Модуль географического поиска информации для платформы ZooSPACE

© Д. М. Скачков © О. Л. Жижимов
Институт вычислительных технологий Сибирского отделения РАН, Новосибирск
danil.skachkov@gmail.com zhizhim@sbras.ru

Аннотация

В статье описывается структура модуля географического поиска информации реализованного на базе платформы ZooSPACE. Информационные системы, подключенные к ZooSPACE являются системами общего назначения, и не содержат географической функциональности. Разработанный модуль позволяет производить географический поиск в таких системах не изменяя их внутреннюю структуру.

Работа выполняется при финансовой поддержке Министерства образования и науки Российской Федерации (грант № 07.514.11.4130) «Разработка принципов и программных средств виртуальной интеграции распределённых источников данных на основе международных стандартов для создания масштабных информационных инфраструктур» шифр «2012-1.4-07-514-0022-004».

История использования географических данных в информационных системах берет начало в 1960 годах. Именно тогда становятся технически возможными и возникают так называемые географические информационные системы или ГИС. Географическая информационная система - это информационная система, обеспечивающая сбор, хранение, обработку и визуализацию пространственных данных и связанной с ними информации. Уже тогда было понятно, что приоритетной задачей картографии является не создание визуальных продуктов, а процессы сбора, преобразования и обработки информации. И фундаментом для этих процессов будут компьютерные системы. С тех пор технологии шагнули далеко вперед и географические данные стали доступны для широкого круга задач. И за счет интернет сервисов, таких как Google Maps™ [1], стало возможным использовать функциональность ГИС в системах, которые для этого не были предназначены изначально. 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***. Но тот факт, что эти системы не были предназначены для работы с географической информацией, еще не говорит, что эта 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. Другое дело, что невозможно эффективно использовать содержащуюся в таких системах географическую информацию. Географическая информация хранится в текстовых полях и пригодна только для простейшего текстового поиска по географическому названию. Результаты такого поиска будут заведомо неточны, поскольку пользователь должен указывать список всех географических объектов из интересующего географического региона. Мало того, что географических объектов огромное количество, названия географических объектов могут изменяться с течением времени, географические объекты могут иметь несколько названий, могут содержаться друг в друге, могут исчезнуть с течением времени. Поэтому, для географического поиска информации в таких информационных системах не обойтись возможностями одного лишь полнотекстового поиска, нужна реализация отдельного механизма, позволяющего скрыть перечисленные сложности от пользователя.

На данный момент активно ведётся разработка платформы ZooSPACE [2] при поддержке Министерства образования и науки Российской Федерации, которая позволит объединять различные информационные системы, в том числе и «негеографические», в единое виртуальное информационное пространство. Разработка модуля,

Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» — RCDL-2013, Ярославль, Россия, 14-17 октября 2013 г.

позволяющего производить географический поиск в пространстве ZooSPACE, улучшит поисковую функциональность платформы.

Задача географического поиска неоднократно обсуждалась как в российских [3-5] так и в зарубежных публикациях [6-10]. Однако не была описана технология реализации географического поиска в существующих информационных системах, которые не имеют географической привязки без изменения целевых систем.

Существующие проекты по реализации географического поиска рассчитаны на то, что записи информационной системы содержат географическую информацию в специально выделенных полях, т.е. изначально привязаны к географическим координатам. Но в большинстве существующих «негеографических» информационных систем такой информации не содержится. И при разработке технологии географического поиска следует учитывать, что изменения в логике работы и структуре данных существующих систем крайне нежелательны и должны быть минимизированы.

Подходы к географическому поиску в информационных системах можно разделить следующим образом:

1. Поиск с помощью атрибутивного и полнотекстового поиска по географическому названию.

2. Поиск с помощью непосредственной индексации записей информационных систем географическими координатами (Рисунок 1).

3. Поиск с помощью метапоисковой машины, использующей:

a. специализированный справочник (тезаурус) географических названий (Рисунок 2);

b. тезаурус географических названий и поисковый географический индекс, организованный с использованием ссылок на записи тезауруса географических названий (Рисунок 3).

Рисунок 1 - Поиск с использованием географического индекса

Метапоисковая машина – поисковая система, не имеющая собственной базы данных и поискового индекса, и формирующая поисковую выдачу из результатов поиска других поисковых систем.

