

В. А. Друщиц, О. А. Курчавова

Геологический институт Российской Академии Наук

Институт проблем информатики Российской Академии Наук

druchits@geo.tv-sign, zatsman@mail.ru

Аннотация

В статье рассматриваются вопросы специальной подготовки документов, предназначенных для включения в электронные коллекции научных документов. Материалом для исследования послужила электронная коллекция по наукам о Земле Российского Фонда Фундаментальных Исследований. Географические названия анализировались с точки зрения их использования в двухязычных электронных формах.

Установлено, что поиск геопространственных данных может быть значительно затруднен в связи с наличием в документах большого числа географических названий и слов, занимающих промежуточное положение между терминами и географическими названиями.

В документах по наукам о Земле географические названия не обеспечивают точность поиска геопространственных данных. Исследуется явление топонимической омонимии.

Рассматриваются проблемы, связанные с использованием географических названий в двухязычных полнотекстовых научных документах.

Для решения проблемы полнотекстового поиска геопространственной информации и увеличения его точности предлагается использовать метод метаданных, который в настоящее время применяется для поиска картографических данных. Обосновывается необходимость выработки структуры метаданных для различных типов геологических объектов.

Доказывается, что применение метаданных на этапе генерации документов для электронных коллекций позволит существенно усилить коммуникативные качества документа.

Введение

Постоянно увеличивающийся объем цифровой информации в исследованиях в науках о Земле порождает очевидную проблему - проблему поиска требуемой информации в хранилищах данных (I) Фундаментальной научной проблемой является лексический поиск геопространственной информации по именам собственным (топонимам) в электронных коллекциях полнотекстовых документов по наукам о Земле (статьи, монографии, отчеты, диссертации и т. д.). Однако часто научные результаты исследований в науках о Земле выражаются в специфической для этих наук форме - географических и тематических картах. Помимо комплексности отражения сложной геопространственной информации карты отличаются (с точки зрения проблемы поиска) тем, что содержат в себе минимум текстовой информации. При этом текстовая информация является основной при организации поиска в современных электронных коллекциях полнотекстовых научных документов. Текстовое описание геопространственной информации, легко отображаемой и интерпретируемой с помощью карт, наталкивается на существенные трудности.

В первую очередь эти трудности связаны с большим количеством топонимов и терминов для описания пространственных и временных характеристик геообъектов, а также таких явлений, как наличие топонимов с одинаковым написанием (омонимов) и терминов, интерпретация которых для специалистов, использующих картографический материал, не представляет сложностей. В то же время, при анализе текста, а тем более организации автоматического поиска в электронных коллекциях, насчитывающих десятки тысяч документов, возможно появление абсолютно непредсказуемого результата.

Современное состояние исследований

Традиционно, имена собственные, включая топонимы, изучались в рамках специального направления в языкознании - ономастике. К числу основных задач ономастики входит сбор полевого и архивного материала, анализ письменных источников, составление специальных словарей и атласов. Ономастика связана с комплексом гуманитарных наук, а также науками о Земле и Вселенной. Топонимия, как часть ономастики, является вспомогательной дисциплиной для специалистов в различных науках, помогая извлекать из анализируемых имен неязыковую (историческую, географическую и т.д.) информацию (I, II).

В науках о Земле топонимы имеют не только координационное, но и смысловое значение, выступая в качестве основы для названий ключевых моментов времени, литологических и структурных подразделений, форм рельефа и месторождений. В ряде случаев топонимы по значению приближаются к терминам (Девон, Гондвана).

С начала 60х гг. XX в. ономастика и связанные с ней научные методы получают особое развитие как в России, так и в других странах. В России основными центрами ономастических исследований стали Институт языкознания РАН и Институт русского языка РАН. За рубежом активно действуют общества и центры ономастических исследований: the American Name Society, the Centre for English Name Studies, University of Nottingham. Издаются журналы Names (США) и Onoma (Бельгия).

В 1973 г. была опубликована работа А.В. Суперанской "Общая теория имени собственного" М., Наука, которая подвела итог всем предыдущим исследованиям, и в которой исчерпывающе описывались как теоретические, так и прикладные задачи ономастики, а также место ономастики среди других наук. В 1985г. вышла в свет коллективная монография

группы ономастики ИЯ РАН "Теория и методы ономастических исследований", а 1986-1988гг. опубликованы работы, посвященные таким вопросам, как практическая транскрипция собственных имен, функционирование названий в своей и чужой иноязычной среде и стандартизация (III).

