

Основанный на онтологии подход к интеграции геоинформационной системы с коллекциями электронных научных публикаций

© В.Т. Вдовицын

Институт прикладных математических
исследований
Карельского научного центра РАН
Петрозаводск
vdov@krc.karelia.ru

© А.К. Полин

Институт геологии
Карельского научного центра РАН

Аннотация

Предлагается онтологически-ориентированный подход к интеграции геоинформационной системы с коллекциями электронных публикаций. В основу подхода положена идея формирования для каждой публикации «географического» индекса, который дополняет ее «текстовой» индекс. Географический индекс формируется автоматически по онтологии географических названий региона, с использованием разработанной нами технологии систематизации и поиска электронных публикаций. Реализация данного подхода позволит пользователям, с одной стороны, по выделенным участкам карты автоматически получить доступ к семантически связанным с этими участками электронным документам. С другой стороны, по найденным по запросу пользователя электронным публикациям осуществить автоматический переход к соответствующим участкам карты, к географическим объектам которых они имеют отношение.

Работа выполнена при поддержке гранта РФФИ № 12-07-00070-а.

1 Введение

В настоящее время проблема разработки эффективных методов и технологий интеграции географических информационных систем (ГИС) с т.н. «негеографическими» информационными системами (электронными библиотеками) остается актуальной. Решение этой проблемы открывает для пользователей новые возможности для организации географического поиска в электронных коллекциях документов, например, появляется возможность автоматического доступа по выделенному участку карты к электронным документам, в тексте которых

имеются ссылки на соответствующие географические объекты.

Существует ряд подходов к решению этой проблемы [1–5]. При создании систем географического поиска необходимо решать задачи, связанные с обеспечением соответствия названий географических объектов, указанных в запросе пользователя, с соответствующими ссылками на эти названия, обнаруженными в текстах документов. При этом рассматриваются различные способы построения географического индекса для текстовых документов с использованием онтологии географических названий объектов. Например, в статье [1] для построения географически-ориентированных систем информационного поиска предлагается онтология географических названий объектов (*an ontology of place*), которая может быть использована для определения меры семантического расстояния между географическими объектами при выполнении запросов. В статье [2] рассматривается проблема идентификации и устранения неоднозначности (*disambiguation*) названий географических объектов в документах при определении географического региона, к которому они имеют отношение. Предлагается подход к решению этой проблемы, основанный на применении т.н. географической сигнатуры (определяется как эталонный список однозначных географических ссылок, обнаруженных в документе), которая может быть использована, в частности, для формирования географического индекса документа коллекции. В статье [3] предлагается структура индекса, которая включает как традиционный инвертированный индекс документа, так и пространственный (*space*) индекс, сформированный по онтологии географических названий объектов. Предлагается архитектура системы географического информационного поиска, которая на основе предложенной структуры географического индекса может выполнять как простые, так и комбинированные пространственные и текстовые запросы. Также следует отметить работы [4,5], в которых исследуются различные аспекты решения проблемы географической привязки текстовых документов электронных библиотек, в частности, посредством тезауруса.

В данной статье рассматривается задача обеспечения более «тесной» интеграции ГИС с коллекциями электронных научных публикаций, которая решается на основе разработанной технологии систематизации и поиска данных с применением онтологии [6–10]. В основе подхода лежит идея автоматического формирования для каждой публикации по онтологии географических названий региона т.н. «географического» индекса, который дополняет ее «текстовой» индекс [3,8]. Реализация данного подхода позволит пользователям, с одной стороны, по выделенным участкам карты автоматически получать доступ к семантически связанным с этими участками электронным документам. С другой стороны, по найденным по запросу электронным публикациям, осуществить автоматический переход к соответствующим участкам карты, к географическим объектам которых они имеют отношение.

В настоящее время в рамках развития сервисов информационно-аналитической системы (ИАС) «Природные ресурсы Карелии» осуществляется реализация и апробация предлагаемой технологии интеграции «негеографической» и географической компонент системы [9,10].

2 Технология систематизации и поиска электронных научных публикаций

Построение эффективных систем информационного поиска на основе онтологий (ontology-based information retrieval) в настоящее время является одним из перспективных направлений исследований и разработок [11–13]. Такие системы информационного поиска используют онтологии, как для индексации информационных ресурсов, так и для организации семантического поиска в больших массивах документов.

В течение ряда последних лет нами разрабатывается онтологически-ориентированная технология систематизации и поиска электронных публикаций. В отличие от традиционно используемой для целей индексирования публикаций модели текста как «набор слов» («bag of words» model) мы используем т.н. концептуальное индексирование (в нашем случае индекс публикации формируется не на наборе слов, а на основе предметной онтологии, моделирующей содержание определенного тематического множества публикаций). При этом в основу используемой нами онтологии положены: рубрикатор (в нашем случае ГРНТИ); набор логических условий для автоматизации процесса предметизации публикаций, а также таксономии терминов определенных научных предметных областей, термины которых связаны отношениями классификации «род–вид», «агрегации», «часть–целое» и синонимии.