Поскольку «негеографические» информационные системы изначально не были предназначены для обработки географической информации, то в таких системах реализован только первый способ поиска (полнотекстовый). Однако, такому способу поиска присущи определенные недостатки. Поиск чрезвычайно сложен для пользователя, поскольку предполагает составление списка всех географических названий объектов указанного географического региона поиска, причем учитывая историю изменения названий, альтернативные названия, названия на других языках, названия исчезнувших географических объектов. Если же не принимать во внимание все эти особенности географических объектов, то результаты поиска будут неполными.

Использование географического индекса решает проблему трудоёмкости географического поиска, поскольку индекс привязывает записи информационной системы к географическим координатам, тем самым решая проблему неоднозначности географических названий и проблему вложенных географических объектов. Для получения записей, относящихся к определённому географическому региону, достаточно найти записи с координатами, лежащими внутри географического региона. Такой поиск осуществляется методами геометрии а не лексического анализа.

Однако, использовать второй вариант организации географического поиска в «негеографических» информационных системах нецелесообразно, поскольку:

1. Не все хранилища данных, на которых построены существующие информационные системы содержат функциональность по обработке и использованию географических координат.

2. Непосредственное задание координат в метаданных объектов означает, что необходимая поисковая логика должна быть реализована во всех целевых информационных системах по отдельности. Реализация такой логики потребует существенных изменений в структуре целевой информационной системы.

Более целесообразным является вариант организации географического поиска через метапоисковую машину с использованием тезауруса географических названий (третий вариант).

Для организации географического поиска с помощью метапоисковой машины, информационная система должна удовлетворять определённым требованиям. Для двух способов организации поиска требования различны.

Для организации поиска способом 3а) информационная система должна удовлетворять следующим требованиям:

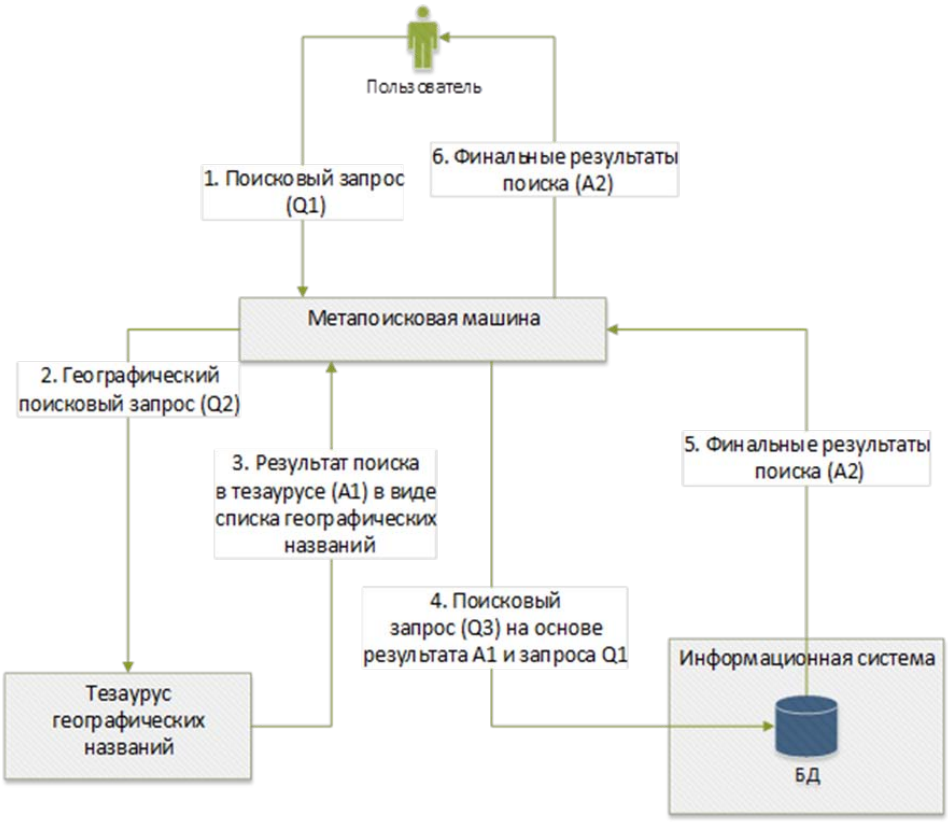

Рисунок 2 - Поиск с использованием тезауруса географических названий

1. должна поддерживать поиск данных по протоколу Z39.50 (обязательный синтаксис запросов RPN-1 [11]);

2. атрибуты, в которых может содержаться географическая информация, должны быть доступны для поиска.