В настоящее время работы по ономастике можно разделить на следующие виды: это составление словарей, атласов, энциклопедий, регистрирующих имена собственные отдельного региона, страны или группы стран (IV). Необходимо отметить такую работу как "Русская ономастика и ономастика России" под ред. О.Н. Трубачева (V); исследования, посвященные именам собственным в творчестве отдельных авторов. Эти исследования носят чисто филологический характер, хотя имена собственные, включая топонимы, изучаются в контексте литературного произведения (т.е. в тексте); исследования посвященные функционированию имен собственных применительно к экстралингвистическим факторам (Циммерлинг А. В. "Идентифицирующие имена в социальном пространстве" Проект N 98-06-80385 ИСБ РАН); исследования, посвященные узким группам ономастической лексики: фамилиям, прозвищам и т. д. (по данным журналов Names, Onoma).

С начала 90х гг. для ономастических исследований активно применяются компьютеры. Словари и атласы оформляются в виде баз данных и информационно-поисковых систем: Информационно-поисковая система шт. Орегон (VI), Компьютерная база данных английских географических названий (VII). Поиск в этих системах производится по атрибутам, как правило, с использованием языка SQL.

Топонимика традиционно связана с топографией. Работа по картографии и связанной с ней вопросами топонимики и стандартизации проводится в таких центрах, как ЦНИИГАИК, МГУ им. Ломоносова, ИГ РАН, ГИН РАН, Евро-Азиатском геофизическом обществе, Географическом обществе РАН (VIII).

Крупномасштабным проектом является Alexandria Digital Library Project, работа над которым ведется в университете г. Санта-Барбара (IX). В рамках данного проекта был создан электронный газеттир географических названий, насчитывающий около 6 млн. топонимов. Данный газеттир служит основой создания цифровой геобиблиотеки, где пользователь может устанавливать соответствие между топонимом, соответствующим ему геообъектом, картографическим изображением и координатами геообъекта на земной поверхности. Таким образом, топонимы также рассматриваются изолированно, вне контекста документов.

Использование вычислительной техники для составления и работы с картами привело к созданию гео-информационных систем. Однако, гео-информационные системы не были рассчитаны на задачи лексического поиска информации, так как предусматривают работу с картами, а не с полнотекстовыми документами.

В 1994 г. Федеральный комитет по географическим данным США (FGDC) создает специальную инфраструктуру для коллективного использования геоинформации на основе современных средств телекоммуникаций. Поиск информации ведется с помощью информационных хранилищ геопространственных метаданных на основе протокола ANSI/NISO Z39.50 и стандарта FGDC/ASTM Content Standard for Digital Metadata (X). Таким образом, каждый геопространственный объект описывается с помощью набора дополнительных сведений - метаданных. Тем не менее, данная структура рассчитана на хранение, поиск, и обмен только картографической информацией, а не полнотекстовыми документами.

Необходимо отметить, что основное положение общей теории имен собственных заключается в том, что "... собственные имена - наиболее удобный способ выделения объекта. Иные пути индивидуализации - числовое обозначение, система координат и

описательная фраза [в текстах] имеют ограниченное применение. Использование имен всеобщее, универсально." [2].

Применительно к электронным коллекциям полнотекстовых документов по наукам о Земле этот тезис становится весьма спорным. Хранение десятков тысяч документов и существование в синхронии топонимий (совокупностей всех географических названий) многочисленных регионов, насчитывающей сотни тысяч имен собственных, делает имя (топоним) недостаточным средством идентификации геопространственного объекта и тем самым не обеспечивает точность лексического поиска геопространственной информации. Точность лексического поиска не может быть обеспечена из-за явления топонимической омонимии, т.е. использования одинаковых географических названий для идентификации различных геопространственных объектов.

В задачи данного исследования входит изучение метода поиска геопространственных объектов в электронных коллекциях карт и адаптация его к полнотекстовому поиску. Необходимо отметить следующее: топонимы и ранее могли входить в поисковые предписания при поиске по ключевым полям. При этом оказывалось, что поиск ведется без учета контекста, что неизбежно сказывалось на точности поиска.

Решение проблемы идентификации геопространственных данных и реализация поиска по топонимам в полнотекстовых документах может быть использовано в электронных коллекциях документов по комплексу наук о Земле и Вселенной и гуманитарных наук, везде, где в научных документах топонимы составляют значительный пласт лексики.