Процедура систематизации публикаций разделяется на два основных этапа: предметизацию и индексацию. В качестве информационной основы предметизации используются термины таксономии

и набор логических условий (логических функций, описывающих связи научных терминов по определенной тематике исследований), с помощью которых осуществляется процесс отнесения публикаций к соответствующим рубрикам (в нашем случае – к рубрикам ГРНТИ). Для формирования этих логических условий, описывающих, по сути, содержание научных публикаций, используются термины таксономий соответствующих научных предметных областей и логические операторы: AND, OR, NOT.

Процесс индексации также разбивается два основных этапа. На первом этапе выполняется предварительная обработка текста (каждая публикация переводится из формата PDF в формат TXT; из текста удаляются «малоинформационные» слова; к тексту и терминам выделенного фрагмента таксономии применяется алгоритм стемминга – «Стеммер Портера»). На втором этапе осуществляется последовательное сканирование обработанного на предыдущем этапе текста публикации и сопоставление каждого выделенного в тексте слова с терминами соответствующего фрагмента таксономии, характеризующего содержание предметной рубрики. По такой схеме последовательно обходится поддеревья всех рубрик, к которым была отнесена публикация на этапе предметизации. При этом каждый раз производится поиск термина онтологии в тексте публикации и если обнаруживается такое вхождение термина, то индексируется не только этот термин, но и все его предки из соответствующего поддерева рубрики. В результате индекс публикации представляет собой упорядоченную совокупность терминов таксономии, и на наш взгляд более детально характеризует ее содержание по сравнению со списком ключевых слов.

Онтология и база индексов публикаций обеспечивают тематический поиск публикаций по запросам пользователей. Нами разработана технология построения и исполнения запросов, суть которой заключается в следующем. Пользователю сначала предлагается выбрать рубрику ГРНТИ, которая должна содержать интересующие его материалы. Далее ему предлагается соответствующий рубрике фрагмент таксономии, в котором он должен отметить интересующие его термины. С использованием указанных терминов система автоматически формирует запрос в виде логического выражения, определяющего конъюнктивные и/или дизъюнктивные связи терминов.

Следует отметить, что поскольку поиск по запросу осуществляется в базе индексов (а не в текстах электронных публикаций), запрос автоматически расширяется включением в него конъюнкции терминов от корня и дизъюнкции терминов и их синонимов вплоть до листьев от указанных пользователем терминов. Тем самым обеспечивается повышение точности ответа на запрос за счет конъюнкции терминов предыдущих уровней таксономии и полноты за счет дизъюнкции терминов одного уровня таксономии и их синонимов. Список названий найденных по запросу публикаций

