Представление онтологии научной коллекции

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

Институт прикладных математических исследований
Карельского научного центра РАН
lebedev@krc.karelia.ru, bragin@krc.karelia.ru, starkova@krc.karelia.ru

Аннотация

Разработана онтология предметной области "Водные ресурсы региона", концептуальная и логическая схема ее представления, технология подготовки, загрузки и применения ее для поиска в коллекции релевантных информационных ресурсов.

Вода - один из важнейших природных ресурсов, от запасов и качества которого зависит здоровье населения, экологическое состояние региона и его экономическое развитие. Сведения о составе и изученности водных ресурсов, их качестве, водообеспеченности и их изменениях под влиянием природных и антропогенных воздействий представляют интерес как для властных структур, так и для научных и общественных организаций. На материалах многолетнего изучения и мониторинга водных объектов Карелии, собранных в Институте водных проблем Севера Карельского научного центра РАН (исследования продолжаются), создается научная интернет/интранет-коллекция о водных ресурсах и их использовании.

Целью создания коллекции является информационное обеспечение по контролю и мониторингу водного баланса, водообеспеченности, водопотребления, водоотведения, качества вод, эффективности водоохранных мероприятий, биопродуктивности водных объектов и их изменениям под влиянием климатических и антропогенных причин в интересах научных исследований, властных структур, хозяйствующих субъектов и общественности.

Содержанием коллекции являются сведения о пространственном расположении, климатических, геологических, гидрологических, гидрохимических и гидробиологических характеристиках водных объектов и водосборов и их изменениях, описания назначения и состояния гидротехнических сооружений, различные аналитические и обобщающие материалы в виде статей, докладов и других текстов. Пространственное расположение объектов отображается на географической карте и сведения о них связаны с картой так, что доступ к ним обеспечивается как через указатель (оглавление), так и через карту [1,2].

Структура системы управления и использования коллекции показана на рис. 1. Она включает программное обеспечение ГИС сервера с инструктивно-методическими материалами по инсталляции и настройке сервера на предметную область (создание "пустой коллекции"), по подготовке и загрузке данных. В качестве Web- сервера используется свободно распространяемый продукт Apache, базы данных и базу метаданных обслуживает РСУБД MySQL, взаимодействие которой с Web-сервером обеспечивается CGI - сценарием через ODBC. Скрипты сервера написаны на C++. Для загрузки данных и ведения коллекции предназначен сайт администратора, обеспечивающий выбор и исполнение функций администрирования. Так, загрузка данных на сервер может выполняться по сети непосредственно с компьютеров, на которых эти данные подготовлены. При этом выполняется регистрация ресурсов и формирование базы метаданных. Помимо этого сайт администратора обеспечивает выполнение функций по защите данных, разделению доступа, по регистрации пользователей и предоставлению им прав доступа.

Доступ к данным осуществляется через сайт пользователя, который осуществляет контроль полномочий пользователей и предоставляет им средства для поиска и выбора релевантных информационных ресурсов. При формировании базы метаданных одновременно создается список ресурсов, из которого производится выборка подписка открытых ресурсов данному пользователю по составу его полномочий, куда включается также список ресурсов свободного доступа. Данный список открывается пользователю для выбора. В реальных условиях количество ресурсов в коллекции исчисляется сотнями, следовательно, выбор релевантных представляет проблему. Поэтому целесообразно предоставить пользователю возможность поиска по ключевым словам. Для решения этой задачи предложено использовать онтологию предметной области.

Ввиду многочисленности водных объектов и различном их значении для природы, науки и хозяйства программы их изучения и изученность также различны. Например, в Карелии насчитывается более 63 тысяч озер, большинство из

Труды 6ой Всероссийской научной конференции "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - RCDL2004, Пущино, Россия, 2004.

