Многоязычный доступ к данным на основе тезауруса географических названий♣

©Лаврёнова О.А.

Российская государственная библиотека
lavr@rsl.ru

Аннотация

Доклад представляет собой описание проекта тезауруса географических названий, формируемого в виде национального нормативного файла как формы представления данных, принятой в автоматизированных библиотечных системах.

Можно выделить следующие наиболее существенные характеристики данного тезауруса**:**

надежность источников информации;

экономичность процессов формирования словаря, реализуемых преимущественно программно;

высокие темпы пополнения базы данных и высокая полнота сведений о географических объектах России;

учет смысловых связей между географическими названиями;

многоязычность;

открытый удаленный доступ.

Российский национальный нормативный файл географических названий получил международный код *rugeo* в системе форматов MARC21.

1 Основные функции тезауруса

Тезаурус географических названий создается в интересах пользователей электронных каталогов, электронных библиотек и других автоматизированных информационных систем (АИС). Он предназначен для поддержки многоязычного доступа к данным с использованием контролируемых точек доступа, представляющих собой названия географических объектов.

Тезаурус **должен выполнять следующие функции** [2]**:**

обеспечивать использование нормализованных наименований географических объектов, прежде всего российских, в географических заголовках библиографических записей (БЗ), предметных рубриках, классификациях, рубрикаторах и других типах метаданных для автоматического контроля их использования и расширения возможностей поиска документов;

создавать условия для централизованного хранения данных о названиях географических объектов, а также возможность получения в свободном удаленном доступе надежных сведений о принятых и не принятых на данный момент времени географических названиях (например, сокращенных, устаревших);

фиксировать отношения условной эквивалентности, иерархические и ассоциативные связи между географическими названиями, расширяя сферу поиска данных;

обеспечивать возможность обмена сведениями о географических названиях с отечественными и зарубежными библиотеками и другими организациями, библиотечными сетями и объединениями.

Это означает, что географический объект на основе этого словаря можно найти не только по его принятому названию, но и по не принятым на настоящий момент названиям. Кроме того, для конкретных регионов можно найти входящие в него географические объекты и т.д.

В настоящее время эти функции реализуются на основе данных о тех регионах РФ, которые отражены в справочнике географических названий электронного каталога (ЭК) РГБ [4]**.** Сведения об остальных регионах России постепенно загружаются в ЭК по мере их обработки в автоматизированном Государственном каталоге географических названий (АГКГН)..

2 Определения

В проекте в целом решено использовать определения основных понятий, данные **в Федеральном законе** Российской Федерации **«О наименованиях географических объектов»** [4]:

- «географические объекты — существующие или существовавшие относительно устойчивые, характеризующиеся определенным местоположением целостные образования Земли: материки, океаны, моря, заливы, проливы, острова, горы, реки, озера, ледники, пустыни и иные

Труды 9ой Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL'2007, Переславль-Залесский, Россия, 2007.

природные объекты; республики, края, области, города федерального значения, автономная область, автономные округа; города и другие поселения, районы, волости, железнодорожные станции, морские порты, аэропорты и подобные им объекты;

наименования географических объектов — географические названия, которые присваиваются географическим объектам и служат для их отличия и распознавания;

установление наименований географических объектов — выявление существующих наименований географических объектов, присвоение наименований географическим объектам и переименование географических объектов;

нормализация наименования географического объекта — выбор наиболее употребляемого наименования географического объекта и определение написания данного наименования на языке, на котором оно употребляется» (Федеральный закон, ст.1 [4]).

3. Развитие проекта

Данный проект РГБ ведет с 2003 года, в частности, при финансовой поддержке Минкультуры и Федерального агентства по культуре и кинематографии.

До **2007 года тезаурус формировался только для географических названий России** по технологии, опирающейся, в первую очередь, на использование данных из автоматизированного Государственного каталога географических названий. При этом в 2006 г. начат эксперимент по дополнению тезауруса названиями на государственных языках субъектов РФ в оригинальной графике.

Особенность **нового этапа развития проекта** - обработка нормативных файлов из зарубежных информационно-библиотечных систем, которая производится в двух направлениях:

внесение в словарные статьи российских географических названий их иностранных аналогов с установлением связей с записями в зарубежных электронных каталогах (ЭК);

заимствование данных о названиях зарубежных географических объектов с переводом сведений на русский язык, вводом в тезаурус русских аналогов соответствующих зарубежных названий и связей с их написанием в иной графике, в частности – оригинальной.

