зонах [1].

На геопортале концепция междисциплинарности реализована в виде различных наборов пространственных данных, формализованных на основе многолетних наблюдений и характеризующих состояние природной среды.

[1] Исследования выполнены при поддержке гранта РФФИ (№ 13-05-12061_офи-м) и гранта Президента «Научные школы» (НШ-2248.2014.5).

Технологии обмена метаинформацией и пространственной информацией, представленной в виде наборов векторных и растровых данных, отрабатывается на базе Геопортала с оперативным наполнением и комплексом обучения и обработки информации космических снимков, созданном в 2011 г. в рамках Программы развития МГУ имени М.В. Ломоносова до 2020 г. совместно с ИТЦ «СканЭкс» - *Геопортал МГУ*.

Использование сетевых технологий является основой геопортального решения проблем интеграции сервисов. Для обмена метаинформацией наиболее приемлема технология, базирующаяся на языке XML, получившем в настоящее время широкое распространение и хорошо себя зарекомендовавшем в ряде проектов по созданию инфраструктур пространственных данных. Спецификация XML позволяет не только описывать информацию, представленную в виде структурированного документа, но и частично определяет поведение XML-процессоров – программ, обеспечивающих доступ к документам. В качестве базовых технологий обмена пространственной информацией могут быть использованы спецификации WMS (Web Map Service), WFS (Web Feature Service), WCS (Web Coverage Service), получившие также широкое применение и одобрение со стороны OGC (Open Geospatial Consortium).

Отработка технологий интеграции данных и функционально-ролевой модели доступа к данным выполнена в виде прототипа геопортала, пилотный вариант которого размещен на облачном сервисе ArcGIS Online. Разработка функционально-ролевой модели и последующее ее применение в рамках Геопортала МГУ является одной из важных задач при определении архитектуры геопортального решения.

Разработка и реализация функционально-ролевой модели доступа к информации геопортала

Функционально-ролевая модель (ФРМ) является одной из составных частей любой геоинформационной системы, особенно, реализованной на архитектуре геопортальных решений. ФРМ представляет собой элемент системы управления, определяющий правила доступа в корпоративных информационных системах с большим количеством пользователей, выполняющих свои служебные обязанности в рамках системы. На ФРМ возлагаются следующие основные функции:

- Обеспечение функциональной надежности – способность информационной системы выполнять поставленные задачи в режимах максимальной нагрузки;
- Обеспечение информационной безопасности - предотвращение или минимизация ущерба (прямого или косвенного, материального, морального или иного), наносимого субъектам информационных отношений посредством нежелательного воздействия на информацию, ее носители и процессы обработки;
- Обеспечение функциональности системы администрирования - разграничение доступа различных групп пользователей, определение уровней доступа информации;

В целом, ФРМ является реализацией определенной политики безопасности - набора правил, определяющих множество допустимых действий в системе, при этом должна быть реализована полная и корректная проверка ее условий. Система считается надежной при условии, что пользователи не имеют возможности нарушить правила политики безопасности. Общим подходом для всех моделей является разделение множества сущностей, состав-

ляющих систему, на множества субъектов и объектов, хотя сами определения понятий «объект» и «субъект» в разных моделях могут существенно различаться.

Среди общего числа математических моделей безопасности компьютерных систем принято выделять три основные ключевые модели. Это модели систем дискреционного, мандатного и ролевого разграничения доступа. Модель дискреционного доступа определяется разграничением доступа между поименованными субъектами и объектами системы. Каждая пара “субъект-объект” характеризуется жестким определением типов доступа к информации. Сам доступ определяется политикой безопасности – имеет ли право доступа субъект, который является либо отдельным субъектом, либо членом группы, взаимодействовать с информационным ресурсом – объектом. При реализации мандатной модели разграничения доступа каждый объект и субъект системы маркируются специальными идентификационными метками, определяющие место объекта или субъекта в иерархической схеме системы. Мандатное управление доступом предусматривает наличие диспетчера доступа. Данный диспетчер контролирует все обращения субъектов к объекту и определяет разграничение доступа к объекту в соответствии с заданными правилами доступа. Ролевая модель разграничения доступа базируется на принципе, когда каждому пользователю или группе пользователей определена функциональная роль с соответствующими разрешениями доступа. В данном случае можно сказать, что роль представляет собой набор правил, определяющих доступ к информационным ресурсам.

Каждая из вышеперечисленных моделей обладает своими преимуществами и недостатками, и все они не стоят на месте, постоянно развиваются. С точки разработки геоинформационной системы, реализованной на архитектуре геопортальных решений

применение ролевой модели является более предпочтительной, так как ролевая модель отвечает требованиям сервис-ориентированной структуры веб-системы с большим количеством потенциальных пользователей. С другой стороны, использование иерархии ролей позволяет эффективно реализовать управлением доступом, соответствующим ролям системы, при этом может быть реализованы дискреционное и мандатное управление доступом.

Выбрав ролевую модель управления доступа в качестве базовой, необходимо сформулировать роли или группы пользователей, которые будут соответствовать задачам геопортала и определять доступ к информации и, как следствие, их функциональные возможности. Кроме того, необходимо разработать четкую иерархическую структуру групп пользователей. В этом случае, скомпоновав пользователей в отдельные группы и их определив место в иерархической системе, формируются связи двух типов: вертикальные и горизонтальные. Горизонтальные связи работают на одном уровне иерархической системы и устанавливают взаимоотношения между пользователями одной группы, либо взаимоотношениями между группами пользователей одного уровня. Вертикальные связи определяют взаимосвязи между группами пользователей, находящихся на разных уровнях. Возможны вертикальные связи не только между соседними уровнями. Обычно, чем выше уровень, тем меньше число групп пользователей. Но, возможны варианты, когда это правило не сохраняется.

Для организации системы многопользовательского онлайн доступа к информации дистанционного зондирования, представленными данными наземных гиперспектральных многоканальных спектрометрических измерений, а также космическими снимками и картами предлагается выделить следующие группы пользователей:

- Обычные пользователи – самая крупная группа пользователей. К этой группе принадлежат пользователи, не имеющие отношения к области исследований. Функциональные возможности этой группы представлены минимальными набором функций, имеют право просматривать документы, находящиеся в общем доступе. Деление на подгруппы нецелесообразно.
- Авторизованные пользователи – группа пользователей, имеющих отношения к области исследования. Предполагается, что информационные источники портала будут использоваться этой группой пользователей целенаправленно. Возможно деление этой группы пользователей на несколько подгрупп, исходя из тематической направленности исследований специалистов для оптимизации поиска информации. Однако, функциональные возможности разных подгрупп будут одинаковы, так как подгруппы находятся на одном уровне иерархической системы. Группа пользователей имеет доступ ко всей информации, представленной на геопортале, имеет право осуществлять поиск по базам данных, скачивать доступную информацию, осуществлять отбор спектров, снимков, результатов их дешифрирования, объектов с известными спектральными свойствами, а также картографических данных.
- Группа редакторов – группа пользователей, ответственных за публикацию информационных источников на геопортале. Данная группа пользователей имеет права доступа для создания, записи и изменения документов на геопортале. Возможно даже изменение и создание отдельных разделов геопортала. Целесообразно деление данной группы на несколько подгрупп, согласно разным тематическим разделам портала. Например, группа картографов

отвечает за наполнение картографических баз данных, картографических материалов и прочее. Группа дистанционного зондирования отвечает за публикацию снимков, результатов их дешифрирования и прочее. Функциональные возможности подгрупп этой группы могут отличаться в зависимости от используемой технологии. Это может быть доступ к базе данных, доступ к картографическому серверу, доступ к отдельным документам.

