Применение онтологии для ведения и доступа к данным коллекции «Природные ресурсы региона»

© В.А. Лебедев, В.Г. Старкова, С.В Брагин

Институт прикладных математических исследований
Карельского научного центра РАН
math@krc.karelia.ru

Аннотация

Описываются подход к концептуализации и структура онтологии «Природные ресурсы региона», реализация ее компьютерного представления и технология применения онтологии для регистрации данных, поиска релевантных, а также технология редактирования самой онтологии.

Введение

Контроль за состоянием и использованием природных ресурсов региона является одной из важных функций регионального и муниципального руководства. В институте прикладных математических исследований разрабатывается интернет-технология контроля за природными ресурсами региона. В настоящее время разработана исследовательская версия [1].

Поскольку природные ресурсы рассредоточены в пространстве региона, в основу разрабатываемой положена геоинформационная (ГИС) технология [2]. Описание природных ресурсов будет отображаться в некотором наборе тематических баз данных, связанных процедурой геокодирования с цифровой векторной картой. На каждый вид ресурса создается отдельная база данных, в которой характеристики ресурса представляет отдельная запись, содержащая различные виды характеристик ресурса: числовые, текстовые и, возможно, графические. Данные об использовании ресурсов должны представляться в виде статистических отчетов (отдельный отчет на каждый ресурс) и аналитических записок. Помимо этого изменение состояния некоторых ресурсов, например, лесов, может контролироваться при помощи данных дистанционного зондирования (космических снимков).

К природным относятся: земельные ресурсы, месторождения полезных ископаемых, лесные, болотные и водные системы, животное и растительное разнообразие, рекреационно-туристические объекты, климатические условия. Контроль за использованием ресурсов осуществляется через мониторинг предприятий и организаций, осуществляющих охрану, использование, добычу, заготовку или восстановление соответствующих ресурсов, региональным статуправлением. Помимо этого для обеспечения инвестиционных интересов должны быть отображены характеристики объектов инфраструктуры: населенные пункты, дороги, водные пути, объекты энергообеспечения и порты с перевалочными пунктами и др.

Весь спектр сведений обо всех ресурсах разных типов и их использовании представляет собой коллекцию данных разнообразной структуры и содержания, ориентация в которой и поиск релевантных является не простой задачей.

Поскольку описания характеристик ресурсов и объектов инфраструктуры и их использования по определению привязаны к географической карте, то обеспечен доступ к ним через карту, благодаря применению технологии ГИС, но возможности поиска объектов, релевантных интересам инвесторов и контролирующих органов, через карту ограничены. Возможен поиск объектов, привязанных к карте. Однако часто необходим поиск данных по связям между объектами и некоторому набору характеристик. В этом случае эффективным представляется поиск с использованием онтологии коллекции, которая должна отражать различные характеристики ресурсов, предприятий и инфраструктуры и отношений между такими объектами. Помимо поиска онтология коллекции может быть использована также при оформлении и загрузке данных (создании и ведении коллекции).

Основание разработки

Онтология в информатике, согласно современным толкованиям, является «точной спецификацией концептуализации предметной

Труды 7ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2005, Ярославль, Россия, 2005.

области» (T. Grubep), но с определенными ограничениями в зависимости от области интересов (M. Ushold) и должна включать словарь терминов и некоторые спецификации их значений [3,4].

Одним из способов представления спецификаций является экстенсиональный, что в условиях нечетких толкований понятий обеспечивает прямую обозримость концепции спецификатора о составе и границах понятийного аппарата предметной области и предоставляет возможность для конструктивной критики и развития. Обычно при выполнении концептуализации используют типы отношений классификации, агрегации и ассоциации [4,5]. Тогда получаемая онтология может быть представлена в виде связного ациклического графа, в котором вершинами являются термины, а дугами – отношения между ними указанных типов. В таком графе часто явно выделяются уровни иерархии отношений, что позволяет специфицировать их свойства с частичным наследованием от корней к листьям.