Такая стратегия привлечения внешних нормативных данных существенно сокращает трудоемкость и, естественно, стоимость процессов создания тезауруса.

4. Структура тезауруса

Что касается российских географических названий, то в тезаурус включаются [1, 2] записи для:

наименований географических объектов Российской Федерации (например, *Москва, Красноярский край, Волга);*

континентального шельфа (например, *Магадан-1, участок примагаданского шельфа Охотского моря; Одопту, нефте-газовое месторождение на северо-восточном шельфе острова Сахалин*)

и исключительной экономической зоны Российской Федерации (например, *Чукотская исключительная экономическая зона*),

а также географических объектов, открытых или выделенных российскими исследователями в пределах открытого моря и Антарктики (например, *Остров Петра Первого в пределах Антарктики*),

если иное не предусмотрено международными договорами Российской Федерации (Федеральный закон, ст.3, п.1 [3]).

Федеральный закон Российской Федерации «О наименованиях географических объектов» устанавливает правовые основы деятельности в области присвоения наименований географическим объектам и переименования географических объектов, а также нормализации, употребления, регистрации, учета и сохранения наименований географических объектов как составной части исторического и культурного наследия народов Российской Федерации. Федеральный закон (ст.8, п.2) регламентирует следующее: «В документах, картографических, иных изданиях на русском языке или на других языках народов Российской Федерации употребляются нормализованные наименования географических объектов».

В связи с этим в РГБ принято следующее **основное требование к нормативных записям для названий российских географических объектов**: нормализованные (принятые) заголовки нормативных записей должны соответствовать нормализованным заголовкам, установленным на федеральном уровне и приведенным в «**Государственном каталоге географических названий России (ГКГН)».**

В качестве средства структурирования словарных статей тезауруса используется принятый в библиотечной сфере формат **MARC21 для нормативных данных**. Каждая словарная статья формируется в виде нормативной записи. Таким образом, данный тезаурус рассматривается в библиотечной сфере как национальный нормативный/авторитетный файл (Authority file) географических названий [2].

При наличии соответствующих данных **нормативная запись включает следующие сведения:**

нормализованное наименование географического объекта;

не принятые на текущий момент наименования географических объектов (иные варианты названий, сокращенные названия);

географические названия на других национальных языках субъектов РФ и иностранных языках в различных алфавитах;

географические координаты;

регистрационный номер географического названия в Государственном каталоге географических названий, который сохраняется даже при изменении названия (для объектов РФ);

источник установления нормализованного названия;

дата (или только год) формального установления нормализованного названия;

род географического объекта;

административный статус населенного пункта;

административно-территориальная привязка географического объекта (наименование субъекта РФ, наименование административного района);

соответствующий индекс ББК (территориальное типовое деление);

примечания различного рода.

Нормализованное **наименование географического объекта** помещается в неповторяющееся поле *151* нормативной записи и, тем самым, **объявляется именем дескриптора** как класса условно эквивалентных понятий, причем **аскрипторы** (не принятые на данный момент названия объекта) помещаются в повторяющееся поле *451*.

В примере (1), приведенном в приложении:
Мурмозеро (юго-западнее оз.Ларинское) – дескриптор,
Мурмозеро (Корвальское), Нурмозеро, Муромозеро, Нурм-озеро – аскрипторы.

Административно-**территориальная привязка географического объекта**, прежде всего, задается вводом в поле *551* нормативной записи названия территории, на которой расположен данный географический объект. Этот элемент записи представляет собой дескриптор, стоящий в иерархическом дереве тезауруса непосредственно выше данного. Дескриптор может входить в несколько деревьев, как это полагается в полииерархическом тезаурусе.

В примере (1), приведенном в приложении:
Ленинградская, область (Россия). Природные объекты – дескриптор, стоящий в иерархическом дереве непосредственно выше дескриптора *Мурмозеро (юго-западнее оз.Ларинское)*.

Наличие в поле *551* записей для всех дескрипторов, стоящих непосредственно ниже по иерархии, чем данное, является указанием для программного обеспечения вывести при поиске по данному дескриптору или его условным синонимам (аскрипторам) всех дескрипторов, расположенных на уровень ниже данного в иерархическом дереве (см. пример 5 в приложении).