Для организации поиска способом 3b) информационная система должна удовлетворять следующим требованиям:

1. должна поддерживать поиск данных по протоколу Z39.50 (обязательный синтаксис запросов RPN-1);

2. атрибуты, в которых может содержаться географическая информация, должны быть доступны для поиска;

3. идентификаторы записей информационной системы должны быть доступны для поиска;

4. если информационная система содержит внутренний географический индекс, то он также должен быть доступен для поиска.

Как уже было сказано, для реализации поиска с использованием метапоисковой машины необходим тезаурус географических названий.

Определим требования, которым должен соответствовать тезаурус географических названий, чтобы его можно было использовать в задаче географического поиска. При разработке требований к тезаурусу необходимо учесть, что в «негеографических» информационных системах зачастую хранятся данные, относящиеся не к настоящему, а к прошлому. И, имея ввиду, что названия и геометрия географических объектов могут изменяться с течением времени, тезаурус должен содержать информацию об этих изменениях. Так как в одном географическом объекте может содержаться множество других объектов, то тезаурус должен содержать сведения об административном подчинении объектов. Дополнительно, можно определять принадлежность одних географических объектов другим используя данные об их географическом положении. Административное подчинение объектов также может меняться со временем, поэтому тезаурус должен содержать информацию и об изменениях отношений между объектами.

Такой тезаурус был описан в предыдущих работах [12, 13] как тезаурус ретроспективного геокодирования. Тезаурус предоставляет доступ к данным по стандартному протоколу Z39.50. При этом, дополнительно, поддерживаются протоколы HTTP/XML/SOAP/SRW и HTTP/SRU за счет возможностей Z39.50 сервера ZooPARK [14].

Атрибуты, по которым может производиться поиск в тезаурусе, описываются профилем доступа к тезаурусу. Приведем примеры поисковых RPN запросов к тезаурусу в соответствии с профилем доступа.

Пример 1: Найти все географические объекты на территории Новосибирской области во временном промежутке с 12 октября 2001 года по 10 января 2007 года.