Как известно, граф является парой <*V*, *E*>, где *V* – множество вершин, *E* – множество дуг. Упорядоченная пара $(a,b) \in E$ называется дугой, где $a,b \in V$, весь граф называется ориентированным. Связный ациклический граф может быть представлен множеством дуг, так как в нем отсутствуют изолированные вершины и, следовательно, все вершины указаны в множестве дуг. Последнее может быть интерпретировано как реляционная база данных, в которой ключевыми колонками являются пары вершин (дуги), дополнительно могут быть указаны пометы вершин и дуг (например, названия уровней иерархии и видов отношений).

В соответствии с указанным подходом в работе [6] было описано представление онтологии «Водные ресурсы региона», ее использование для поиска релевантных данных в коллекции и представлена технология формирования поискового образа с использованием онтологии и последующего поиска релевантных данных. При этом запрос строится по схеме:

A&B&…&C& (F|L|K…)&(M|…|N)…,

где буквами обозначены термины, & – конъюнкция, | – дизъюнкция. Поскольку под термином может пониматься как одно слово, так и словосочетание, выбранная схема поискового образа не содержит только средств указания расстояний между терминами. В то же время допускаются и простейшие запросы в один термин. Преимущество использования онтологии для формирования запроса в том, что из нее извлекаются стандартизованные лексические представления терминов и заносятся в запрос без ошибок. Это же преимущество можно использовать и при оформлении ключевых слов текстов, названий колонок баз данных, снимков и карт в процессе их загрузки в коллекцию, что в работе [6] не было реализовано.

К недостаткам этой работы следует отнести также невозможность формирования запросов и оформления информационных ресурсов, содержащих сведения о связанных (ассоциированных) объектах, отнесенных к разным классам. В представленной ниже реализации учтены эти замечания.

Структура онтологии

Структура онтологии принята такой же, как в работе [6]. Представляется она в виде реляционной базы данных, содержащей таблицу помеченных дуг графа с указанием уровня иерархии и таблицу синонимов.

Первый уровень включает следующие типы объектов: сводные данные по региону, земли, водные объекты, биоресурсы, месторождения полезных ископаемых, населенные пункты, предприятия и организации, транспортные, энергообеспечивающие объекты и объекты связи. Следующий уровень представляет разделение типов ресурсов на классы объектов в соответствии с основными научными дисциплинами. Например, земли подразделяются на категории: лесные, сельскохозяйственные, поселений и предприятий, коммуникаций, мест захоронений и др.; биоресурсы – на древесные, недревесные (ягоды, грибы, лекарственные растения), рыбные, охотничьи и т.д. На третьем уровне представлены темы описания ресурсов; например, темы месторождений минеральных ресурсов – черные, цветные и драгоценные металлы, химическое сырье, строительные, поделочные и драгоценные неметаллы. На четвертом уровне отображаются характеристики ресурсов, например, месторождения характеризуются местоположением, видом ресурса, составом пород, процентом содержания, качеством, запасом, площадью, глубиной залегания, мощностью пласта, степенью разведанности и т.д. Далее располагаются значения некоторых характеристик, выражаемых словесно и названия объектов (листья). Отметим, что количество уровней иерархии может быть расширено.

Следует отметить также принципы построения таблицы синонимов. В каждом гнезде синонимов выделяется синоним-доминант, который включается в состав онтологии. В таблице синонимов доминанты помещаются в левую колонку, а в правую прочие синонимы. В качестве доминанта выбирается наиболее употребительная или обобщенная форма термина. Например, в гнезде: «углекислый газ, диоксид углерода, углекислота, CO_2» в качестве доминанта выбирается первый термин. Примером обобщенной формы является лексема «название», которая включается в термины «названия ресурсов». В этом случае заменой термина «ресурс» конкретными названиями ресурсов получаем гнездо неполных

синонимов, которое будет полезно как при загрузке ресурсов, так и при построении запроса.

Построенная часть онтологии включает 120 тем, более 600 характеристик, более 450 значений 47 характеристик и около 150 синонимических гнезд.

Поиск данных

Для использования онтологии в составе коллекции данных о природных ресурсах и их использовании разработано три сервиса:

1) построение запроса и поиск данных о ресурсах, 2) оформление метаданных о ресурсах, 3) редактирование онтологии.

