Электронная коллекция информационных ресурсов по топонимии Европейского Севера России [*)].

Вдовицын В.Т., Керт Г.М.,
Беляева Н.А., Луговая Н.Б., Сорокин А.Д., Чуйко Ю.В.

Институт прикладных математических исследований
Институт языка, литературы и истории
Карельский научный центр РАН
185610: г. Петрозаводск, Пушкинская 11
E-mail: {vdov,kert}@krc.karelia.ru

АННОТАЦИЯ

В работе рассматриваются вопросы создания и развития электронной коллекции информационных ресурсов по топонимии Европейского Севера России, которая представлена в виде тематического Web-сайта с интегрированными базами данных и программами их обработки. Эта система предназначена для информационной поддержки исследований и разработок российских топонимистов с использованием Интернет-технологии.

Ключевые слова: электронная коллекция, Web-сайт, база данных.

This paper reviews problems of creation and development of electronic collection of information resources on toponymics of European North of Russia, which is represented as a thematic Web-site with integrated databases and programs to process them. The system is designed to support russian toponymists' researches and developments with the aid of Internet technologies.

Key words: electronic collection, Web-site, database.

[*)] Работа поддержана грантом РГНФ № 00-04-12020в

1. Введение.

В настоящее время одной из перспективным форм организации и проведения совместных научных исследований и разработок является создание в среде Интернет так называемых виртуальных лабораторий (collaboratory) [1-4]. При этом Интернет-технологии обеспечивают оперативный обмен информацией между исследователями, доступ к инструментарию, распределенным базам данных и знаний, вычислительным ресурсам, а также проведение тематических телеконференций и выпуск электронных сборников публикаций.

В качестве примера такой системы можно привести проект Worm Community System (http://www.canis.uiuc.edu), в рамках которого ученые-биологи (molecular biologists) могут с помощью Интернет просматривать соответствующие данные и знания, редактировать и пополнять их, обмениваться мнениями. Следует отметить, что эту систему можно рассматривать как пример электронной библиотеки, содержащей электронные информационные ресурсы в определенной области биологии (nematode wormC.elegans).

В данной работе рассматриваются вопросы создания и развития электронной коллекции информационных ресурсов по топонимии Европейского Севера России, которая представлена в виде тематического Web-сайта с интегрированными базами данных и программами их обработки. Эта система предназначается для информационной поддержки исследований и разработок российских топонимистов с использованием Интернет-технологии и может стать первым шагом на пути создания тематической collaboratory.

2. Цели и задачи проекта

Данный проект направлен на создание и развитие специализированного тематического Web-сайта с интегрированными базами данных по топонимии Европейского Севера России и программами их обработки с целью содействия развитию российской топонимической науки, координации теоретических и прикладных исследований и разработок российских топонимистов в различных регионах страны, оперативного обмена информацией, организации и проведения телеконференций и выпуска сборников электронных публикаций.

Необходимость использования компьютеров в исследовании топонимии обуславливается, в первую очередь, стремительным увеличением вводимого в научный оборот материала, возможностью его системной классификации в структурном, временном и пространственном аспектах.

Основное назначение, функция топонимов – это выделение, индивидуализация, идентификация именуемых объектов среди подобных. В настоящее время в условиях становления информационного общества трудно переоценить роль и значение топонимов как точных ориентиров на местности. Поэтому одной из первых задач является создание региональных сводов, фиксирующих точное написание и произношение топонимов. Также представляет интерес решение следующих задач в этой области:

■ выявление субстратного слоя топонимии, его фонетических и морфологических особенностей;

■ определение ареалов топонимии, этимологизируемой с помощью живых языков Европейского Севера России (прибалтийско-финские, саамский, коми, ненецкий, русский и пр.);

■ исследование ареалов смешения топонимии двух и более языков;

■ определение абсолютных и относительных частот топонимов каждого языка на единицу площади и др.

Эти задачи требуют для своего решения значительных усилий как со стороны ученых-топонимистов, так и специалистов в области математики и информатики. Создание тематического Web-сайта в сети Интернет позволит, на наш взгляд, объединить и скоординировать усилия многих специалистов, заинтересованных в решении этих задач.

3. Структура тематического Web-сайта и используемые технологии

Исходя из сформулированных задач нами предложена структура информационного содержания тематического Web-сайта по топонимии Европейского Севера России (http://toris.krc.karelia.ru), включающая следующие основные разделы:

■ БАЗЫ ДАННЫХ (база данных, содержащая информации о топонимах; библиографическая база данных);

■ ФОРУМ, для организации и проведения тематических электронных конференций;

■ ЭЛЕКТРОННЫЕ ПУБЛИКАЦИИ, для размещения электронных материалов по вопросам исследования топонимии;

■ ССЫЛКИ на информационные ресурсы по топонимии в Интернет, КАРТА САЙТА, НОВОСТИ и КОНТАКТНАЯ информация.

Тематический Web-сайт создан в среде ОС UNIX с использованием Web-сервера Apache, СУБД PostgreSQL и технологии CGI сценариев для обеспечения доступа к базам данных.

4. База данных по топонимии

Структура описания русской, прибалтийско-финской и саамской топонимии, разработанная на основе многолетних исследований д.ф.н. Керта Г.М., учитывает структурные, семантические и иные свойства топонима, а также экстралингвистические признаки объекта, именуемого данным топонимом [5, 6]. При этом описание прибалтийско-финских и саамских топонимов разделяется на части (определяемую и определяющую) и на компоненты - значимые и неэтимологизируемые. Вычленение неэтимологизированных компонентов позволит ввести в научный оборот всю субстратную (невыявленную) топонимию. В соответствии со значением топонимная лексика с некоторыми коррективами, обусловленными ее спецификой, распределяется на семантические классы по классификации Р. Халлига и В. фон Вартбурга, изложенной в монографии "Система понятий как основа для лексикографии" (Rudolf Hallig und Walter von Vartburg. Begriffssystem als Grundlage für die Lexikographie, Berlin, 1952).