которых имеют площадь менее 1 кв. км. и на карте М1:1000000 изображено около 5200. Из этого количества по полной программе подвергаются мониторингу несколько десятков, несколько сотен контролируются эпизодически или однократно. По большинству имеются фрагментарные сведения. В таких условиях в составе коллекции целесообразно иметь онтологию, которая отображает полный перечень водных объектов (реки, озера, водохранилища, водосборы, подземные воды) и полную программу их исследований и мониторинга. В то же время она должна показывать фактическую изученность водных объектов. Такой подход позволяет создать некоторую структуру, отвечающую на ряд важных вопросов о содержании коллекции без просмотра самой коллекции. Например, такие вопросы:

- Какие озера изучаются по полной программе?
- Какие характеристики имеются по такому-то объекту?
- По каким объектам имеются гидробиологические данные?
- и т.п.

Далее рассматривается подход к построению онтологии и организации сервиса взаимодействия с ней пользователей.

В соответствии с определением [3] онтология есть результат концептуализации предметной области. В результате инвентаризации выявляются списки понятий предметной области и связей между ними. В нашем случае устанавливаем списки водных объектов и их классов, темы (дисциплины) их изучения, состав характеристик и процессов, подлежащих контролю и изучению. Процессы подразделяются на краткосрочно циклические и длительные. К первым относятся сезонные (термические, гидродинамические, вегетационные и некоторые другие); ко вторым - процессы эвтрофирования, загрязнения (в частности, закисления), изменения климата. Некоторые из первых могут быть представлены в виде моделей. Однако большинство процессов представляются в виде временных рядов изменения определенных характеристик. В коллекции все процессы будут представляться временными рядами, характеризующими их статистиками и графическими материалами.

В результате инвентаризации получен ряд функциональных зависимостей: объекты-классы (многие к одному), объекты-темы (многие ко многим), характеристики-темы (многие к одному), характеристики-значения (один - ко многим). Зависимость типа ¡многие ко многим¿ неконструктивна. Поэтому выполнена ее декомпозиция на две: объекты-подклассы (многие к одному) и темы-подклассы (многие к одному). Дополнительно определена зависимость "классы-подклассы". Содержательно подкласс объектов определяется соответствующей программой изучения, объединяющей некоторую совокупность тем. Функция "значения-характеристики" содержит данные о характеристиках с символьными значениями. Если представить множества значений левых и правых частей этих зависимостей в виде доменов, то зависимости представляются в виде бинарных отношений (реляционных таблиц). При этом оказывается, что от характеристик к объектам и обратно каждая тройка колонок обладает свойством транзитивности. Это свойство обеспечивает возможность построения реляционной базы данных онтологии, формирования запросов и получения ответов по принципу работы оператора Select. На рис. 2 представлен фрагмент зависимости "темы- характеристики".

Логически рассматриваемая онтология является графом сравнительно сложной структуры без циклов. Такого рода графы обычно редуцируются в дерево посредством расщепления вершин с более чем одной входящей дугой. В этом случае на уровне листьев наблюдаются блоки повторяющихся имен, поэтому выбор релевантных осуществляется при движении по дереву от корней к листьям, как это показано ниже.

Особенность использования данного дерева заключается в том, что выбрав темы и соответствующие им характеристики и значения, необходимо выбрать также объекты, описания которых могут содержать данные по выбранным темам. Указанные выше зависимости должны трактоваться как дуги данного дерева, представленные парой имен смежных вершин. При этом должна быть обеспечена его связность. Каждая зависимость представляет собой конкретный уровень в дереве, а связность обеспечивается совпадением имен левой и правой вершин смежных дуг.

В этих условиях распределение множества дуг по уровням дерева как раз соответствует табличной структуре реляционной базы данных.

База данных онтологии содержит порядка 5200 озер, 3360 рек, около 20 водохранилищ, более 520 гидротехнических сооружений, 65 тем, 270 характеристик, около 50 значений, более 100 синонимических гнезд.