Онтология представляет собой реляционную базу данных, приложения для работы с которой оформлены как скрипты JavaSkript или C++ и взаимодействуют с ней по CGI-сценариям.

Сервис построения запросов и поиска релевантных данных вызывается из формы доступа к данным. Он включает набор динамических форм для прохождения всех уровней иерархии онтологии от корня к листьям (поиск в глубину). При этом на каждом уровне производится выбор некоторого подмножества терминов, по которому формируется оператор Select и производится отбор терминов следующего уровня, связанных с этим подмножеством (рис. 1). Имитация ассоциативных связей обеспечивается возможностью выбора нескольких альтернатив на каждом уровне. После прохождения всех уровней иерархии онтологии полученный набор терминов включается в оператор Select, который производит отбор синонимов. Набор терминов вместе с синонимами образует исходный набор терминов для формирования запроса и высвечивается на экране в виде списка. Пользователь последовательно переносит желаемые термины из левой части таблицы в правую. При этом термины в правой части заключаются в скобки (которые изображаются пустыми строками), что означает соединение в конъюнкцию. При необходимости обеспечить выбор хотя бы одного термина из списка (например, синонимов) термины этого списка заключаются в общие скобки, как показано на рис. 1. Это означает, что термины в общих скобках соединены дизъюнкцией. Таким образом, обеспечивается построение запроса по схеме, показанной в разделе «Основание разработки».

После сформирования запроса запускается скрипт поиска релевантных данных, который осуществляет сравнение терминов запроса с записями базы данных наборов ключевых слов информационных ресурсов. Эта база данных представляет собой таблицу, в первой колонке которой указаны имена информационных ресурсов и далее – списки ключевых слов, характеризующих содержание информационных ресурсов. Эта база данных формируется при загрузке и регистрации данных (карт, слоев, баз данных, текстов, снимков). (Во время регистрации также используется онтология, как показано ниже). По результатам поиска формируется список релевантных информационных ресурсов: текстов, баз данных, карт и снимков.

Отметим, что после очередного поиска в базе данных ключевых слов (успешного или неуспешного) возможен возврат в исходный набор терминов, сформированный в результате прохождения онтологии, и на его основе формирование новых запросов.

Регистрация данных

Данные о природных ресурсах региона готовятся в различных организациях, поэтому их загрузка и регистрация (а также сопровождение коллекции) осуществляются по сети с помощью разработанной интернет-технологии [7].

Одной из важных задач подготовки информационных ресурсов является лингвистическое обеспечение, под которым понимается унификация употребления и написания терминов и названий предметной области применительно к подготавливаемым данным. Онтология, которая составляется специалистами предметниками, является унифицированным и стандартизованным словарем, поэтому его целесообразно использовать при регистрации русских названий колонок таблиц баз данных и списков ключевых слов текстов, перенося соответствующие термины и названия из онтологии в регистрационные таблицы. Для выполнения этой функции разработан специальный сервис, дополняющий технологию администрирования коллекций, изложенную в работе [7].

На рис. 2 показана форма для регистрации названий колонок некоторой базы данных. В правой части экрана высвечивается окно для выбора терминов онтологии, аналогичное тому, что показано на рис. 1. После сформирования исходного списка терминов для регистрируемых данных последовательным указанием требуемых терминов переносим их в соответствующие строки формы регистрации (см. рис. 2). Таким образом отпадает необходимость в ручном наборе названий, избегая тем самым произвола в написании терминов и названий и ошибок при наборе.

Редактирование онтологии

Разработка онтологии некоторой предметной области является довольно сложной задачей, при решении которой нужно учитывать множество аспектов, в том числе влияют и личные предпочтения составителей. Проект онтологии следует предоставить для изучения и внесения поправок другим специалистам. Для удобного выполнения этих работ разработан сервис модификации онтологии, принятой нами структуры.

Модификации в онтологию можно вносить, начиная с любого уровня, в том числе и начиная с пустой онтологии, но при этом необходимо ясно представлять весь путь до вносимого изменения.