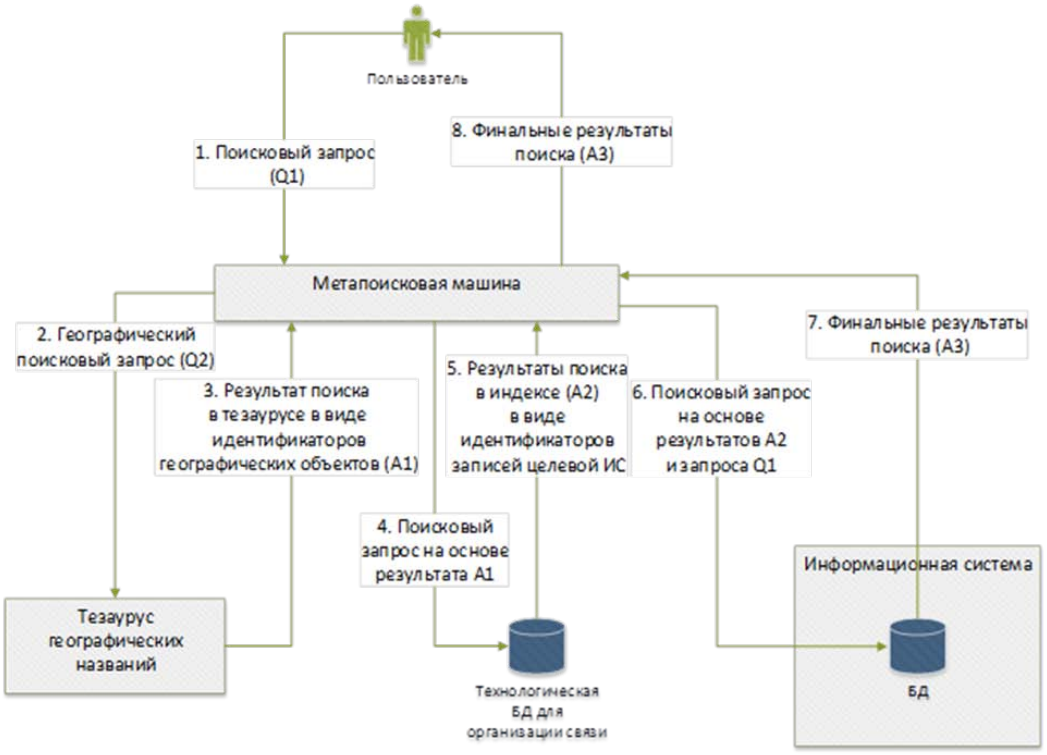

Рисунок 3 - Поиск с использованием тезауруса и географического индекса

@and

@attr 1=59 @attr 2=3 @attr 4=108 {Новосибирская область}

@attr 1=31 @attr 2=16 @attr cip 4=210 {2001-10-12,2007-01-10}

Пример 2: Найти все географические объекты на территории, ограниченной прямоугольником с географическими координатами углов (53.3590, 75.2152) и (57.2273, 85.1248) во временном промежутке с 12 октября 2001 года по 10 января 2007 года.

@and

@attr 1=59 @attr cip 2=7

@attr cip 4=202 {((53.3590,75.2152),(57.2273,85.1248))}

@attr 1=31 @attr cip 2=16 @attr cip 4=210 {2001-10-12,2007-01-10}

Метапоисковая машина используется для реализации географического поиска следующими способами:

1. с помощью специализированного тезауруса географических названий (Рисунок 2);
2. с помощью тезауруса географических названий и поискового географического индекса, организованный с использованием ссылок на записи тезауруса географических названий (Рисунок 3).

Основная идея использования метапоисковой машины в преобразовании запроса, содержащего географическую компоненту, в запрос, не содержащий географическую компоненту, чтобы он мог быть обработан целевой информационной системой. Два способа построения метапоисковой машины, определяют, каким образом будет изменён оригинальный поисковый запрос.

Метапоисковая машина для географического поиска должна отвечать следующим требованиям:

1. должна передавать поисковые запросы без географической компоненты в целевую информационную систему без изменений;
2. должна предоставлять возможность пометки части поискового запроса как географического с возможностью указания, по каким поисковым атрибутам будет производиться поиск;
3. должна уметь обработать географическую часть запроса (согласно способам географического поиска)
4. должна объединять результаты выполнения географической части запроса с частью оригинального запроса, не относящейся к географии, и строить на основе полученного результата новый запрос, который может быть обработан целевой информационной системой.

Описание алгоритма работы метапоисковой машины без использования географического индекса (Рисунок 2):

1. Пользователь направляет поисковый запрос Q1 в поисковую машину.
2. Если в Q1 есть географический запрос, то из запроса Q1 выделяется географический запрос Q2.

3. Если в Q1 не содержится географический запрос, то поиск производится только в целевой информационной системе, без использования тезауруса и возвращаются результаты выполнения запроса Q1.

4. Запрос Q2 направляется в тезаурус ретроспективного геокодирования.

5. Результатом поиска A1 в тезаурусе является список названий географических объектов.

6. Производится генерация различных словоформ географических названий из результатов A1, в результате создаётся список S1.

7. Поисковая машина заменяет географическую часть в запросе Q1 на поисковый запрос (группу поисковых запросов) со списком географических названий S1 по тем-же атрибутам, что и в географическом запросе. Новый запрос обозначен Q3.

8. Запрос Q3 выполняется на целевой информационной системе.

9. В результате выполнения запроса Q3 поисковая машина получает финальные результаты A2.

10. Результаты A2 возвращаются клиенту.

Под словоформами в шаге алгоритма 6 подразумеваются географические названия в различных падежах русского языка. Данный шаг нужен, поскольку в записях информационной системы названия объектов могут встречаться в различных падежах. Словоформы географических названий генерируются отдельным модулем [15].

Описанная технология была экспериментально реализована в качестве модуля платформы ZooSPACE. ZooSPACE является технологической платформой массовой интеграции распределённых гетерогенных источников данных, поддерживающей создание и функционирование широкомасштабных информационных инфраструктур на основе подхода виртуальной интеграции данных [2].

Прежде, чем описывать реализацию, сформулируем некоторые предварительные требования для пользовательских интерфейсов и структуре запросов:

1. Запрос на поиск должен формулироваться в синтаксисе RPN-1 в текстовом представлении PQF. Это необходимо для обеспечения общей интероперабельности системы в целом.