Для получения ответов на осмысленные вопросы относительно содержания коллекции, необходимо, чтобы наборы ключевых слов статей и баз данных как минимум, содержали списки характеристик и значений, представленных в них, и названий объектов, к которым они относятся (Наборы ключевых слов могут содержать также слова, относящиеся к темам, подклассам и классам, но никаких других слов, кроме синонимов, включать в эти наборы не следует). Тогда, представив набор из ключевых слов информационных ресурсов в виде зависимости: набор ключевых слов - название информационного ресурса, создаем возможность получать из онтологии ответы на вопросы о содержании в коллекции тех или иных данных.

Проведенный анализ показал, что среди названий объектов часто наблюдается омонимия. Для

исключения ее целесообразно к таким названиям добавлять расширения, уточняющие названия по местоположению объектов. Вместе с тем в терминологии часто встречаются синонимы, употребляемые в текстах и базах данных наряду с основными (доминантами). Для ликвидации разрыва между списком характеристик и наборами ключевых слов включается зависимость, типа ¡один ко многим¿, доминанты-синонимы, которая применяется для расширения запросов.

Таким образом, онтология коллекции данных о водных ресурсах состоит из онтологии предметной области и функциональной зависимости "наборы ключевых слов - названия информационных ресурсов". Первая представляет собой декомпозиционную базу данных, содержащую ряд двухграфных таблиц, соответствующих указанным выше бинарным зависимостям. Вторая является двухграфной таблицей, в которой первая графа содержит названия информационных ресурсов, а вторая, многозначная, содержит наборы ключевых слов, определяющих содержание этих ресурсов.

Для занесения и модификации регистрационных сведений в базу метаданных разработаны специальные формы и скрипты. В составе БМД имеется таблица ключевых слов. Загрузка ресурсов на сервер выполняется по сети при помощи программы "Загрузчик", являющейся частью сайта администратора коллекции. Все эти средства используются также для создания онтологии коллекции и взаимодействия с нею (рис. 1).

Помимо перечисленных средств разработан сервис формирования схемы базы данных онтологии.

Первоначально онтология сформирована при помощи пакета Excel и подвергнута тестированию на связность по алгоритму прохождения всех путей в графе (поиска в глубину). В процессе тестирования были устранены неполнота и несоответствие имен (разрывы в графе). В процессе загрузки выполнена конвертация данных в формат MySQL. Для добавления и изменения данных в онтологию разработаны специальные формы и используются скрипты модификации БМД.

Для использования онтологии при работе пользователя с коллекцией разработаны сервисы для формирования запросов и поиска релевантных запросу ресурсов. Сервис запросов содержит форму для построения запроса к онтологии и для построения запроса к БД ключевых слов информационных ресурсов (рис. 3), а также форму, визуализирующую отклик в виде списков релевантных статей и баз данных, для вызова и просмотра этих ресурсов. В случае, если списки оказываются пустыми, предусмотрен возврат на предыдущую форму для редактирования запроса (например, посредством отбрасывания некоторых слов из запроса).

Алгоритм сервиса формирования запроса основан на циклическом применении механизма заполнения полей в заготовке оператора Select данными, последовательно получаемыми из формы построения запроса. Выбор значений в форме и активизация соответствующей колонки определяют двухграфную таблицу базы данных и значения строк для отбора. После этого оператор Select срабатывает, производит отбор строк и этот отклик передается обратно на форму запроса, подготавливая данные для следующего шага. Формирование запроса к онтологии выполняется следующим способом. На визуализированной форме (рис. 3) представлены колонки: объекты, подклассы, классы, темы, характеристики и значения. Название каждой из колонок является кнопкой, запускающей скрипт отбора данных. (Колонки функций: "доминанты- синонимы" не показаны, так как не требуют от пользователя выбора). В колонке "классы" сразу представлен список классов (см. рис. 3), в котором следует выбрать только один. После этого пользователь может выбрать либо интересующие его подклассы объектов, либо темы из представленных в соответствующих колонках, предварительно активизировав нажатием соответствующей кнопки одну из них. Тем самым запускается скрипт выборки. Для определенности рассмотрим процесс построения запроса через темы. Выбрав класс "Озера", пользователь хочет узнать, какие характеристики содержат темы "Качество воды" и "Биопродуктивность". При активизации колонки "Характеристики" высвечиваются соответствующие списки (см. рис. 2, 3). Чтобы занести все или некоторые из характеристик в запрос, он должен их отметить. Далее он может добавить значения некоторых характеристик. Активизировав колонки "Подклассы" и "Объекты" он может выбрать озера, программа исследования которых содержит указанные им темы. Отметим, что построение запроса он может начать с колонки "Подклассы", а не "Темы", но ни с какой другой, т.к. только эти колонки непосредственно связаны с "Классами".