Формулировка **дескриптора имеет сложную структуру**: непосредственно название объекта (*Ленинградская*) + род географического объекта *(область)* через запятую + уточнение географического названия в скобках: *Ленинградская, область (Россия), Нерль, река (впадает в Угличское вдхр.), Нерль, река (левый приток р. Клязьма).*

Уточняющие пометы в скобках необходимы, поскольку многие географические названия повторяются в различных и даже в одних и тех же регионах (например, две реки с названием *Нерль* можно найти на карте Ярославской области), Можно выделить следующие типы помет:

название того региона, к которому географический объект относится, например: *Москва, город (Россия), Москва, город (США);*

иерархическая последовательность названий тех регионов, к которым географический объект относится, например: *Переславль-Залесский, город (Россия, Ярославская область, Переславский район);*

уточнение расположения объекта относительно других, более крупных или известных объектов, например: *Нерль, река (впадает в Угличское вдхр.), Нерль, река (левый приток р. Клязьма), Мурмозеро (юго-западнее оз.Ларинское).*

5. Источники данных

Используются следующие **источники данных о названиях географических объектов РФ** [1, 2]:

автоматизированный Государственный каталог географических названий России (АГКГН) - фрагменты в виде файлов;

Общероссийский классификатор объектов административно-территориального деления (ОКАТО);

библиографические записи, формируемые при каталогизации документов в РГБ, а затем и в других библиотеках;

нормативные записи из электронных каталогов других библиотек;

справочный аппарат каталогов РГБ;

справочные издания.

АГКГН**,** который создается **Центральным институтом геодезии, аэросъемки и картографии (ЦНИИГАиК),** выбран в качестве основного источника данных, так как он представляет собой первоисточник сведений и абсолютно надежен, но поскольку он только создается, данные приобретаются РГБ для тезауруса по мере формирования данных в АГКГН. В этой базе данных регистрируется нормализованное название объекта, **полученное обязательно из официального источника**, который также фиксируется в базе данных. Из **АГКГН** РГБ получает все основные сведения, которые нужны для тезауруса, а именно: нормализованное название, источник установления нормализованного названия и дата его, род географического объекта, сведения об административно-территориальной привязке географического объекта, географические координаты объекта; варианты названий, источники установления варианта названия и т.д. Все данные представлены в АГКГН в виде специальных системных таблиц. Специалисты РГБ конвертируют данные непосредственно из этих таблиц в формат MARC21 for Authority Data.

То обстоятельство, что основной массив данных в нормативном файле создается автоматически на основе надежных источников при минимальных затратах труда сотрудников библиотеки, является **ключевым моментом в данном проекте**. Программное обеспечение, разработанное заведующей сектором РГБ Аветисовой Т.В., позволяет автоматически сформировать правильные нормативные записи в формате MARC21, провести их формальный контроль, а затем – преобразовать в их в формат RUSMARC для передачи в Сводный электронный каталога библиотек России (СКБР) сети ЛИБНЕТ [2].

6. Доступ

Тезаурус **географических названий установлен в открытом доступе** на **Web-сайте РГБ** в составе системы АЛЕФ как общедоступный электронный ресурс и одновременно как средство поиска в электронном каталоге Библиотеки [4], а также может быть использован для других задач обработки информации. Файл постоянно пополняется. На 1 мая 2007 г. получены данные о 139000 географических объектов (58 регионов РФ) и загружены в ЭК РГБ данные более чем о 80000 географических объектах.

После входа в систему АЛЕФ, на основе которой работает ЭК РГБ, следует выбрать в верхнем меню функцию «Базы данных», затем в списке найти раздел «Справочники» и выбрать «Географические названия». Поиск ведется по произвольным словам, сочетаниям слов и их частей с учетом иерархических и ассоциативных связей между лексическими единицами тезауруса. Каждую нормативную запись как результат поиска можно вывести на экран в краткой (стандартной) форме, полной форме с наименованиями типов элементов данных и в формате MARC 21/Authorities/.

Названия географических объектов на иностранных языках и национальных языках субъектов федерации, отличных от русского, в настоящее время в основную базу данных тезауруса не загружены.

7. Эксперименты с данными на других языках и в иной графике

Первый **эксперимент по работе с зарубежными нормативными файлами** географических названий проводится в порядке научного сотрудничества с Библиотекой Конгресса США (БК), которая предоставила возможность РГБ скачать все свои нормативные записи, содержащие географические названия (более 150 тысяч). РГБ проводит исследование соответствия файлов и разрабатывает автоматизированные методы создания собственного многоязычного файла с использованием сторонних данных, а также методы обеспечения сетевого взаимодействия при обращении к файлам различных библиотек. Подготовленные методы и выводы предполагается использовать в проекте Виртуального международного нормативного файла (VIAF) и при организации сотрудничества по разработке других видов многоязычных нормативных файлов в РГБ (например, имен лиц, наименований организаций).