2. Запрос должен содержать часть, относящуюся к географическому (геометрическому) поиску в соответствии с правилами, обсуждавшимися выше.

3. Запрос должен содержать инструктивную часть, предписывающую включение требуемого тезауруса для промежуточной обработки запроса метапоисковой машиной.

4. Запрос должен содержать целевые поисковые атрибуты для поиска в целевой базе данных.

5. Все части запроса должны быть локализованы в единичном структурном блоке запроса RPN (простом или составном) для возможности включения в более сложные запросы.

6. Разбор и исполнение запроса не должны приводить к нештатным ситуациям для поисковых машин, не поддерживающих описываемую технологию.

Эти требования могут быть реализованы множеством способов, для платформы ZooSPACE был выбран способ без использования технологической базы данных.

1. Единичный структурный блок запроса представляет собой объединение двух операндов (APT в терминах RPN) при помощи оператора OR (ИЛИ).
2. Первый операнд содержит геометрическую часть со всеми поисковыми атрибутами, например

@attr 1=2060 @attr 2=7 @attr 4=202 {((52.2, 80.3), (53.8, 81.0))}

- найти все прямоугольные объекты, попадающие в область, ограниченную указанным термом.

3. Второй операнд, например,

@attr 1=4 @attr 6=10 {**geo_module.py**}

указывает на необходимость исполнения инструкции с именем {**geo_module.py**}, которая должна сформировать список географических объектов в соответствии с первым операндом, и на целевой поисковый атрибут @attr 1=4 (Title). Конструкция @attr 6=10 маркирует часть запроса как географическую.

4. Операнды объединяются оператором @or (OR)

Запрос в поисковую машину должен приходить в следующем виде:

@or

@attr 1=2060 @attr 2=7 @attr 4=202 {((52.2, 80.3), (53.8, 81.0))}

@attr 1=4 @attr 6=10 {**geo_module.py**}

Поисковая логика выделяет географическую часть запроса и передаёт его в качестве параметра в модуль, указанный в качестве терма второго операнда (geo_module.py). Модуль производит обращение к тезаурусу, и возвращает список географических названий:

Новосибирск; Новосибирска; Новосибирску; Новосибирском; Новосибирске; Луневский; Луково; Совхоз Луговской; Луговая; Станция Ложок; Локти; Логовой; Льниха; Лиственый; …

На основе списка географических названий формируется запрос к целевой базе данных. Первая часть географического запроса отбрасывается, а вместо названия модуля вставляются географические названия. Части запроса соединяются логической операцией AND:

@and @and @and @and …

@attr 1=4 { Новосибирск }

@attr 1=4 { Новосибирска }

@attr 1=4 { Новосибирску }

@attr 1=4 { Новосибирском }

@attr 1=4 { Новосибирске }

@attr 1=4 { Луневский }

@attr 1=4 { Совхоз Луговской }

@attr 1=4 { Луговая }

…

При этом результирующий запрос имеет стандартную форму и может быть исполнен любым сервером Z39.50 или SRW/SRU.

Для формирования подобных запросов в ZooSPACE предусмотрены графические WEB интерфейсы, изображенные на Рисунке 4.

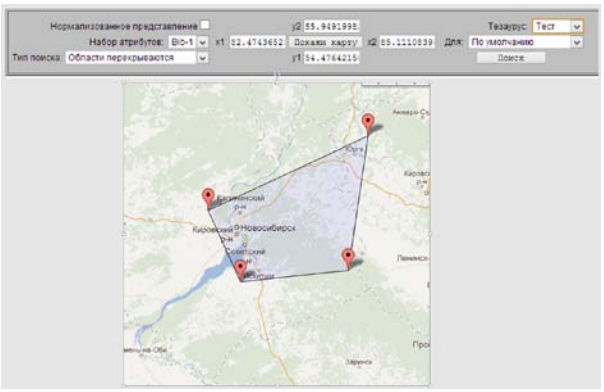

Рисунок 4 Интерфейс построения географического запроса в ZooSPACE

Описанный и реализованный в модулях платформы ZooSPACE подход в настоящее время проходит тестирование в реальной информационной системе.

В таблице 1 представлен фрагмент результатов поиска.