После того, как пользователь отметил значения во всех интересующих его колонках, необходимо составить список ключевых слов для поиска. В список нужно добавить синонимы, нажав кнопку "Сформировать список ключевых слов". И тем самым запустить скрипт формирования списка, поиска синонимов и добавления их в список. Весь список вместе с синонимами визуализируется в нижнем окне формы (рис. 3). При этом все термины и названия (слова и словосочетания) разделяются запятыми, а подсписки синонимов заключаются в квадратные скобки как признак того, что из каждого такого подсписка достаточно присутствие одного термина в искомом информационном ресурсе.

Далее пользователь должен оценить обязательность присутствия каждого термина в запросе. Некоторые он может убрать, другие заключить в круглые скобки как признак их необязательности. Кроме того из некоторых терминов он может сформировать подсписки альтернатив, заключив их в квадратные скобки.

После сформирования запроса нажатием кноп-

ки "Запрос" происходит запуск скрипта, осуществляющего поиск информационных ресурсов в БД ключевых слов. Он выполняется последовательным сопоставлением обязательных терминов запроса с наборами ключевых слов этой БД. При полном совпадении обязательных терминов запроса с терминами ресурса (сравниваются только основы слов) имя ресурса помещается в отклик. При этом по расширению имени ресурса определяется, статья это или база данных. Отобранные ресурсы через визуализированную форму можно просмотреть. Если списки ресурсов оказываются пустыми, предусмотрен возврат для редактирования запроса.

Таким образом, онтология способствует формированию запросов к коллекции, предоставляя пользователю списки терминов в их логической взаимосвязи в структуре описания предметной области. База данных онтологии и программные сервисы подготовлены. Предусмотрены средства модификации содержания онтологии.

Литература

[1] Лебедев В.А., Старкова В.Г., Брагин С.В. Публикация в Интернет коллекций типа справочников Труды IV Всероссийской конференции по электронным библиотекам. Дубна, 2002.

[2] Лебедев В.А., Брагин С.В., Старкова В.Г. Геоинформационная коллекция по водным ресурсам // Труды Института прикладных математических исследований. Петрозаводск, 2003.

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

Presentation of the ontology of the "Regional Water Resources" scientific collection

V. Lebedev, V. Starkova, S. Bragin

The ontology of the subject area "Regional Water Resources", its conceptual and logical schemes, the technology for preparing and uploading the collection and using it to find relevant resources were developed.