- Группа разработчиков портала – группа пользователей, имеющих практически полный доступ к геопорталу. По своим функциональным возможностям обладают более полными правами доступа, чем группа редакторов. Основная задача этой группы пользователей – создание геопортала, как информационного ресурса сети Интернет. Пользователи этой группы имеют право не только менять структуру геопортала. В отличие от группы редакторов, данная группа может добавлять, удалять редактировать отдельные функциональные блоки, публиковать дополнительные сервисы и осуществлять мониторинг их использования разными группами пользователей. В эту же группу, в качестве отдельной подгруппы, могут входить специалисты, отвечающие определение тематических разделов геопортала, за подготовку информационных материалов для последующей передачи их группе редакторов, а также контролирующих корректность опубликованной информации. Необходимо отметить наличие в этой группе специалистов по веб – дизайну, разрабатывающих графическое и цветовое оформление геопортала.
- Системные администраторы – самая малочисленная группа пользователей. Основная задача этой группы обеспечение надежной работоспособности сервера, системное

администрирование, включая установку необходимого программного обеспечения, ведение учетных записей пользователей и определение политики безопасности.

Реализация пилотного проекта

Разработанная функционально-ролевая модель геопортала, представленная в виде совокупности групп пользователей и определение их основных функций, позволит существенно упростить и облегчить трудоемкую задачу по администрированию системы, таким образом, внося свой вклад в процесс разработки, управления и оптимизации геопортала.

Модель реализована на программной платформе Esri ArcGIS и включает две составляющих: картографический портал ArcGIS Portal, предоставляющий доступ к опубликованным данным, а также портал метаданных Esri Geoportal Server, описывающий опубликованные данные спектрометрирования, карты и прочие элементы в соответствии со стандартами OGC (Open Geospatial Consortium) и специализированным профилем метаданных для спектрометрических данных - GEOMS (Generic Earth Observation Metadata Standard). Использование указанного профиля метаданных для спектрометрической информации позволит, в конечном итоге, интегрироваться в мировую систему исследования окружающей среды на основе данных дистанционного зондирования.

Профиль метаданных включает группы метаданных:

- EarthObservationMetadata – общие свойства, такие как идентификатор данных, ссылка доступа и информация об архивировании,
- EarthObservationEquipment – описание оборудования, использованного для проведения спектрометрирования, включая платформу, название инструмента, тип сенсора, его технические и физические характеристики,

– EarthObservationResult – описание результатов спектрометрирования,
– Footprint – описание области спектрометрирования,
– EarthObservationObject – описание объекта – новая группа метаданных, созданная специально для профиля спектрометрирования.

На основе интеграции данных наземного и космического *гиперспектрального зондирования* созданы наборы тестовых спектральных образов объектов для повышения достоверности дешифрирования снимков; выполнена их каталогизация на примерах построения экологических трансект в разных регионах (рис.1).

Рис.1. Экологическая трансекта. Кольский полуостров, 2013 г.

Для работы с этими данными разработана структура базы данных, специализированный профиль метаданных для описания данных спектрометрирования, и архитектура геопортала. База данных геопортала содержит следующие элементы: точки спектрометрирования, таблицы спектрометрирования, фотографии образцов проб, космические снимки, схемы дешифрирования. Таблицы спектрометрирования и фотографии образцов привязаны к точкам по уникальным идентификаторам. Для облегчения работы с данными в режиме онлайн основные спектрометрические характеристики вынесены в отдельную таблицу и использованы при публикации картографического сервиса для геопортала. Полные данные спектрометрирования с графиками доступны по ссылке в виде файлов Microsoft Excel, предоставленных на FTP-сайте (рис.2).

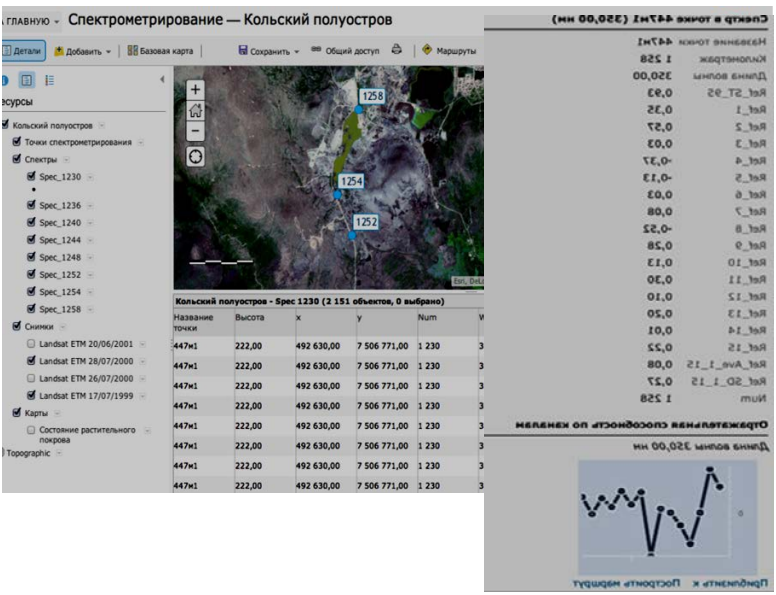

Рис. 2 Веб-сервис с доступом к таблицам спектрометрирования, снимкам и картам.

Литература

1. Зимин М,В., Тутубалина О.В., Голубева Е.И., Рис.Г.У. Методика наземного спектрометрирования растений Севера для дешифрирования космических снимков //Вестник МГУ. Серия 5: География, №3, 2014.
2. Лурье И.К., Аляутдинов А.Р., Осокин С.А. Интеграция географических информационных ресурсов и обеспечение онлайн-доступа к ним для решения научных и образовательных задач в журнале «Электронные библиотеки» *http://www.elbib.ru/index.phtml?page=elbib/rus/journal*, том 16, № 4 2013.

РОССИЙСКИЕ НАУЧНО-ОБРАЗОВАТЕЛЬНЫЕ ГЕОПОРТАЛЫ И ГЕОСЕРВИСЫ КАК ЭЛЕМЕНТЫ ИНФРАСТРУКТУРЫ ПРОСТРАНСТВЕННЫХ ДАННЫХ[1]

А.В. Кошкарев
Институт географии РАН, г. Москва
akoshkarev@yandex.ru

Выполнен анализ современного состояния работ в области создания инфраструктур пространственных данных (ИПД) в России. Даны примеры разработок геопорталов и сетевых геосервисов в организациях Российской академии наук и университетах в интересах наук о Земле.

