

Организация ретроспективного тезауруса для использования в задаче географического поиска в «негеографических» информационных системах

Скачков Данил Михайлович

Институт вычислительных технологий СО РАН (Новосибирск), Россия

e-mail: danil.skachkov@gmail.com

Выполнено при частичной поддержке СО РАН (IV.31.1.1, ИП-2012-17, ПИП-2012-73), РФФИ (10-07-00302-а, 12-07-00472-а), Президиума РАН (Проекты 2012-14.3, 2012-15.2), ФЦП шифр номер 2012-1.4-07-514-0022-004

До середины 1960 годов карты являлись всего лишь способом хранения символической информации о географических объектах. 1960-е годы были ознаменованы появлением географических информационных систем или ГИС. ГИС это информационная система, обеспечивающая сбор, хранение, обработку и визуализацию пространственных данных и связанной с ними информации.

Уже тогда было заявлено, что приоритетной задачей картографии является не создание визуальных продуктов, а процессы сбора, преобразования и обработки информации. И основаны эти процессы будут на компьютерных системах [1]. Со временем количество областей, в которых использовались ГИС, увеличивалось. Но, все же, ГИС оставались в сферах, напрямую связанных с географией. Много изменилось с появлением географических интернет сервисов, таких как Google Maps [2]. Они дали возможность интегрировать функциональность географических информационных систем в системы, прямым образом не связанные с географией. 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. Ведь тот факт, что данные системы напрямую не связаны с географией, не означает, что географическая информация там не содержится. Любая статья была где-то написана и опубликована, любой экспонат музея был где-то найден, тексты научных трудов зачастую содержат названия географических объектов. И это только несколько примеров того, что «негеографические» системы на самом деле содержат географическую информацию.

Чтобы иметь возможность использовать эту информацию, мы должны интегрировать географические метаданные в записи целевой системы. Есть несколько вариантов такой интеграции, которые были рассмотрены в [3]. В итоге наиболее оптимальным решением оказался вариант интеграции с использованием тезауруса географических наименований. Однако в ходе решения данной задачи возникает ряд сложностей. Во-первых, географические объекты, на которые ссылаются информационные объекты целевой системы, могут не существовать в настоящее время или иметь иное название и расположение, поэтому нам нужен тезаурус, содержащий исторические данные о географических объектах (так называемый тезаурус ретроспективного геокодирования [4]). Во-вторых, различные географические объекты могут иметь одинаковые названия, что также препятствует правильной индексации. В-третьих, необходимо учитывать взаимное расположение географических

объектов, и, в частности, вариант, когда один географический объект содержит другой географический объект.

В данной работе рассматриваются указанные проблемы и приводится вариант организации ретроспективного тезауруса, а также способ его интеграции в «негеографическую» информационную систему. Также приводится алгоритм индексации существующих данных в информационной системе, подробно описанный в [5]. В заключение рассматривается процесс поиска в проиндексированной информационной системе, и приводится пример результатов поиска при экспериментальной интеграции.

Литература:

1. Abresch J., Hanson A., Heron S., Reehling P. Integrating Geographic Information Systems into Library Services: A Guide for Academic Libraries // <http://elib.sbras.ru:8080/jspui/handle/11055/20000> - ISBN 978-1-59904-726-3
2. Карты Google <http://maps.google.com/>
3. Скачков Д.М., Жижимов О.Л. Об использовании ретроспективного геокодирования для географического поиска в электронных библиотеках // XIII Всероссийская научная конференция «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL39;2011 (Воронеж, Россия, 19.10 - 22.10.2011): Труды конференции. - Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2011. - С.51-58. - ISBN 978-5-9273-1875-9.
4. Жижимов О.Л., Скачков Д.М. О профиле доступа к данным тезауруса для ретроспективного геокодирования и географического поиска в электронных библиотеках // XVIII Международная конференция «Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса»- Крым-2011 (Судак, Украина, 04.06 - 12.06.2011): Материалы конференции. - М.: ГПНТБ России, 2011. - ISBN 978-5-85638-150-3. - Гос. регистр. № 0321100651. - <http://www.gpntb.ru/win/inter-events/crimea2011/disk/059.pdf>
5. Барахнин В. Б., О. Л. Жижимов, А. А. Куперштох, Д. М. Скачков, А. М. Федотов. Алгоритм извлечения из текстовых документов географических названий, отражающих содержание // Вестник Новосибирского государственного университета. Серия: Информационные технологии. Том 10. Выпуск 1. - Новосибирск: Новосибирский государственный университет, 2012. - С.109-120. - ISSN 1818-7900.