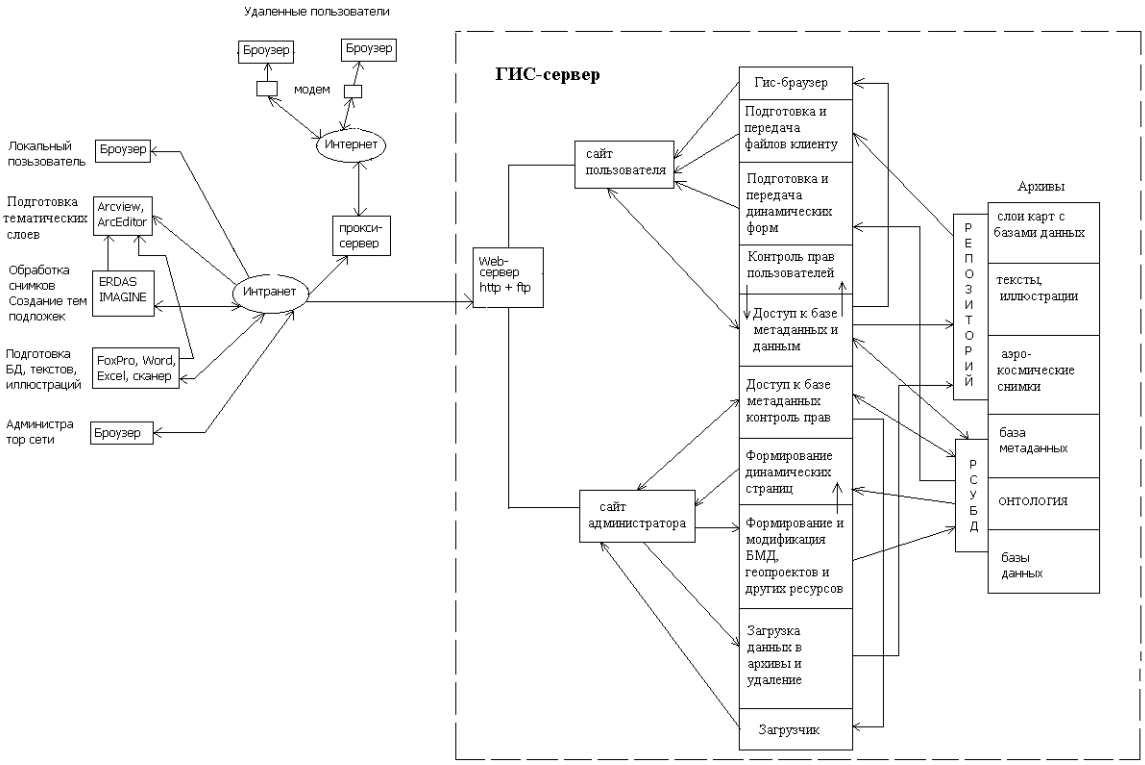
Удаленные пользователи
Броузер
Броузер
модем
Интернет
прокси-сервер
Локальный позьзователь
Броузер
Подготовка тематических слоев
Arcview, ArcEditor
Обработка снимков Создание тем подложек
ERDAS IMAGINE
Интранет
Подготовка БД, текстов, иллюстраций
FoxPro, Word, Excel, сканер
Администратор сети
Броузер
Web-сервер http + ftp
ГИС-сервер
сайт пользователя
сайт администратора
Гис-браузер
Подготовка и передача файлов клиенту
Подготовка и передача динамических форм
Контроль прав пользователей
Доступ к базе метаданных и данным
Доступ к базе метаданных контроль прав
Формирование динамических страниц
Формирование и модификация БМД, геопроектов и других ресурсов
Загрузка данных в архивы и удаление
Загрузчик
Архивы
РЕПОЗИТОРИЙ
РСУБД
слои карт с базами данных
тексты, иллюстрации
аэро-космические снимки
база метаданных
онтология
базы данных

Рис. 1.

Темы	Характеристики
Качество воды	Класс воды
	Прозрачность
	Цветность
	Взвеси
	Минерализация
	Гумификация
	Кислотность (pH)
	Металлы
Изменение качества воды	Временной интервал
	Сточные воды
	Металлы
	Нефтепродукты
	Фенол
	Фурфурол
	Лигносульфонаты
	Калий
	Азот
	Фосфор
	Изменение трофности
	Закисление
Химический баланс	Речной приток
	Атмосферные осадки
	Сточные воды
Биопродуктивность	Трофность
	Среднегодовая продукция
	Среднегодовая биомасса
	Биомасса детрита
Фитопланктон	Основные группы
	Численность
	Первичная продукция
	P/B коэффициент
	Фитосинтез
	Хемосинтез
	Сезонная динамика продукции
	Утилизация солнечной энергии

Рис. 2.