Ключевые слова: инфраструктура пространственных данных, геопортал, сетевой сервис, науки о Земле, образование

The analysis of current statement of the Spatial Data Infrastructures (SDIs) in Russian are performed. Any examples are given to illustrate experiences, that carried out in the scientific institutes of Russian Academy of Sciences and State Universities regarding to SDI components, such a geoportals and networked spatial data services in Earth sciences.

Keywords: spatial data infrastructure, geo-portal, networked service, Earth sciences, education

10. Введение

Среди многообразия инфраструктур пространственных данных (ИПД) разного уровня и назначения (национальных, региональных, локальных, корпоративных) можно выделить ИПД научно-исследовательского типа, то есть научные, образовательные

[1]Работа выполнена при финансовой поддержке РФФИ (проект 13-05-12047 офи_м).

и научно-образовательные инфраструктуры, создаваемые вузами и академическими учреждениями или их консорциумами для обслуживания информационных потребностей, прежде всего в науках о Земле, в той их части, которая относится к сфере пространственных данных и связанных с ними сервисов (геосервисов), предполагая их возможную интеграцию с иными инфраструктурными информационными системами, например, национальными. Процесс «геопорталостроительства» в России развивается медленно и без особых успехов; это утверждение, основанное на анализе отечественного и зарубежного опыта в деле ИПД, неоднократно подтвержден его аналитическими обзорами последних лет [5, 6, 9, 12, 13]. Настоящий обзор сужает область анализа, ограничив его проблемами геоинформационного научного сообщества. Анализ основан на актуальных ресурсах Сети, публикациях последних лет и материалах прошедших совсем недавно конференций и семинаров, тематика которых имеет самое прямое отношение к затрагиваемой теме и позволяет дать представительный «срез» инициатив, успехов и проблем на пути повышения эффективности научных исследований в науках о Земле за счет ее дальнейшей информатизации путем интеграции сетевых информационных ресурсов. Из них можно назвать Международную научно-практическую конференцию «Современные технологии в деятельности особо охраняемых природных территорий» (ГИС-Нарочь 2014, Республика Беларусь, Нарочь,12-16 мая 2014 г.) [11], круглый стол «Геоинформационные системы. Инфраструктура пространственных данных» на Всероссийском форуме в области информационных и коммуникационных технологий «IT Диалог 2014» (Санкт-Петербург, 26-27 июня 2014 г.), Международную конференцию «ИнтерКарто/ИнтерГИС-20: Устойчивое развитие территорий: геоинформационное обеспечение (Белгород, Харьков (Украина), Кигали (Руанда) и Найроби (Кения), 23

июля-8 августа 2014 г.) [2] и Международную конференцию «Современные информационные технологии для фундаментальных научных исследований в области наук о земле» (Петропавловск-Камчатский, 8-13 сентября 2014 г.) [10].

11. Немного истории

Анализ был бы не полон и, главное, неконструктивен, если не рассматривать ИПД-инициативы на фоне обширного международного опыта, тем более, что кое-что давно используется в отечественных разработках. Здесь, прежде всего, стоит упомянуть североамериканские и европейские программы [5].

Первой из национальных ИПД, положившей начало перехода от эпохи ГИС к эпохе ИПД, является ИПД США NSDI, созданная в 2000 г. в соответствии с Распоряжением Президента США У.Д. Клинтона от 13 апреля 1994 г. Перспективы ее развития определены в 1998 г. в докладе NAPA (Национальная академия общественного управления) «Географическая информация в XXI веке» (http://www.napawash.org/pc_management_studies/napa_report.html) . Для прямого онлайнового доступа к пространственным данным или поиска необходимых данных в NSDI было создано шесть шлюзов с выходом в сеть национальных центров информационного обмена («клиринговых центров» — от англ. clearinghouse), объединяющую сотни серверов на территории США. На начало 2006 г. «архитектура» NSDI как «Системы Систем» представляла собой объединение Федерального комитета по географическим данным FGDC, координирующего органа NSDI, программ «Национальная карта» и «Национальный атлас», а также геопортала GOS (Geospatial One-Stop), реализующего принцип «одного окна». Геопортал был разработан в начале нулевых годов в рамках президентской программы «Электронного правительства» США как одна из 24 его инициатив. С мая 2009 г. для доступа к геоин-

формационным ресурсам NSDI используется новый сервис Data.gov (http://www.data.gov/), объединивший метаданные о более чем 400 тыс. федеральных наборов данных 172 агентств и организаций. Значительная часть данных доступна также на геопортале Geospatial Platform (http://www.geoplatform.gov). Нечто подобное – «портал открытых данных России», где можно отыскать и пространственные данные, – появился совсем недавно (http://data.gov.ru).

Еще более впечатляющий и достойный воспроизведения пример – программа создания Европейской ИПД, известная сейчас под именем программы Европейского союза INSPIRE (Infrastructure for Spatial Information in Europe), объединяющая усилия и ресурсы ее стран-участниц [3]. Как идея и инициатива Европейской комиссии она известна с 2001 г. Работы над нею начались в 2005 г., в соответствии с ее рабочей программой, в которой выделены три этапа ее разработки: предварительный этап (2005-2006 гг.), переходный этап (2007-2009 гг.) и этап реализации вплоть по 2019 г. Важнейшим событием в истории развития программы и началом разработки ее нормативной правовой и нормативно-технической базы явилось утверждение и вступление в силу 15 мая 2007 г. Директивы INSPIRE (Директива 2007/2/ЕС Европейского парламента и Совета Европы от 14 марта 2007 г.). Первая версия текста Директивы INSPIRE была опубликована в 2004 г.; ее содержание достаточно близко к утвержденной и действующей ныне. Она нацелена на решение проблем, связанных с теми данными, которые необходимы для мониторинга окружающей среды и улучшения ее состояния, включая воздушную и водную среды, почвы и природные ландшафты. Тем не менее, она не предназначена для выработки какой-либо новой программы сбора пространственных данных в странах-членах ЕС, лишь предлагая разработку эффективных средств ис-

пользования имеющихся данных и требуя документирования пространственных данных в виде метаданных, предоставления сервисов доступа к ним, обеспечения их интероперабельности (взаимосовместимости) и устранения препятствий в их использовании [4].