Фактический материал и методы его обработки

В целях изучения проблем поиска геопространственной информации был исследован корпус электронных документов по наукам о Земле, входящий в состав электронной коллекции отчетов Российского фонда фундаментальных исследований. Был собран большой фактический материал, содержащий результаты проверки орфографии Форм 501 и 502 отчетов РФФИ (1400 документов) по следующим специальностям: геология, геохимия, геофизика, океанология, горные науки. Работа велась в рамках инициативного проекта "Анализ обоснованности введения новой терминологии в геолого-геофизических моделях" N 98-05-64549" (XI). В процессе анализа ошибок были выявлены две группы слов, имеющих особо важное значение для формирования поисковых запросов при автоматическом поиске геопространственной информации. К первой группе относятся термины, ко второй - топонимы. Было установлено, что при написании отчетов грантодержатели испытывали трудности как при написании известных терминов, зафиксированных в словарях, так и при подборе адекватного нового термина. Много трудностей связано с переводом терминов и терминологических словосочетаний на английский язык, т.е. наличие параллельных текстов на русском и английских языках (XII, XIII).

Сходные проблемы возникают и при употреблении топонимов, особенно тех названий, возникающих при крупномасштабном картировании, так как они обычно не указываются в словарях, справочниках и атласах. Наиболее частые ошибки - это ошибки в орфографии, транслитерации и использование прямой транслитерации вместо английского эквивалента (Beloe More вместо the White Sea). Было также выявлено явление топонимической омонимии (одинаковых названий для идентификации разных геопространственных объектов). Трудности написания топонимов и терминов накладываются на систему "топоним+термин", так как в науках о Земле имеется около 10 классов объектов, атрибутом которых выступает

топоним. Перечисленные проблемы могут серьезно повлиять на точность и полноту полнотекстового поиска.

Географические названия составляют значительную часть словарного состава языка. Эта часть постоянно расширяется, пополняется и изменяется. Частота встречаемости топонима зависит от размера географического объекта, его значимости для данного континента или данного государства, а также его исторического значения. Например, штат Вашингтон, г. Вашингтон, о. Вашингтон, гора Вашингтон.

Многие трудности написания топонимов в документах РФФИ связаны с необходимостью перевода на английский язык. Существуют два основных способа передачи географических названий на другие языки: 1) перенос в оригинальном написании при единстве алфавитов или транслитерация при различии алфавитов; 2) полное заимствование.

Необходимо также принимать во внимание тот факт, что в России документы написаны с использованием кириллицы. При этом топонимы приходится транслитерировать с помощью латинского алфавита. Более того, географические названия, употребляемые народами бывшего СССР, также транслитерируются с помощью кириллицы и латиницы.

Реальность функционирования топонимов гораздо богаче выявленных закономерностей. Рассмотрим ряд примеров (12, XIV, XV, XVI). Типичным примером для функционирования географического названия, несмотря на существующие правила, в русской среде есть несколько вариантов транслитерации топонимов: для горного перевала в Ошской области Чигирчик - Чыйырчик - Чийрчик; Джалтыр - Жалтыр.

Отличаются по написанию в различных источниках географические названия французского происхождения: остров Беннетта - Benetta - Bennetta; поднятие Анжу - Anzhu High - Anjou High. Для России характерно также параллельное существование русского и местного национального названия: Гнилое море - Сиваш; Голодная степь - Бетпак-Дала. Еще одна особенность употребления топонимов в России - это их частое изменение. В

результате в различных источниках могут упоминаться различные названия одного и того же объекта: Самара - Куйбышев - Самара; Кировобад - Пяндж.

Часто встречаются полностью совпадающие в определительной части названия для различных геологических объектов. При неточном переводе могут возникнуть разночтения русской и английской версии текста. Например, Михайловская свита (Ю. Урал) - Михайловская толща (Подмосковный бассейн); Березовская свита (Ю. Урал) - Березовские слои (Тиман) - Березовский горизонт (Ю. Урал); Коркинская свита (Дальний Восток), Коркинский горизонт (г. Красноярск); Лихвинский подъярус - Лихвинская свита - Лихвинское оледенение. Возникают определенные трудности при переводе топонимов в различных областях. Например, в зоне Арктического шельфа России в районе Бренцева моря некоторые структуры морского дна в англоязычном варианте имеют два и более эквивалентов: Медвежинская впадина - Medvezhinskaya depression - Bjornoya Basin - Bear trough. В Восточно- Сибирском море: Благовещенский прогиб - Blagovetshensky deep - Annunciation deep.

