Разработка географического тезауруса под управлением системы автоматизации библиотек ИРБИС64

Баженов С. Р.,
ГПНТБ СО РАН

25 сентября 2019 г.

Географическая информационная система - это информационная система, обеспечивающая сбор, хранение, обработку и визуализацию пространственных данных и связанной с ними информации.

Сегодня, за счет того, что технологии шагнули далеко вперед, географические данные стали доступны для широкого круга задач.

Это так называемые «негеографические» информационные системы, к которым относятся, например, **электронные каталоги, базы данных научно-технической информации, архивы** с информацией о цифровых и нецифровых объектах

Любая статья была где-то написана и опубликована, любой экспонат музея был где-то найден, тексты научных трудов зачастую содержат названия географических объектов.

Географическая информация хранится в текстовых полях и пригодна только для простейшего текстового поиска по географическому названию.

Результаты такого поиска будут заведомо неточны, поскольку названия географических объектов могут изменяться с течением времени, географические объекты могут иметь несколько названий, географические объекты могут исчезнуть с течением времени.

Задача создания прототипа ретроспективного тезауруса географических названий в ИРБИС64 была поставлена в рамках интеграционного проекта СО РАН № 0334-2018-0002.

Одной из основных задач этого проекта является разработка алгоритмов извлечения географических названий из произвольных текстовых документов. Решением этой задачи занимаются Институт вычислительных технологий СО РАН и Институт систем информатики СО РАН.

ГПНТБ СО РАН поставлены задачи создания экспертных наборов данных на основе различных текстовых документов и информационных ресурсов, содержащих географическую информацию, в том числе ретроспективную для отладки разработанных алгоритмов индексации и создание прототипа географического тезауруса с данными определенного региона.

Тезаурусом называется совокупность терминов (лексических единиц), описывающих конкретную предметную область, с указанием семантических отношений (связей).

Возможно ли создание, ведение и использование тезаурусов в ИРБИС64?

Медицинский тезаурус MeSH (Medical Subject Headings)

- Разрабатывается в Национальной медицинской библиотеке США с 1960 год.
- Используется для задач систематизации и поиска в большинстве медицинских библиотек страны.

Структура тезауруса:

- Дескрипторы (рубрики)
- Модификаторы (подрубрики)
- Синонимы (перекрестные ссылки).

Дескрипторы и синонимы связаны перекрестными ссылками.

В ИРБИС64 имеется специально разработанный аппарат для создания, ведения и использования этого тезауруса

- **Основной термин: АБСОРБИРУЕМЫЕ ИМПЛАНТАТЫ - Дескриптор**
- **Английский эквивалент: ABSORBABLE IMPLANTS**
- **Коды MeSH:**
- E07.695.025
- **Синонимы:**
- БИОАБСОРБИРУЕМЫЕ ИМПЛАНТАТЫ
- ИМПЛАНТАТЫ АБСОРБИРУЕМЫЕ
- **Русские подрубрики: ВВ, ВЕ, ВИ, ИП, ИС, КЛ, МИ, ПП, ПР, ПС, СК, СР, СТ, ТЕ, ЭК**
- **Английские подрубрики: AE, CL, CT, EC, HI, MI, PS, PX, SD, SN, ST, TD, UT, VE, VI**
- **Аннотация: коорд с субстанцией, если уместно**

Сельскохозяйственный тезаурус

ЦНСХБ использует Тезаурус в своей базе данных «АГРОС» в автоматизированной системе «Артефакт» для индексирования входного документального потока.

В стандартном ИРБИС64 отсутствует возможность работы с этим тезаурусом, но в ИРБИС64 имеется и аппарат для создания «произвольного» тезауруса.

ГПНТБ СО РАН имеет опыт использования этого аппарата на примере создания Сельскохозяйственного тезауруса.

В прошлом году мы рассказывали о начале работ по этому проекту.

В 2019 году работы по созданию прототипа географического тезауруса были продолжены.

В рамках решения задачи создания экспертных наборов данных была создана БД документов для тестирования разработанных алгоритмов индексации данных общим объемом в 500 документов, содержащая статьи из журналов, газет, сборников, однотомные и многотомные издания с рефератами. В 16 из них имеются полные тексты.

Описания содержат географические термины, выделенные из текстов экспертом, необходимые для проверки работы алгоритмов индексации.

Из поставленных задач были сформулированы основные требования к структуре тезауруса:

- производить прямое и обратное геокодирование;
- производить ретроспективное прямое и обратное геокодирование;
- предоставлять доступа к информации по стандартному протоколу;
- позволять учитывать в процессе поиска административную принадлежность географических объектов;
- предоставлять данные в схеме, максимально приближенной к какой-либо стандартной;
- производить поиск по стандартному набору атрибутов, характерных для тезаурусов.