Россия не может похвастаться какими-либо успехами в деле строительстве ИПД. Оно началось с подготовки еще десятилетие назад Концепции ИПД РФ, одобренной распоряжением Правительства Российской Федерации от 21 августа 2006 г. # 1157-р (http://base.consultant.ru/cons/cgi/online.cgi?req=doc;base=EXP;n=372580). К числу реальных инициатив, которые можно записать в «актив» ИПД РФ, нужно отнести лишь ее геопортал, введенный в эксплуатацию в марте 2012 г. и официально открытый с июля 2012 г. (http://nsdi.ru). Правда, по состоянию на сентябрь 2014 г. он находится в режиме тестирования. Заслуживают внимания региональные инициативы по разработке законов и постановлений о создании ИПД субъектов РФ и устройстве их геопорталов. Первым из них был геопортал электронного правительства Самарской области (http://geosamara.ru). На сегодняшний день наиболее полнофункциональным можно считать геопортал Воронежской области (http://map.govvrn.ru). Наличие функций поиска данных по метаданным позволяет отнести к числу «истинных» и «Отраслевой узел Единого геоинформационного пространства города Москвы» (http://egip.mka.mos.ru/egip/egip.nsf/). Из недавних примеров нужно упомянуть геоинформационный портал Чувашии (http://sdi.cap.ru/geoportal/catalog/main/home.page) в составе ИПД Чувашской Республики и геопортал Республики Коми (http://gis.rkomi.ru/). К сожалению, часть региональных информационных систем ограничивает или не обеспечивает доступ к своим информационным ресурсам широкому кругу пользователей по соображениям секретности своих, в основном картографических,

данных. Около восьми лет назад был утвержден российский национальный профиль стандарта ИСО на содержание пространственных данных ГОСТ Р 52573-2006 «Географическая информация. Метаданные», однако, до сих пор большинство разработчиков геопорталов используют его оригинал, а именно ISO 19115:2003 «Geographic information – Metadata» (ИСО 19115:2003 «Географическая информация – Метаданные»). Не определен перечень базовых пространственных данных ИПД РФ, и само это понятие обсуждается уже около десяти лет. Это практически все, что можно записать в «актив» ИПД РФ национального и регионального уровня.

12. Научно-образовательные геоинформационные ресурсы

За пределами вертикальной линейки ИПД РФ, в том числе в научно-образовательной сфере, усилия и успехи тоже фрагментарны. Еще в 2010 г. на Всероссийском семинаре «Современные информационные технологии для фундаментальных научных исследований РАН в области наук о Земле» в г. Владивостоке была выдвинута идея создания Академической ИПД, оформленная в виде стратегии ее создания и так и не получившая до сих пор своего одобрения в академических верхах [7].Тем не менее, ее отдельные компоненты: геопорталы с доступом к тем или иным сервисам и их прототипы – создаются.

Их география достаточно широка.

Из дальневосточных примеров нужно назвать, прежде всего, геопортал Института вулканологии и сейсмологии ДВО РАН (http://geoportal.kscnet.ru/), предоставляющий доступ к информационным ресурсам (данным, каталогам, архивам, метаданным, репозитарию публикаций) по тематике деятельности института (рис. 1). В качестве технологической платформы используется

популярный продукт с открытым исходным программным кодом GeoNetwork. Важно, что геопортал поддерживает поиск данных по их метаданным – основой критерий, по которому судят о том, является ли геопортал «истинным», или представляет собой лишь средство картографической визуализации данных, а то и вовсе сайт или портал с громким названием «геопортал», «ГИС-портал» или «геоинформационный портал».

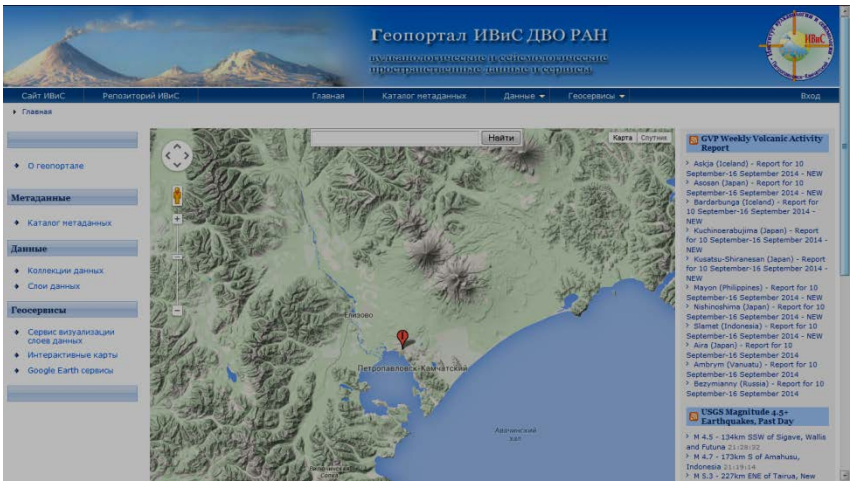

Рис. 1. Главная страница геопортала ИВиС ДВО РАН.

Давно известен онлайновый «Электронный атлас Северо-Востока России», проект лаборатории ГИС-технологий Северо-Восточного комплексного НИИ ДВО РАН, ранее находившийся по магаданскому адресу (http://atlas.magis.ru), а ныне переместившийся на сайт Дальневосточного геологического института ДВО РАН (http://ags.febras.net/atlas_ne_ngp/) во Владивостоке.

Прототип геопортала с каталогом метаданных, возможностями картографической визуализации и доступом к информационным ресурсам можно найти в другой научной организации Владивостока – Тихоокеанском институте географии ДВО РАН (рис. 2).

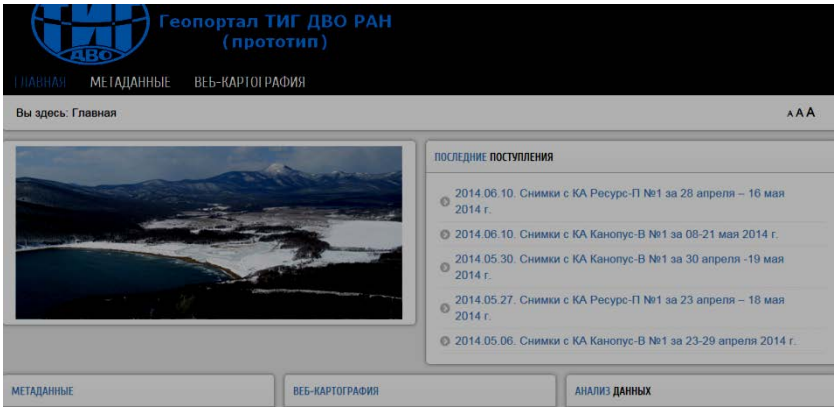

Рис. 2. Главная страница геопортала ТИГ ДВО РАН: http://gis.dvo.ru/.

Из работ Сибирского отделения РАН нужно отметить геопортал Института вычислительной математики СО РАН (рис. 3). Геопортал служит программно-технологической основой ресурсоемких информационно-аналитических систем регионального уровня для задач различной тематики – информационной поддержки отраслевого управления (в сфере здравоохранения и образования), экологического мониторинга и оценки состояния окружающей природной среды, прогноза социально-экономического развития региона, централизованного информационного обеспечения картографическими данными и т.д.