В науках о Земле имеются стандартные термины и авторские термины, отражающие личную точку зрения исследователя о природе объекта: грабен Св. Анны - мегабассейн Св. Анны - трог Св. Анны.

Заключение

Одним из способов решения проблемы поиска геопространственной информации является метод метаданных - текстовой информации, сопровождающей специализированные данные, и имеющей целью облегчить автоматизированный поиск геопространственной информации. Традиционно метаданные включают в себя описание содержания, качества, поисковых признаков и других сведений о географической информации.

Таким образом, метаданные устанавливают соответствие между традиционной языковой областью коммуникации и специфическими способами передачи информации в науках о Земле.

Современные информационные технологии позволяют на этапе генерации электронного документа по наукам о Земле создать набор метаданных, однозначно идентифицирующих геопространственную информацию, что позволит сделать электронный полнотекстовый документ самодостаточным, а не привязанным конкретному атласу или карте (как это сделано в бумажных газеттирах). Такой документ не теряет информационных качеств ни при поиске геопространственных данных, ни при передаче его по компьютерным сетям.

Этот подход можно считать новым, так как впервые для решения задач, связанных с текстами по наукам о Земле, будут применяться методы, используемые для идентификации геопространственных данных в геолого-географических картах. Решение этой задачи позволит осуществить основное требование, предъявляемое к электронным коллекциям полнотекстовых научных документов, которые в настоящее время являются самым современным средством хранения и коллективного доступа к геопространственным данным (электронная библиотека по наукам о Земле СО РАН, электронная коллекция заявок и отчетов на дисках типа CD-ROM РФФИ): информационный поиск по свободной лексике языков, использующих кириллицу и латинский алфавит (XVII).

Такая постановка задачи стала возможной во-первых, благодаря успехам в электронной картографии в 90х гг., а во-вторых, благодаря развитию информационных технологий в области реализации полнотекстового поиска по свободной лексике в электронных коллекциях научных документов.

Литература

973. I. Суперанская, А.В., Общая теория имени собственного, М.: Наука, 1973.
974. II. Суперанская, А.В., и др., Теория и методика ономастических исследований, М.: Наука, 1985.
III. Суперанская, А.В., Ономастика, типология, стратиграфия, М.: Наука, 1988.

976. Illustrated Dictionary of Place Names. United States and Canada, Ed. Kelsie B. Harder, NY, 1976.
977. V. Русская ономастика и ономастика России, Под ред. Трубачева О.Н., М.: Школа-пресс, 1994.
978. Lewis L., McArthur, The GNIS and the PC: Two Tools for Today's Toponymic Research, Names (USA), 1995, vol. 43, N 4.

VII. Hough, Carole, A Database for English Placenames, Names (USA), 1995, vol. 43, N 4.

VIII. Суперанская, А.В. Что такое топонимика? - М.: Наука, 1985.

1. Linda L. Hill, James Frew, and Qi Zheng, Geographic Names. The Implementation of a Gazetteer in a Georeferenced Digital library, D-Lib Magazine, 1999, vol. 5, N 1.
2. Evans, John D., Query Language Requirements for a Distributed Geospatial Clearinghouse//W3C Query Languages Workshop Standards\XML - (<http://www.w3.org/tands/QL/QL98/pp/fgdc.html>).
3. XI. Друщиц, В. А., Курчавова О.А., Лингвистические проблемы новой терминологии в науках о Земле и поисковый барьер в электронных библиотеках, Труды международного семинара по компьютерной лингвистике и ее приложениям, ДИАЛОГ'99, Таруса, 1999.

XII. Yining, Yang, Carbonel, Jaime G., et al., Translingual Information Retrieval: Learning from Bilingual Corpora, Artificial Intelligence 103, 1998, pp.323-345.

XIII. Stalter, W.G., Beijert, J.-H., et al., Twenty -One-: Cross-Language Disclosure and Retrieval of Multimedia Documents on Sustainable development, Computer Networks and ISDN Systems 30, 1988, pp. 1237-1248.

XIV. Алексеев, М.Н., Голоднюк, Т.Н., Друщиц, В.А., Русско-английский геологический словарь, М.: Руссо, 1998.

972. Glossary of Geology Ed. Margaret Gary, et al, Washington, DC. 1972.

XVI. The World Atlas. Index-Gazetteer, Ed., Baranov, A.N., Moscow 1968.

XVII. Антопольский, А.Б., Вигурский, К.Б., Электронные библиотеки// Информационные ресурсы России, 1999, N 4.