Таблица 1 Фрагмент результатов поиска

Новосибирскому заводу "Труд" - 100 лет
Новосибирский мегаполис
Влияние концентрации аэрозолей на качество атмосферы в г. Новосибирске
Новосибирский ученый удостоен «Глобальной энергии»
Кемеровская область заключила соглашение о создании на ее территории индустрии венчурного инвестирования и развития инновационных отраслей экономики
Кемеровская область занимает второе место по объему инвестиций в Сибирском федеральном округе
….

Описанные выше подходы к решению проблемы географического поиска в "негеографических" информационных системах лишь приближают нас к намеченной цели. Они не в состоянии решить проблему в целом, но, тем не менее, позволяют существенно расширить функциональность поисковых интерфейсов информационных систем и тем самым предоставить пользователю новые возможности по поиску информации.

Литература

[1] Карты Google [Электронный ресурс] URL: http://maps.google.com/

[2] Жижимов О.Л., Никульцев Н.С., Никульцева Е.В., Федотов А.М., Шокин Ю.И. Технологическая платформа интеграции разнородных распределенных данных ZooSPACE // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 20 междунар. конф. «Крым 2013». Судак, Украина. 2013.

[3] GeoNetwork [Электронный ресурс] URL: http://geonetwork-opensource.org/

[4] Атаева О.М., Кузнецов К.А., Серебряков В.А., Филиппов В.И. Портал интеграции пространственных данных "ГеоМета" // Труды XXII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL’2010. Казань. 2010. С. 219-224.

[5] Атаева О.М., Каленкова А.А., Серебряков В.А. MultiMeta - Система интеграции пространственных данных и ресурсов электронных библиотек // Труды XXIII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2011. Воронеж. 2011. С. 26-29.

[6] Tochtermann K. et al. Using semantic, geographical, and temporal relationships to enhance search and retrieval in digital catalogs //Research and Advanced Technology for Digital Libraries. – Springer Berlin Heidelberg, 1997. – С. 73-86.

[7] Hill L. L., Zheng Q. Indirect Geospatial Referencing Through Place Names in the Digital Library: Alexandria Digital Library Experience with the Developing and Implementing Gazetteers: Analysis and Preliminary Evaluation of the Classical Digital Library Model //Proceedings of the Annual Meeting-American Society for Information Science. – Information Today; 1998, 1999. – Т. 36. – С. 57-69.

[8] Volz R., Kleb J., Mueller W. Towards Ontology-based Disambiguation of Geographical Identifiers //I3. – 2007.

[9] Martins B., Silva M. J., Chaves M. S. Challenges and Resources for Evaluating Geographical IR. – 2005.

[10] Jones C. B. et al. Spatial information retrieval and geographical ontologies an overview of the SPIRIT project //Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. – ACM, 2002. – С. 387-388.

[11] Жижимов О.Л., Мазов Н.А. Применение протокола Z39.50 для построения распределенных информационных систем // IV рабочее совещание по электронным публикациям-EL-PUB-99. Новосибирск, Академгородок. 1999. С. 21-23.

[12] 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" - RCDL'2011 (Воронеж, Россия, 19.10 - 22.10.2011): Труды конференции. Воронеж. 2011. С. 51-58.

[13] Скачков Д.М., Жижимов О.Л. Об интеграции географических метаданных посредством ретроспективного тезауруса // Информатика и ее применения. 2012. № 3.

[14] Жижимов О.Л., Мазов Н.А. Сервер ZooPARK: вчера, сегодня, завтра // Библиотеки и информационные ресурсы в современном мире науки, культуры, образования и бизнеса: 14–я междунар. конф. «Крым 2007». Судак, Украина. 2007. С. 168-171.

[15] Барахнин В.Б., Жижимов О.Л., Куперштох А.А., Скачков Д.М., Федотов А.М. Алгоритм извлечения из текстовых документов географических названий, отражающих содержание // Вестник НГУ. 2012. Т. 10. № 1. С. 109-120.

Geographical information search module for the ZooSPACE platform

Danil M. Skachkov, Oleg L. Zhizhimov

The article describes the structure of geographic information retrieval module implemented on the ZooSPACE platform. Information systems connected to ZooSPACE are a general-purpose systems, and do not contain geographic features. The developed module allows geographical search in these systems without changing their internal structure.