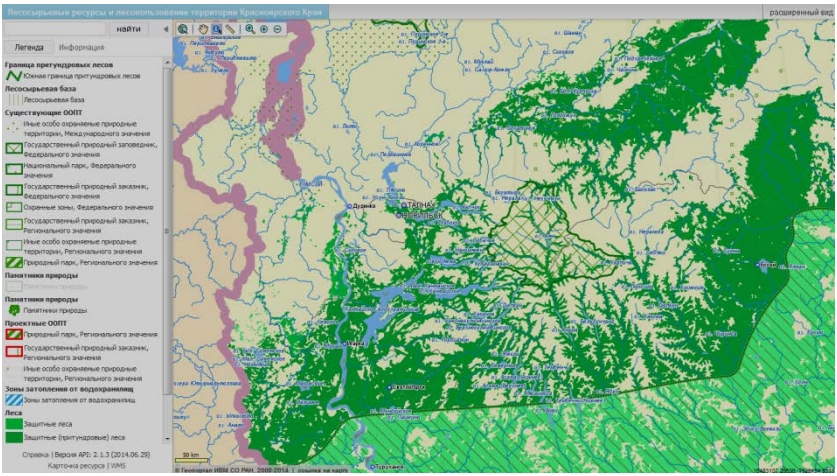

Рис. 3. Картографическая визуализация на геопортале ИВМ СО РАН (http://gis.krasn.ru/): лесосырьевые ресурсы и лесопользование территории Красноярского края.

Это лишь немногие примеры, отражающие современное состояние работ в области создания геопорталов в рамках инфраструктур научных информационных ресурсов и систем. Заслуживает внимания также геопорталы Института географии РАН (http://asdi.igras.ru), Вычислительного центра им. А.А. Дородницына РАН (http://www.geometa.ru), Геофизического центра РАН (http://gis.gcras.ru/geoportal/catalog/main/home.page) и некоторых другие, достойные внимания с точки зрения усвоения и адаптации их опыта в деле «геопорталостроительства».

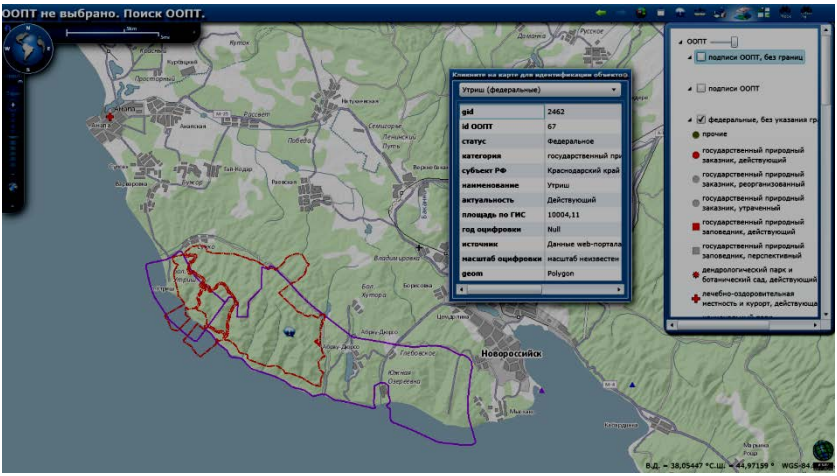

Рис. 4. http://oopt.aari.ru/oopt_map.

Еще один пример иллюстрирует тезис о том, что наука была и остается главным источником тематических данных или, по крайней мере, она вполне способна систематизировать и организовать их в операбельные наборы. Это сервис Арктического и антарктического научно-исследовательского института (г. Санкт-Петербург) в системе Роскомгидромета РФ об особо охраняемых природных территориях (ООПТ) России (рис. 4).

По существу он альтернативен кадастру ООПТ, который ведется Минприродой РФ и гораздо более информативен по сравнению с ним, поскольку содержит не только координатные описания границ ООПТ федерально и регионального и, отчасти, местного уровня, отсутствующие в кадастре, но и огромную справочную информацию по краснокнижным видам и нормативной правой базе. И таких примеров, когда академическая или вузов-

ская наука успешно решает ведомственные задачи, множество. Это и цифровые тематические данные, и производные от них тематические картографические произведения, до которых ведомствам, которые отвечают за свои кадастры, реестры и иные ведомственные (в том числе государственные) ресурсы, зачастую нет никакого дела.

Несколько примеров из сферы образования.

Здесь нужно сказать, прежде всего, о консорциуме «Унигео», созданном в 2011 г. по инициативе ИЦТ СканЭкс и объединяющем 26 университетов (данные на 2013 г.), от Белгорода до Хабаровска с Центрами приема и обработки космической информации и геопорталами, включая один из первых – геопортал Московского государственного университета им. М.В. Ломоносова (http://www.geogr.msu.ru:8082/api/index.html). Здесь же в прошлые годы была создана единственная в своем роде локальная ИПД базы полевых студенческих практик «Сатино» и соответствующий геопортал, ныне, по-видимому, доступный только в университетской локальной сети [8].

Институтами Иркутского научного центра СО РАН разрабатывается ИПД Иркутского регионального научно-образовательного комплекса [1], активно развивается геопортал Тверского государственного университета (http://geoportal.tversu.ru/).

13. Заключение

Основные выводы, которые можно и нужно сделать относительно реализации отдельных компонентов отечественных научно-образовательных ИПД.

Нормативная правовая база. Все работы ведутся в правовом вакууме, поскольку концепция ИПД РФ устарела, а новое федеральное законодательство, то есть поправки в Федеральный закон «О геодезии и картографии», проект которого на сегодня извес-

тен в четырех версиях. Термина «ИПД» мы не найдем ни в одном федеральном законе. Исключение составляет региональное законодательство.

Стандарты. В России нет системы национальных стандартов, гарантирующих интероперабельность пространственных данных и связанных с ними сервисов в сетевой среде. Многие годы бездействует Технический комитет по стандартизации ТК 394 «Географическая информация/геоматика» Ростехрегулирования РФ, который является российским аналогом Технического комитета ИСО ТК 211 «Географическая информация/геоматика» (ISO/TC 211 «Geographic information/Geomatics), источника международных стандартов в области геоинформатики и ИПД серии ИСО 19100. Выход видится в использовании оригинальных зарубежных стандартов, что допускается Федеральным законом «О техническом регулировании».

Данные. Наука – основной производитель уникальных тематических пространственных данных, собираемых, хранимых и обрабатываемых сотнями учреждений вычислительно-математического, информационно-технологического, географического и геоэкологического профиля, где использование геоинформационных технологий давно стало обычным делом, а «цифра» если не вытеснила, то серьезно потеснила «бумагу». Сейчас они доступны и в Сети, и в облаках, но оказывается, что по большому счету они никому не нужны, недовостребованы ни наукой, ни практикой. Причин такой ситуации много, но одна из них очевидна – это отсутствие информации о них, то есть метаданных. И это при том, что затраты на их подготовку занимают ничтожную долю от затрат на сбор и организацию самих данных.