Операция модификации предусматривает возможность удаления, замены терминов (изменением написания) и добавления новых. Отметим, что внесение дополнений в любой уровень, кроме листьев, влечет необходимость построения путей до листьев включительно, т.е. внесение одного нового слова может потребовать внесения еще определенного множества новых слов в вышележащие уровни, что необходимо ясно представлять, начиная операцию.

Большой строгости требует операция удаления, так как произвольным удалением можно разрушить структуру онтологии. Поэтому предусмотрен программный анализ возможного нарушения связности графа онтологии при удалении терминов. Результаты анализа выдаются на экран.

Заключение

Представленная здесь технология применения онтологии при создании и ведении коллекции данных о природных ресурсах региона и при поиске релевантных данных разработана на основе объектной технологии программирования с использованием средств JavaSkript, C++ и динамического HTML. Технология проверяется на фрагментах коллекции о водных ресурсах, населенных пунктах и дорогах Республики Карелия. Онтология в упомянутой выше комплектации в настоящее время проходит стадию проверки и редактирования.

Литература

[1] Лебедев В.А., Брагин С.В., Старкова В.Г. Геоинформационные коллекции о природных ресурсах региона в Интернет. // Материалы международной научно-практической конференции «Рациональное природопользование: ресурсо- и энергосберегающие технологии и их метрологическое обеспечение». Москва, 2004.

[2] Цветков В.Я. Геоинформационные системы и технологии. М., 1998.

[3] Когаловский М.Р. Энциклопедия технологий баз данных. М., 2002.

[4] Россеева О.И., Загорулько Ю.А. Организация эффективного поиска на основе онтологий. www.dialog-21.ru/Archive/2001/Volume2

[5] Бездушный А.Н., Гаврилова Э.А., Серебряков В.А., Шкотин А.В. Место онтологий в единой интегрированной системе РАН. www.benran.ru

[6] Лебедев В.А., Старкова В.Г., Брагин С.В. Представление онтологии научной коллекции «Водные ресурсы региона». // Труды Шестой Всероссийской конференции по электронным библиотекам. Пущино, 2004.

[7] Лебедев В.А., Старкова В.Г., Брагин С.В. Технология администрирования геоинформационными коллекциями. // Труды Института прикладных математических исследований Карельского научного центра РАН, вып. 5. Петрозаводск, 2004.

Using ontology to maintain and access data of the “Natural resources of the region” collection

V. Lebedev, V. Starkova, S. Bragin

The approach to the conceptualization and the structure of the “Natural resources of the region” ontology, its presentation on the computer and the technology for its application to register data, search for relevant data, as well as the technology for editing the ontology itself are described.