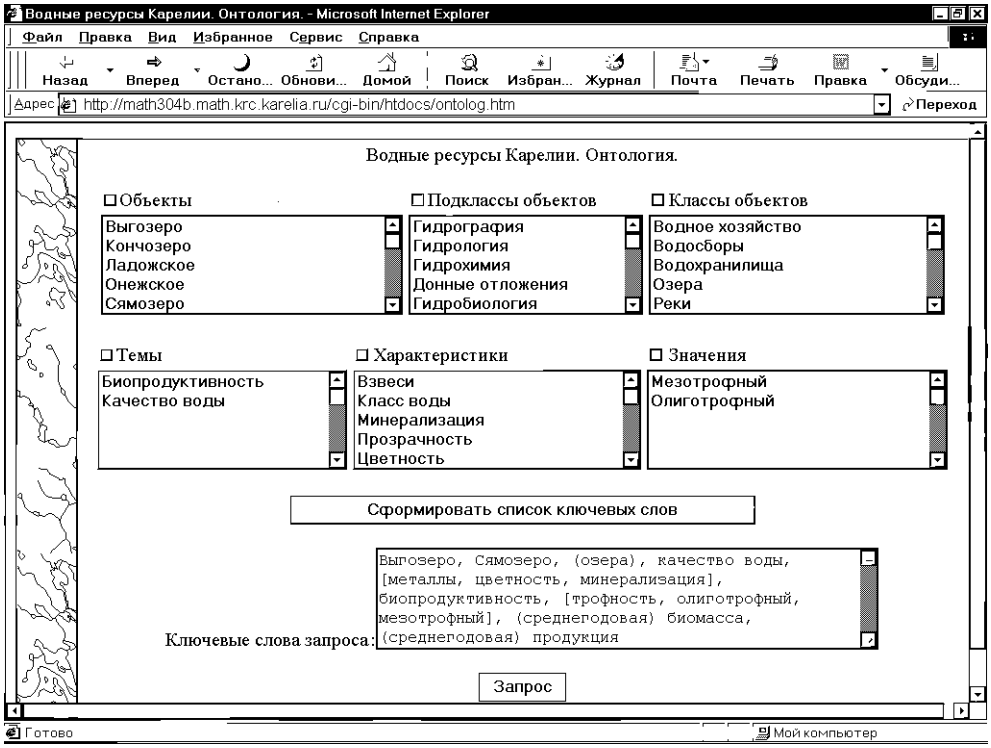
Водные ресурсы Карелии. Онтология. - Microsoft Internet Explorer
Файл Правка Вид Избранное Сервис Справка
Назад Вперед Остано... Обнови... Домой Поиск Избран... Журнал Почта Печать Правка Обсуди...
Адрес http://math304b.math.krc.karelia.ru/cgi-bin/htdocs/ontolog.htm
Переход
Водные ресурсы Карелии. Онтология.
Объекты
Выгозеро
Кончозеро
Ладожское
Онежское
Сямозеро
Подклассы объектов
Гидрография
Гидрология
Гидрохимия
Донные отложения
Гидробиология
Классы объектов
Водное хозяйство
Водосборы
Водохранилища
Озера
Реки
Темы
Биопродуктивность
Качество воды
Характеристики
Взвеси
Класс воды
Минерализация
Прозрачность
Цветность
Значения
Мезотрофный
Олиготрофный
Сформировать список ключевых слов
Ключевые слова запроса:
Выгозеро, Сямозеро, (озера), качество воды,
[металлы, цветность, минерализация],
биопродуктивность, [трофность, олиготрофный,
мезотрофный], (среднегодовая) биомасса,
(среднегодовая) продукция
Запрос
Готово
Мой компьютер

Рис. 3.