Геопорталы. К немногим из достижений научно-образовательного сообщества в сфере ИПД можно отнести геопорталы. Повышению их роли будет способствовать устранение

ряда явных недостатков, среди которых нужно упомянуть необновляемость контента, короткий жизненный цикл (эфемерность), ограничения доступа (обязательность регистрации пользователя, полная закрытость) и ограниченность поддерживаемых сервисов (по числу и качеству). До сих пор у нас нет геопортала, обеспечивающего доступ к тому минимальному набору сервисов, который имеет каждый европейский национальный геопортал в рамках реализации программы INSPIRE.

Литература

1. Бычков И.В., Ружников Г.М., Хмельнов А.Е., Гаченко А.С., Федоров Р.К. Инфраструктура пространственных данных Иркутского регионального научно-образовательного комплекса // Интернет и современное общество. Сборник научных статей. Материалы XIV всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 12-14 октября 2011 г. СПб., 2011. – С. 33-35.
2. ИнтерКарто/ИнтерГИС-20: Устойчивое развитие территорий: геоинформационное обеспечение. Материалы Международной конференции, Белгород, Харьков (Украина), Кигали (Руанда) и Найроби (Кения), 23 июля – 8 августа 2014 г. – 636 с.
3. Кошкарев А.В. Директива INSPIRE и национальные инициативы по ее реализации // Пространственные данные, 2009, №2. – С. 6–11. Копия: http://www.gisa.ru/54638.html.
4. Кошкарев А.В. От первых инициатив по созданию инфраструктур пространственных данных – к Директиве INSPIRE // Вестник геодезии и картографии, январь 2014 г. – С. 2.
5. Кошкарев А.В. Проблемы становления российских ИПД // ИнтерКарто/ИнтерГИС-20: Устойчивое развитие территорий: геоинформационное обеспечение. Материалы Международной конференции, Белгород, Харьков (Украина), Кигали (Руанда) и Найроби (Кения), 23 июля – 8 августа 2014 г. – С. 137-151.

6. Кошкарев А.В., Ротанова И.Н. Проблемы российских региональных ИПД // Геоинформационное картографирование в регионах России: материалы V Всероссийской научно-практической конференции (Воронеж, 19-22 сентября 2013 г.) / Воронежский государственный университет. – Воронеж: Изд-во «Цифровая полиграфия», 2013. – С. 77-90.
7. Кошкарев А.В., Ряховский В.М., Серебряков В.А. Инфраструктура распределенной среды хранения, поиска и преобразования пространственных данных // Открытое образование, 2010, № 5. – С. 61-73.
8. Лурье И.К., Аляутдинов А.Р. Осокин С.А. Интеграция географических информационных ресурсов и обеспечение онлайн-доступа к ним для решения научных и образовательных задач: http://www.elbib.ru/content/journal/2013/201304/LAO/LAO.ru.html.
9. Ротанова И.Н., Кошкарев А.В., Медведев А.А. Использование материалов дистанционного зондирования Земли для цифрового моделирования рельефа в составе региональных инфраструктур пространственных данных // Вычислительные технологии. – 2014. – Т. 19. – № 3. – С. 38–47.
10. Современные информационные технологии для фундаментальных научных исследований в области наук о земле: Материалы. Международной конференции, Петропавловск-Камчатский, 8-13 сентября 2014 г. – Владивосток: Дальнаука, 2014. – 178 с.
11. Современные технологии в деятельности ООПТ. ГИС-Нарочь, 2014: Материалы международной научно-практической конференции (тезисы): https://www.dropbox.com/s/m6rdaao2p58a6gz/sbornik_tezisov. pdf.
12. Koshkarev A.V., Rotanova I.N. Position and role of Russian research and education community in formation and development of spatial data infrastructure // Современные информационные технологии для фундаментальных научных исследований в области наук о земле: Материалы. Международной конференции, Петропавловск-Камчатский, 8-13 сентября 2014 г. – Владивосток: Дальнаука, 2014. – С. 126.

13. *Koshkarev A.V., Rotanova I.N. Projects on implementation* of spatial data infrastructure in the Russian Federation: a review based on available sources // Конференција Математичке иинформационе технологије (2013 ; ВрњачкаБања, Бечићи) Zbornik radova Konferencije MIT [Matematičke i informacione tehnologije] 2013 : [[održane] u Vrnjačkoj Banji od 5. do 9. septembra i u Bečičima od 10. do 14. septrembra 2013. godine] / [urednik Dragan Aćimović]. – Kosovska Mitrovica : Prirodno-matematički fakultet ; Novosibirsk ; Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences, 2014 (Kraljevo : Ofsetpres). – P. 348–358.

СЕРВИСЫ ВВОДА И РЕДАКТИРОВАНИЯ РЕЛЯЦИОННЫХ ДАННЫХ НА ОСНОВЕ БАЗОВЫХ ПРОСТРАНСТВЕННЫХ ДАННЫХ

Р.К. Фёдоров, А.С. Шумилов, Е.Н. Фёдорова

Институт динамики систем и теории управления СО РАН, Иркутск, Россия

fedorov@icc.ru

В рамках геопортала ИДСТУ СО РАН разработаны сервисы ввода и редактирования реляционных данных на основе базовых пространственных данных, позволяющие многопользовательскую работу через Интернет.

Ключевые слова: Базовые пространственные данные, OGC, WMS, WPS, SLD, Mapserver.

Within the geoportal ISDCT SB RAS editing services for relational data based on the basic spatial data has been developed. They allow multi-user work via the Internet.

1. Введение

В настоящее время активно развиваются научные исследования состояния и динамики природных экосистем и их компонентов, носящие междисциплинарный характер. Одной из начальных задач любых исследований является сбор данных. Обмен, совместный ввод, редактирование и анализ данных различными коллективами исследователей являются затруднительными по ряду причин. В частности исследователи вводят данные в различных программных системах, в том числе в геоинформационных, используются различные классификаторы и атрибутный состав. Объединение данных не является тривиальным. В некоторых случаях требуется организация регулярного обмена данными.

Развитие Интернет технологий, создание хранилищ данных и центров обработки данных, наличие базовых пространственных данных позволяют разрабатывать полнофункциональные информационные системы, работающие через Интернет. Создание Интернет сервисов ввода и редактирования данных позволяет обеспечить более эффективное взаимодействие между исследователями на уровне данных.

2. Сервисы ввода и редактирования реляционных данных

В рамках геопортала ИДСТУ СО РАН разработаны сервисы ввода и редактирования реляционных данных, содержащих пространственные атрибуты. Достоинствами разработанных сервисов являются:

- многопользовательская работа через Интернет, одновременно несколько пользователей могут вводить, редактировать и просматривать данные;
- пользователь может самостоятельно создавать таблицы и определять атрибутный состав таблиц;
- сервисы автоматически осуществляют ввод и отображение на карте пространственных данных;
- возможно применение различных фильтров, в том числе пространственных;
- осуществляется разграничение прав доступа.