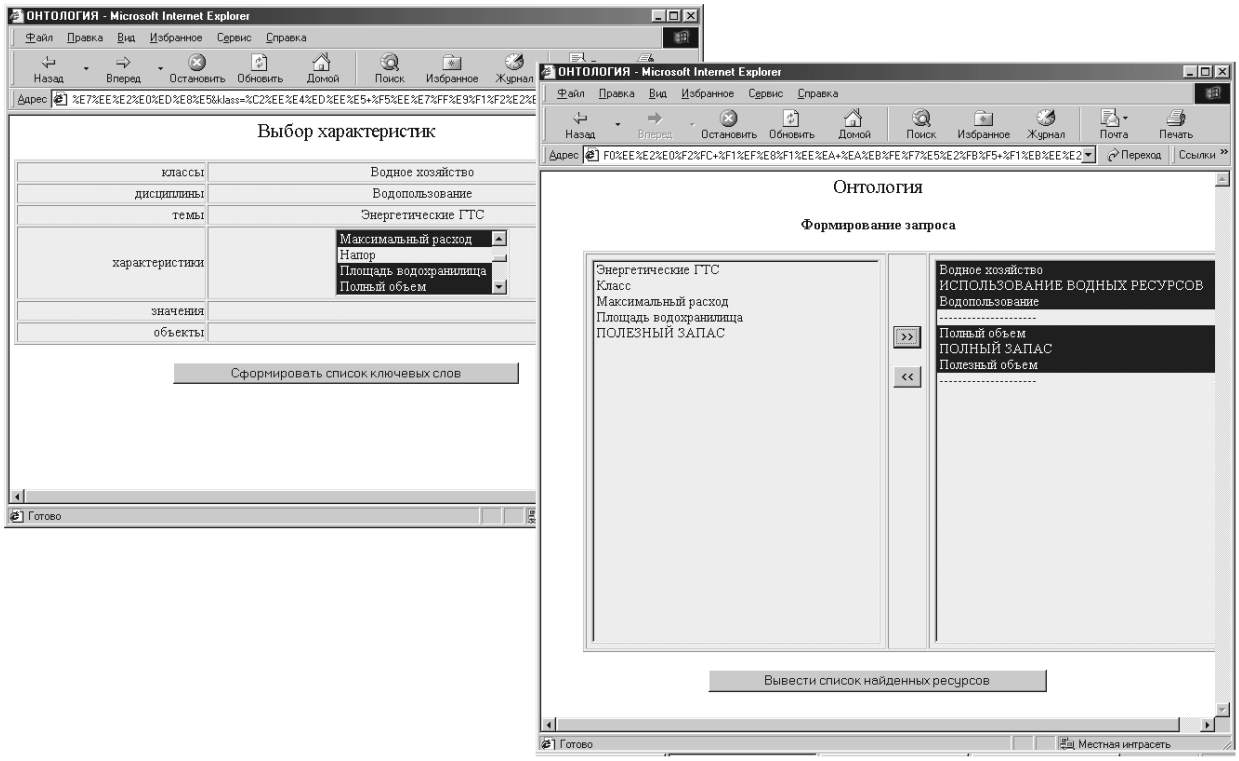
ОНТОЛОГИЯ - Microsoft Internet Explorer
Выбор характеристик
классы
Водное хозяйство
дисциплины
Водопользование
темы
Энергетические ГТС
характеристики
Максимальный расход
Напор
Площадь водохранилища
Полный объем
значения
объекты
Сформировать список ключевых слов
Готово
ОНТОЛОГИЯ - Microsoft Internet Explorer
Онтология
Формирование запроса
Энергетические ГТС
Класс
Максимальный расход
Площадь водохранилища
ПОЛЕЗНЫЙ ЗАПАС
Водное хозяйство
ИСПОЛЬЗОВАНИЕ ВОДНЫХ РЕСУРСОВ
Водопользование
Полный объем
ПОЛНЫЙ ЗАПАС
Полезный объем
Вывести список найденных ресурсов
Готово
Местная интрасеть

Рис. 1.