Для создания структуры географического тезауруса и определения источников его наполнения были исследованы существующие схемы данных и существующие тезаурусы географических наименований

Тезаурус географических названий Российской государственной библиотеки.

Тезаурус не содержит ретроспективных данных в записях. Невозможно получить ни данных о предыдущих названиях, ни данных о предыдущих координатах объектов.

Координаты географических объектов заданы в виде координат точек, что не совсем соответствует действительности. Из записей могут быть получены данные о иерархических связях с помощью обработки. Явным образом иерархические связи не указаны.

Служба геокодирования API Карт Google

- позволяет определить координаты объекта, а также найти адрес наиболее близкий к указанным координатам .
- Координаты объекта приводятся в виде точки и в виде прямоугольной области. В записях содержатся иерархические связи с другими объектами.
- В то же время стоит отметить, что тезаурус содержит данные не только о крупных географических объектах, но также и об адресах. А также есть возможность обратного геокодирования.

Служба Яндекс.Карт

- предлагает своим пользователям сервис геокодирования. Он позволяет определять координаты и получать сведения о географическом объекте по его названию или адресу и наоборот, определять адрес объекта на карте по его координатам (обратное геокодирование).
- Ответ службы геокодирования не содержит сведения о предыдущих состояниях географического объекта. Координаты географического объекта представлены в виде точки и прямоугольника. Есть информация об административной принадлежности географического объекта.

ОКАТО - общероссийский классификатор объектов административно-территориального деления

В классификаторе принята иерархическая система классификации. Всё множество объектов административно-территориального деления подразделяется на группы согласно территориальному делению и эти группы располагаются по трём уровням классификации в соответствии с административной подчинённостью, причём в каждый уровень включаются объекты, непосредственно подчинённые объектам предыдущего уровня. Классификатор не содержит географическую (геометрическую) привязку.

Государственный каталог географических названий РОСРЕЕСТР

- содержит полный реестр официальных географических названий по регионам с точечными координатами.

К сожалению, требованиям, изложенным выше, не удовлетворяет ни одна из перечисленных выше баз данных.

Тем не менее, в каждой из них есть своя уникальная часть, которая необходима для полноценного описания географического объекта.

Поэтому наиболее продуктивным было бы построение собственной базы данных географических названий методом объединения записей из различных источников при сохранении оригинальной идентификации объектов

- В результате исследования разработана структура прототипа тезауруса, включающая два типа документов: **географические объекты** и **административные единицы**.

Структура тезауруса включает возможности:

- учитывать многоязычность (русский, английский, казахский и пр.);
- фиксировать геометрию объекта в геометрических примитивах (точка, окружность, прямоугольник, полигон) в фиксированных системах координат, принятых в ГИС системах;
- учитывать зависимость географического названия от времени (например, один и тот же географический объект может иметь разные названия в разные временные отрезки времени, один и тот же географический объект может иметь разную геометрию в разное время);
- учитывать нормативные документы по смене названия и/или геометрии географического объекта.

- В рамках проекта также решается задача извлечения географических названий из текстов для автоматического индексирования документов географическими терминами.

Проверка извлечения географических названий - Google Chrome

Не защищено | db4.sbras.ru/ZooSPACE/gw/geo_text1.php

Проверка извлечения географических названий из текста

Тип конфигурации: (•) Spell ( ) Stem

Коды: **ГРНТИ:** 39.19.25

Авторы: Романова Е.В.

Заглавие: Лишайники городов-спутников город Новосибирска

Реферат: Исследована лихенофлора четырех населенных пунктов в окрестностях город Новосибирска: город Бердск, город Обь, поселок Краснообск и Наукоград Кольцово. Выявлено 79 видов лишайников из 18 семейств и 38 родов, проведены таксономический, географический и экологический анализы. На основе индекса полеотолерантности проведено лихенофлористическое зонирование каждого населенного пункта. Ил.2. Табл.3. Библ. 22

Ключевые слова: ЛИХЕНОБИОТА; ГОРОДА-СПУТНИКИ; НОВОСИБИРСК;