Основой работы сервисов являются структурные спецификации таблиц в формате JSON. На основе структурной спецификации создаются таблицы БД, генерируется пользовательский интерфейс и определяется логика работы сервисов. Структурная спецификация таблицы содержит: название таблицы и набор атрибутов. Каждый атрибут в свою очередь характеризуется: названием, именем в базе данных, типом данных, единицами изме-

рения (для числовых данных), элементом управления и его свойствами. Элемент управления необходим для формирования пользовательского интерфейса добавления, редактирования и отображения данных. Свойства элемента управления позволяют настраивать пользовательский интерфейс в зависимости от характеристик данных, например, единицы измерения для числовых данных или определять тип пространственных данных. В рамках геопортала разработан каталог описания таблиц, хранящий метаданные и структурные спецификации. Структурные спецификации таблиц в рамках каталога упорядочиваются в виде иерархий и применяют механизмы наследования и полиморфизма в терминах объектно-ориентированного подхода.

Каждому пользователю геопортала предоставляется схема в СУБД PostgreSQL [1], в которой он может создавать таблицы с помощью специальной Web-формы (см. рис.1). При создании таблицы может определить произвольное количество атрибутов.

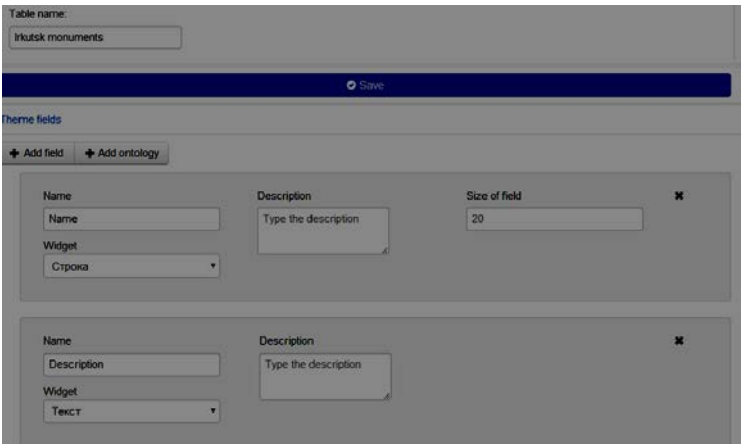

Рис. 1. Интерфейс создания таблицы.

Рассмотрим подробнее элементы управления, которые определяются для каждого атрибута таблицы. Каждый элемент управления реализует методы для ввода, отображения, фильтрации данных. Набор элементов управления является расширяемым. Достаточно унаследовать базовый класс widget (рис.2). Метод initwidget() предназначен для генерации пользовательского интерфейса для ввода и редактирования данных атрибута. Метод getUserVal() предназначен для формирования строки со значением для отображения пользователю. Например, если это атрибут является ссылкой на другую таблицу (reference key), то пользователю отображается строковое значение, полученной со ссылаемой таблицы. Методы ViewpropForm() и Getprop() предназначены для формирования специфичных свойств, необходимых для реализации специфических операций. Например, в свойствах атрибута структурной спецификации может содержаться список возможных значений атрибута или имя таблицы справочника.

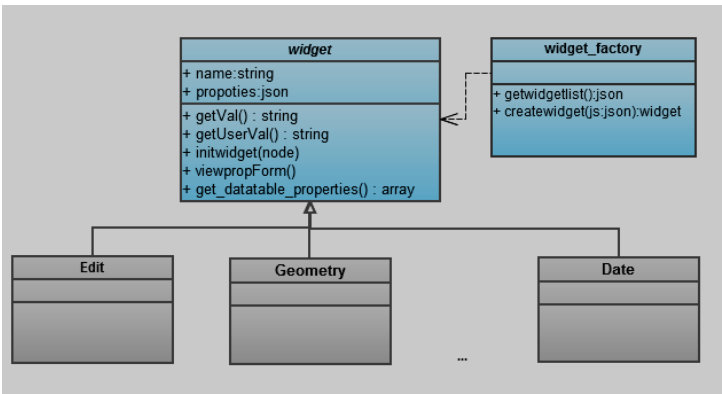

Рис. 2. Диаграмма классов компонента widgets.

Каждому элементу управления установлен в соответствие тип данных СУБД PostgreSQL. При сохранении структурной спецификации в СУБД PostgreSQL создается таблица.

Основываясь на структурных спецификациях, формируется интерфейс ввода и редактирования таблицы. Пространственные атрибуты отображаются на карте. Ввод и редактирование данных осуществляется в таблице или на форме. Для каждого атрибута используется элемент управления, указанный пользователем. По всем атрибутам можно выполнять сортировку и фильтрацию.

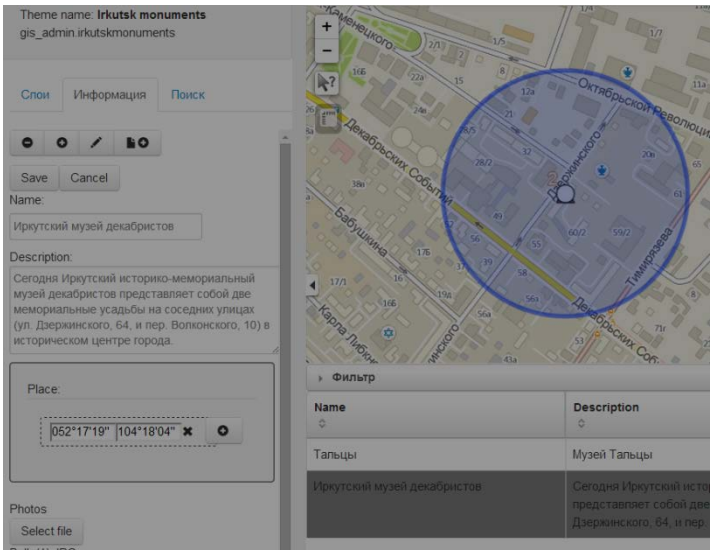

Рис. 3. Редактирование таблицы.

Применение элементов управления позволяет без программирования создавать гибкий и удобный пользовательский интерфейс для работы с реляционными таблицами с любой линейной

структурой атрибутов. Взаимодействие с картой, таблицами справочниками реализовано с помощью элементов управления. Разработано более двадцати различных элементов управления, позволяющих работать со стандартными типами данными: number, string, date, Boolean и т.д.

Рассмотрим далее наиболее интересные элементы управления.

Элементы управления point, line, polygon. Эти элементы управления предназначены для работы с пространственными данными. Позволяют вводить координаты точечных, линейных и площадных объектов. При наличии у таблицы атрибутов, использующих перечисленные элементы управления, создается карта и добавляется для каждого атрибута специальный слой. Ввод и редактирование данных можно осуществлять в двух режимах: последовательно указывая координаты всех точек пространственного примитива, либо с помощью мыши.