Проект широко обсуждался на различных конференциях и получил поддержку на Конгрессе ИФЛА-2005, где был представлен в форме доклада [1].

Одна из проблем работы с многоязычными данными заключается в следующем: географические имена в различных языках могут представлять собой как формы транслитерации, построенные по разным стандартам, так и формы транскрипции в соответствии с неизвестными правилами, формы, полученные в результате перевода на другой язык, смешанные формы, варианты, сложившиеся различными путями исторически. Кроме того, существуют специальные латинские формы русских географических названий, созданные отечественными географическими агентствами для российских карт в латинской графике.

Примеры [2]**:**

Россия - *Rossiya* (транслитерация по ISO 9-95), *Russland* (вариант в немецком языке), *Russia* (итальянский вариант), *Russia* (вариант в английском языке), *Rossija* (специальная форма, установленная географическим агентством для российских карт в латинской графике);

Шупашкар - название города на чувашском языке в оригинальной графике, *Чебоксары* - русский вариант названия.

Приложение. Примеры нормативных записей из файла географических названий

Примечания:

записи приводятся в форматах MARC21 и RUSMARC в сокращенном виде – без формальных данных;

по сравнению с записями в ЭК РГБ, в примерах примечания из поля *678* переведены в поле *370*, как это планируется сделать в ближайшее время.

(1) Пример записи в формате MARC21

151## |аМурмозеро (юго-западнее оз.Ларинское)
451## |wa |аМурмозеро (Корвальское)
451## |wa |аНурмозеро
451## |wa |аМуромозеро
451## |wa |аНурм-озеро
551## |wg |аЛенинградская, область (Россия). Природные объекты
670## |а АГКГН |b Мурмозеро (Корвальское) (31.12.2000) |b Нурмозеро (31.12.1936) |b Муромозеро (31.12.1988) |b Нурм-озеро (31.12.1941) |b Мурмозеро (31.12.1967)
670## |аАГКГН (ЦНИИГАиК)

670## |aР-36-143,144/1967
670## |aР-36-143,144/1980,2000
670## |a10-верстная карта
670## |aКарты поверхностных вод Европейской части СССР, 1936
670## |aР-36-В,Г/1988
670## |aР-36-143,144/1941
670## |aР-35,36/1943
670## |aАдминистративная карта Ленинградской области, 1945
670## |aАдминистративная карта Ленинградской области, 1986
670## |aОтвет Бокситогорского районного отдела народного образования, 1991
670## |aОтвет начальника почтового отделения связи н.п. Красный Бор от 25/10/ 1991
670## |aСловарь названий гидрографических объектов России и других стран-члено в СНГ, 1999.
680## |iозеро
680## |iюго-западнее оз.Ларинское
680## |i60 град. 6 мин. С.Ш., 35 град. 1 мин. В.Д.

(2) Пример записи в формате MARC21

151 |a Нерль, река (впадает в Угличское вдхр.)
451 |w a |a Нерль (Векса Плещеевская), река
451 |w a |a Нерль Волжская, река
670 |a АГКГН |b Нерль (Векса Плещеевская) (31.12.1966) |b Нерль Волжская (31.12.1936) |b Нерль (25.2.1946)
680 |i 57 град. 6 мин. С.Ш., 37 град. 40 мин. В.Д.
551 |a Ярославская, область (Россия). Природные объекты |w g
551 |a Тверская, область (Россия). Природные объекты |w g

(3) Пример записи в формате MARC21

151 |a Нерль, река (левый приток р. Клязьма)
451 |w a |a Нерль Клязьминская, река
670 |a АГКГН |b Нерль Клязьминская (31.12.1936) |b Нерль (9.2.1945)
670 |a АГКГН (ЦНИИГАиК)
680 |i 56 град. 11 мин. С.Ш., 40 град. 44 мин. В.Д.
551 |a Ярославская, область (Россия). Природные объекты |w g
551 |a Ивановская, область (Россия). Природные объекты |w g
551 |a Владимирская, область (Россия). Природные объекты |w

(4) Пример записи в форме вывода на экран в формате MARC21 (система АЛЕФ)