База данных по топонимам состоит из 13 таблиц - основной "Топоним на языке представления" и двенадцати вспомогательных: "Географический объект"; "Компонент"; "Язык"; "Структурная формула"; "Падеж"; "Диалект"; "Семантическая формула"; "Часть речи"; "Объект"; "Область/республика"; "Район"; "Населенный пункт". Структура базы данных представлена на рис. 1. Разработаны интерфейсные формы доступа к данным, позволяющие пользователям формулировать запросы, пополнять, редактировать и удалять данные из базы данных. В настоящее время в базу данных введены описания свыше тысячи топонимов.

Рис.1 Структура базы данных по топонимии.

Структура базы данных библиографического справочника по прибалтийско-финской топонимии состоит из двух основных частей. В первой части структуры записи представлены поля, описывающие библиографические данные о публикациях согласно российскому коммуникативному формату представления библиографических записей (RUSMARK). Во второй части представлены поля, описывающие каждую публикацию с точки зрения ее содержания; например, объект исследования, семантика, этимология, субстрат, история изучения, язык исследования. При этом предполагается, что для каждой публикации эксперты будут заполнять вышеперечисленные поля с использованием специально разработанных интерфейсных форм. Таким образом, в системе реализована возможность получения информации о публикациях по запросам, в которых могут быть указаны содержательные признаки публикации (например, можно формулировать

запросы на поиск публикаций, в которых рассматриваются вопросы семантики топонимов и т.д.).

5. Анализ данных

В данной работе под анализом данных мы подразумеваем применение KDD-методов (Knowledge Discovery in Databases) к анализу информации о топонимах с целью получения правил, выражающих так называемые ассоциативные связи между характеристиками описания топонимов в базе данных [7].

Ассоциативным правилом (т.е. правилом, выражающим ассоциативную связь между характеристиками описания топонимов в базе данных) будем называть импликацию следующего вида

{Antecedent => Consequent | c, s},

где:

Antecedent и Consequent – непересекающиеся множества характеристик описания топонимов, определяющие, соответственно, посылку и следствие правила;

c – фактор уверенности правила;

s – степень поддержки правила.

При этом утверждается, что данное ассоциативное правило удовлетворяет исходному набору записей базы данных с фактором уверенности c ($0 \le c \le 1$), если в данном наборе по крайней мере c% записей, содержащих посылку правила, содержат и следствие. Степень поддержки правила s определяется как отношение количества записей, содержащих посылку и следствие правила, к общему количеству всех записей в исходном наборе.

Для решения этой задачи разработан исследовательский прототип программной системы DMiner, в основу которой положен алгоритм генерации значимых ассоциативных правил [8]. Система реализована с использованием пакета разработчика на языке Java JDK 1.2.2 (SDK 2). В настоящее время ведутся эксперименты, связанные с обнаружением значимых для предметников ассоциативных связей между характеристиками описания топонимов в базе данных, а также с учетом пожеланий пользователей совершенствуется интерфейс и отрабатывается методика работы с системой.

Литература

1. R.T. Konzes, J.D. Myers, W.A.Wolf Collaboratories: doing science on the Internet. IEEE Computer, vol.29, August 1996.
2. Vladimir T. Vdovitsyn, Vladimir V. Tarasov Multi-Agent System for International Support of Collaborative Researchers Work in a Computer Network // Proc. of FDPW'99. Developments in Distributed Systems and Data Communications. Vol.2. Petrozavodsk. 1999. P. 139-145.
3. Вдовицын В.Т., Керт Г.М., Сорокин А.Д., Русаков С.М. Информационная технология для поддержки совместной работы исследователей в сети Internet – перспективы развития TORIS //Тр. межд. телеконф. "Информационные технологии в гуманитарных науках", Татарстан. 1999. http://www.kcn.ru/tat_ru/universitet/gum_konf/index.htm.
4. Вдовицын В.Т., Сорокин А.Д., Тарасов В.В. Создание и развитие электронных информационных ресурсов на основе Интернет-технологий для поддержки научных исследований в КарНЦ РАН //Материалы Всерос. объединенной конф. "Технология современного общества – Интернет и современное общество", Санкт-Петербург. 2000. С. 133-135.
5. Сорокин А.Д., Вдовицын В.Т., Луговая Н.Б. Создание и развитие электронных информационных ресурсов в КарНЦ РАН //Сб. докл. Второй Всерос. научн. конф. "Электронные библиотеки: перспективные методы и технологии, электронные коллекции". Протвино, ГИЦ ИФВЭ. 2000. С. 3-5.
6. Керт Г.М., Вдовицын В.Т., Веретин А.Л., Луговая Н.Б. Компьютерный банк топонимии Европейского Севера России: TORIS //Тр. межд. телеконф. "Информационные технологии в гуманитарных науках", Татарстан. 1999. http://www.kcn.ru/tat_ru/universitet/gum_konf/index.htm.
7. G. Piatetsky-Shapiro (Editor), Knowledge Discovery in Databases, AAAI/MIT Press, 1991.
8. Rakesh Agrawal, Tomasz Imielinski, Arun Swami Mining Association Rules between Sets of Items in large Databases // Proc. of the 1993 ACM SIGMOD Conf. Washington DC, USA, May 1993