Элемент управления political_division. Большинство существующих тематических данных связано с базовыми пространственными данными. В рамках геопортала формирование тематических данных на основе пространственных данных, административное деление которые являются частью базовых пространственных данных (БПД), производится с помощью специально разработанного элемента управления «political_division». Данный элемент управления позволяет выбирать объекты административного деления и отображать границы выбранных административных объектов. Для базовых пространственных данных в PostgreSQL созданы таблицы, содержащие семантику и геометрию объектов. БПД образуют иерархии объектов (федеральные округа, регионы, районы и т.д.). Информация об иерархической подчинённости задается с помощью атрибутов таблиц. При настройке элемента управления необходимо указать уровень иерархии административного деления (см. Рис. 4). Соответственно в таблице выбор объектов административного деления произво-

дится на указанном уровне. Далее пользователь должен указать конкретные объекты выше уровнем иерархии административного деления. Например, если в таблицы будут указываться данные по районам, то нужно указать тип район и что будут выбираться районы Иркутской области Сибирского федерального округа.

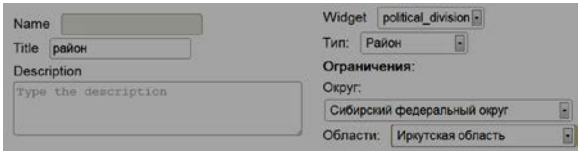

Рис. 4. Настройка элемента управления political_division.

На форме пользователь может выбрать объект по его названию, используя контекстный поиск (autocomplete) (см. рис. 5).

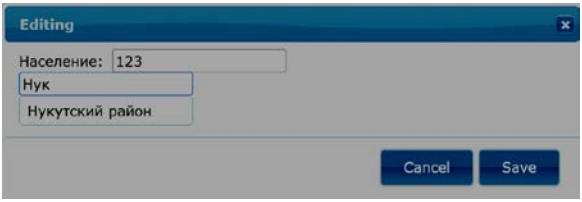

Рис. 5. Ввод данных с помощью элемента управления «political_division».

Элемент управления classify. Для работы с произвольными таблицами без иерархической зависимости применяется элемент управления «classify». Данный элемент управления позволяет использовать в качестве таблицы справочника произвольную таблицу, зарегистрированную в каталоге. Если в таблице справочника имеется атрибут с пространственными данными, то создается соответствующий слой на карте.

Элемент управления image. Данный элемент управления позволяет в качестве значения атрибута использовать набор изображений. При редактировании пользователь может загрузить изображения в систему хранения геопортала, указать к ним различные комментария. Отображение изображений производится в виде слайдов.

Отображение данных таблиц осуществляется с помощью Mapserver [2] на серверной стороне (применяется стандарт WMS [3]), на клиентской с помощью библиотеки leaflet [4]. Для Mapserver создается специальный файл настроек (MAP), в котором формируется запрос на получение данных из таблиц БПД и пользовательской таблицы. Учитываются все указанные пользователем фильтры. Вид отображения данных задается с помощью SLD [5]. Разработан Web-редактор, позволяющий создавать SLD файлы для пользовательских таблиц.

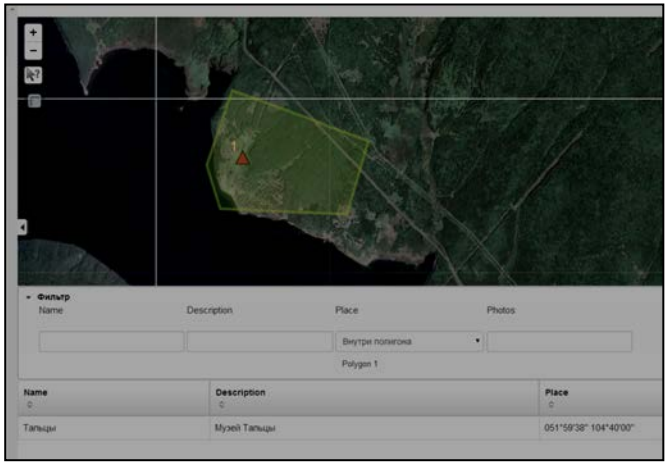

Рис. 6. Отображение данных на карте и применение пространственного фильтра.

Созданные таблицы с помощью сервисов ввода и редактирования можно применять в качестве входных данных в WPS-сервисах [6]. При передаче таблицы формируется строка соединения драйвера GDAL для непосредственной работы WPS-сервиса с базой данных. В текущей версии таблицы PostgreSQL доступны только для локальных WPS-сервисов, но в дальнейшем будет реализация для регламентированного доступа к базе данных извне.

3. Заключение

Разработанные сервисы обладают следующими достоинствами:

– открытость научных пространственных и тематических информационных ресурсов для проведения междисциплинарных исследований учреждениями науки и образования;

– применение единой системы базовых пространственных данных (БПД) для географической привязки баз данных и знаний по экосистемам Байкальской Сибири;

– внедрение международных стандартов на географическую информацию и на представление и обмен пространственными данными;

– удобство работы, предоставление удобного, интуитивно понятного пользовательского интерфейса; возможность быстрой публикации в Интернет как данных, так и метаданных;

– обеспечение надежного и регламентированного хранения данных на сервере и предоставление функций работы с файловой системой сервера;

– возможность одновременной работы нескольких пользователей с одним набором пространственных данных (добавление новых записей), что обеспечит интеграцию работы разных исследователей;

– возможность использования разного рода карт, что поможет анализу зависимости распространения видов (групп видов) от различных факторов – почвы, ландшафты, антропогенная нагрузка и т.д.;

– использование единых международных классификаторов.

Разработанная система используется сбора данных для инвентаризации и анализа фиторазнообразия Байкальской природной территории.

Литература

1. PostgreSQL [Электронный ресурс] // The PostgreSQL Global Development Group [сайт]. URL: http://www.postgresql.org/ (дата обращения: 04.09.2014).
2. Mapserver [Электронный ресурс] // Mapserver Consortium [сайт]. URL: http://mapserver.org / (дата обращения: 04.09.2014).
3. Geospatial and location standards [Электронный ресурс] // Open Geospatial Consortium [сайт]. URL: http://www.opengeospatial.org/ (дата обращения: 04.09.2014).
4. LeafLet [Электронный ресурс] // An Open-Source JavaScript Library for Mobile-Friendly Interactive Maps [сайт]. URL: http://leafletjs.com/ (дата обращения: 04.09.2014).
5. The OGC Announces Styled Layer Descriptor & Symbol Encoding Specifications [Электронный ресурс] // Open Geospatial Consortium [сайт]. URL: http://www.opengeospatial.org/pressroom/pressreleases/761 (дата обращения: 04.09.2014).
6. OpenGIS Web Processing Service (WPS) Implementation Specification, v1.0.0. Release date: June 08, 2007. [Электронный ресурс] // Open Geospatial Consortium [сайт]. URL: http://www.opengeospatial.org/standards/wps (дата обращения: 04.09.2014).

СОДЕРЖАНИЕ I ТОМА