151	|a Сергиев Посад, город (Россия, Московская область, Сергиево-Посадский район)
451	|w a |a Сергиево-Посад, город
451	|w a |a Загорск, город
451	|w a |a Сергиев, город
451	|w a |a Сергиевский Посад, город
451	|w a |a Святотроицкая Сергиева Лавра и Посад, город
670	|a АГКГН |*b* Сергиево-Посад (31.12.1999) |*b* Загорск (31.12.1939) |*b* Сергиев (31.12.1926) |*b* Сергиевский Посад (31.12.1899) |*b* Святотроицкая Сергиева Лавра и Посад (31.12.1804) |*b* Сергиев Посад (21.9.1991)
670	|a Указ Президиума Верховного Совета РСФСР № 1675 от 21/09/1991
670	|a Справочник административно-территориального деления Московской области, стр. 9, 1999
670	|a Материалы переписи населения, 1939
670	|a Материалы переписи населения, 1926
670	|a 10-верстная карта
670	|a Атлас Российской Империи, 1804
680	|i 56 град. 18 мин. С.Ш., 38 град. 8 мин. В.Д.
551	|a Сергиево-Посадский, район (Россия, Московская область) |w g

(5) Пример фрагмента вывода при поиске в системе АЛЕФ следующего (вниз) уровня иерархии для дескриптора «*Ярославская, область (Россия)»*

Заголовок	Ярославская, область (Россия)
Прим.польз	область субъект федерации центр - город Ярославль
Шире	Россия
Уже	Большесельский, район (Россия, Ярославская область)
Уже	Борисоглебский, район (Россия, Ярославская область)
Уже	Брейтовский, район (Россия, Ярославская область)
Уже	Гаврилов-Ямский, район (Россия, Ярославская область)
Уже	Даниловский, район (Россия, Ярославская область)
Уже	Любимский, район (Россия, Ярославская область)
Уже	Мышкинский, район (Россия, Ярославская область)
Уже	Некоузский, район (Россия, Ярославская область)
Уже	Некрасовский, район (Россия, Ярославская область)
Уже	Первомайский, район (Россия, Ярославская область)
Уже	Переславский, район (Россия,

	Ярославская область)
Уже	Пошехонский, район (Россия, Ярославская область)
Уже	Ростовский, район (Россия, Ярославская область)
Уже	Рыбинский, район (Россия, Ярославская область)
Уже	Тутаевский, район (Россия, Ярославская область)
Уже	Угличский, район (Россия, Ярославская область)
Уже	Ярославский, район (Россия, Ярославская область)
Уже	Дарвинский, заповедник (Россия, Ярославская область)

(6) Пример записи, преобразованной в формат RUSMARC

#215 \aАцинский, хребет\z(31.12.1976 -)
#3001\aАсиновский (31.12.1966 - 31.12.1976), Асинский (31.12.1938 - 31.12.1966)
#3301\ахребет\aЧитинская область
#3301\алевобережье р. Чикой
#3301\a50 град. 8 мин. С.Ш., 109 град. 30 мин. В.Д
.#415 \aАсиновский, хребет\5a
#415 \aАсинский, хребет\5a
#515 \aЧитинская область \5g

Литература

[1] Лавренова О.А. Национальный нормативный файл географических названий России // Международный библиотечный и информационный конгресс: 71-я генеральная конференция ИФЛА. 14-18 августа 2005. Осло, Норвегия. - Материалы. – Русская версия: http://www.ifla.org/IV/ifla71/papers/015r_trans-Lavrenova.pdf.
Английская версия: http://www.ifla.org/IV/ifla71/papers/015e-Lavrenova.pdf

[2] Лавренова О.А. Национальный файл географических названий - новый проект РГБ // Библиотековедение. – 2006. - №2. – С. 46-53

[3] Российская Федерация. Федеральный закон от 18 декабря 1997 г. № 152-ФЗ. «О наименованиях географических объектов»

[4] Файл нормативных/авторитетных записей географических названий России в электронном каталоге РГБ: http://aleph.rsl.ru/ (БАЗЫ ДАННЫХ/СПРАВОЧНИКИ/ Географические названия).

The multilingual Access to the Data on the Base of the Geographic Names Thesaurus

Lavrenova Olga
The Russian State Library

The paper deals with the RSL project of the geographic names thesaurus in the form of a national authority file. The most essential characteristics of the thesaurus are:
effectiveness of the generation technology;
good rate of data base growth and completeness of information concerning geographic objects in Russian regions;
taking into account of the semantic relationships between geographic names;
conception of an open multilingual access to the data.
The Russian National Authority File of Geographic Names has received the international code “rugeo” in the MARC21 format system.