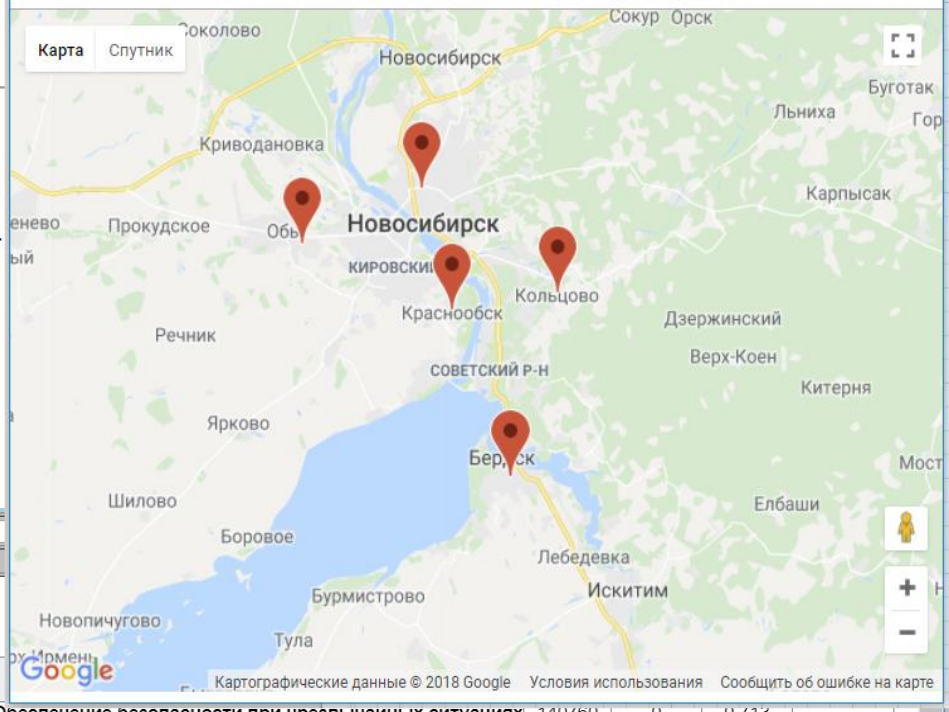
Включено в: Растит. мир Азиатской России. (ISSN 1995-2449). - 2008. - N 2. - С. 33-40.

Закрыть | Раскрыть сокращения | Выделить названия | Убрать многозначно

1496421 1 24	Обь
1496747 1 13,22,66	Новосибирск
1502091 1 25	Краснообск
1502847 1 27	Кольцово
1510350 1 23	Бердск
2017370 1 72	Россия

db4.sbras.ru/ZooSPACE/gw/gis/g1.htm?s1=54.9872,82.716...

Не защищено | db4.sbras.ru/ZooSPACE/gw/gis/g1.htm?s1=54.9872,...



Искать: Краснообск

Идентификатор					
EX	**РЖ Обеспечение безопасности при чрезвычайных ситуациях**	140760	0	0.713	
GG	**РЖ Геология и геофизика**	1564490	1	2.096	SRU
GR	**РЖ География**	418474	5	1.17	SRU
OC	**РЖ Охрана окружающей среды**	744806	0	1.634	

Запись: 5 из 5 **Представление:** Обычное **Формат:** XML **Схема:** F << < > >>

Коды: **ГРНТИ:** 39.19.25

Авторы: Романова Е.В.

Заглавие: Лишайники городов-спутников г. Новосибирска

Реферат: Исследована лихенофлора четырех населенных пунктов в окрестностях г. Новосибирска: г. Бердск, г. Обь, пос. Краснообск и Наукоград Кольцово. Выявлено 79 видов лишайников из 18 семейств и 38 родов, проведены таксономический, географический и экологический анализы. На основе индекса полеотолерантности проведено лихенофлористическое зонирование каждого населенного пункта. Ил.2. Табл.3. Библ. 22

Ключевые слова: ЛИХЕНОБИОТА; ГОРОДА-СПУТНИКИ; НОВОСИБИРСК;

Включено в: Растит. мир Азиатской России. (ISSN 1995-2449). - 2008. - N 2. - С. 33-40.

Выбор источников данных | Простой | Расширенный | Эксперт | Карта | ГТ | Начать с: 1 порций 1

Выводы

- БД документов, содержащая статьи из журналов, газет, сборников, однотомные и многотомные издания с рефератами, в которых имеются географические термины, выделенные из текстов экспертом, позволят отладить алгоритмы индексации, разрабатываемые в других блоках проекта.

Географический тезаурус будет использоваться при поиске информации для выбора названий объектов, независимо от их переименования и получения геометрической информации об объектах с их расположением на географической карте.

В 2020 году планируется закончить разработку системы создания, ведения тезауруса и использования его для поиска в документных базах данных.

Спасибо за внимание!

Докладчик: Баженов Сергей Романович
E-mail: bazhenov@spsl.nsc.ru
www.spsl.nsc.ru