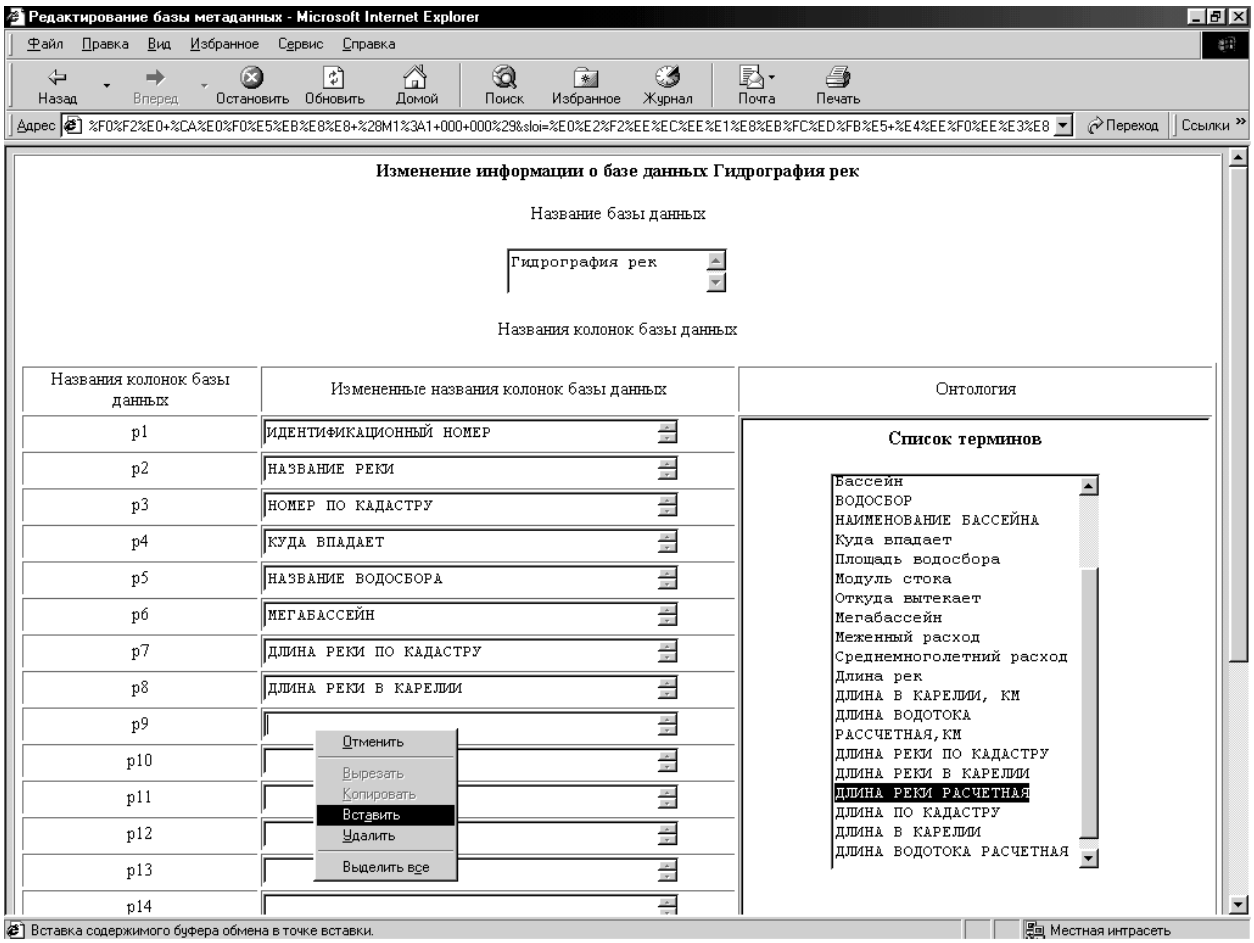
Редактирование базы метаданных - Microsoft Internet Explorer
Изменение информации о базе данных Гидрография рек
Название базы данных
Гидрография рек
Названия колонок базы данных
Названия колонок базы данных
Измененные названия колонок базы данных
Онтология
p1
ИДЕНТИФИКАЦИОННЫЙ НОМЕР
p2
НАЗВАНИЕ РЕКИ
p3
НОМЕР ПО КАДАСТРУ
p4
КУДА ВПАДАЕТ
p5
НАЗВАНИЕ ВОДОСБОРА
p6
МЕГАБАССЕЙН
p7
ДЛИНА РЕКИ ПО КАДАСТРУ
p8
ДЛИНА РЕКИ В КАРЕЛИИ
p9
p10
p11
p12
p13
p14
Отменить
Вырезать
Копировать
Вставить
Удалить
Выделить все
Список терминов
Бассейн
ВОДОСБОР
НАИМЕНОВАНИЕ БАССЕЙНА
Куда впадает
Площадь водосбора
Модуль стока
Откуда вытекает
Мегабассейн
Меженный расход
Среднемноголетний расход
Длина рек
ДЛИНА В КАРЕЛИИ, КМ
ДЛИНА ВОДОТОКА
РАССЧЕТНАЯ,КМ
ДЛИНА РЕКИ ПО КАДАСТРУ
ДЛИНА РЕКИ В КАРЕЛИИ
ДЛИНА РЕКИ РАСЧЕТНАЯ
ДЛИНА ПО КАДАСТРУ
ДЛИНА В КАРЕЛИИ
ДЛИНА ВОДОТОКА РАСЧЕТНАЯ
Вставка содержимого буфера обмена в точке вставки.
Местная интрасеть

Рис. 2.