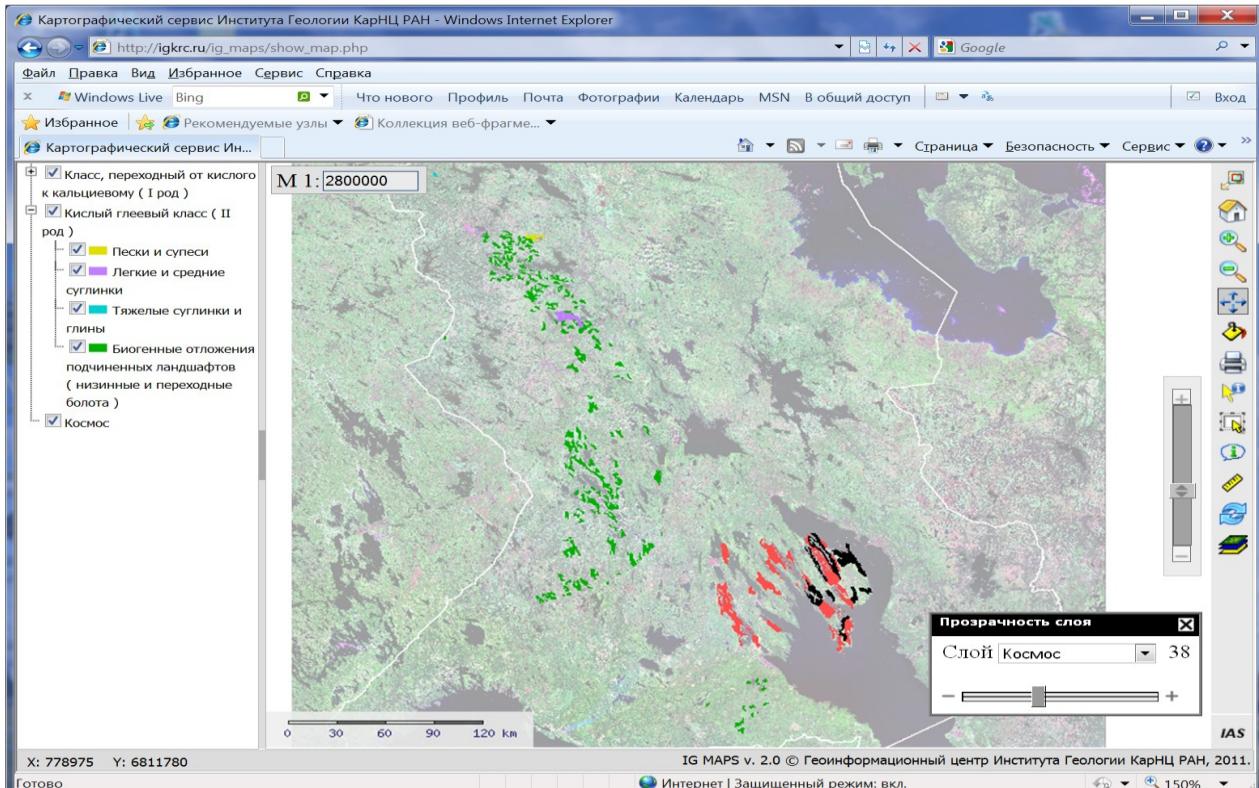


Рис. 1 Страница картографического сервиса

специальным образом ранжируется и выводится пользователю в виде гиперссылок для последующего просмотра или сохранения текстов публикаций в «личном» кабинете пользователя [6].

Использование базы индексов как результата систематизации публикаций непосредственно для их поиска обеспечивает с одной стороны устранение полисемии терминов (т.е. устраняет многозначность терминов за счет «отсечения» других предметных областей в процессе построения запроса), а с другой определяет конкретную предметную область запроса. Тем самым обеспечивается как релевантность, так и пертинентность найденных системой по запросу документов.

Следует также заметить, что пользователю на наш взгляд гораздо проще и точнее выразить свои информационные потребности путем указания терминов в таксономии по сравнению с заданием списка ключевых слов. При этом ему не надо формировать логические условия отбора данных с использованием логических операторов: AND, OR, NOT (система делает это автоматически). Детали разрабатываемой технологии более подробно представлены в работах [6–10].

3 Интеграция ГИС с электронными коллекциями научных публикаций в ИАС «Природные ресурсы Карелии»

В настоящее время в крупных научных центрах и вузах накоплено огромное количество самой разнообразной научной информации (коллекции

электронных научных публикаций, базы данных, ГИС и т.п.), рациональное использование которой в интересах развития науки, образования и экономики страны в целом остается проблемой. Для ее решения требуется разработка и применение интегрированных информационных систем, обеспечивающих накопление, систематизацию и эффективный доступ пользователей к разнородным информационным ресурсам. Одним из примеров такого рода систем является разрабатываемая нами информационно-аналитическая система «Природные ресурсы Карелии», которая включает две основные компоненты – сервисы накопления, систематизации и поиска электронной научной информации, а также набор тематических карт, отражающих результаты научных исследований, и необходимые для работы с ними картографические сервисы [8–10]. Эта система предназначена для информационной поддержки ученых и специалистов, как при проведении научных исследований, так и для оценки состояния окружающей среды и экологических последствий, планируемых и проводимых на территории Карелии и сопредельных регионов мероприятий в сфере промышленности, лесного, сельского и рыбного хозяйства.

Картографический сервис ИАС «Природные ресурсы Карелии» базируется на свободном программном обеспечении. Для Web-сервера используется Ubuntu Server v.7.10, а основой картографического сервиса – MapServer v.4.10.3.

Первый этап работы с сервисом заключается в инициализации пользователя. После этого пользова-

Поиск по таксономии терминов

выделить все | убрать выделение

- 39. География (656)
 - 39.29. Топонимика (656)
 - 39.29.15. Региональная топонимика (656)
 - Республика Карелия (656)
 - Озера (388)
 - Падмозеро (6)
 - Путкозеро (7)

Для ключевых слов: 39. География И 39.29. Топонимика И 39.29.15.
Региональная топонимика И Республика Карелия И Озера И Падмозеро (6)

- 1. Зоопланктон водных объектов бассейна Онежского озера (156)
- 2. Альгофлора озер и рек Карелии. Таксономический состав и экология (133)
- 3. Библиография работ по водорослям Европейского Севера России (Республика Карелия, Мурманская область) (110)
- 4. Планктонная фауна водоемов Заонежского полуострова (100)
- 5. Деградация поздневалдайского оледенения в бассейне Онежского озера (91)
- 6. Современное состояние макрообиотоса озер Заонежского полуострова (87)

Для ключевых слов: 39. География И 39.29. Топонимика И 39.29.15.
Региональная топонимика И Республика Карелия И Озера И Путкозеро (6)

- 1. Зоопланктон водных объектов бассейна Онежского озера (157)
- 2. Альгофлора озер и рек Карелии. Таксономический состав и экология (148)
- 3. Кумжи, радужная форель, гольцы и перспективы их использования в озерах Северо-Запада России (102)
- 4. Планктонная фауна водоемов Заонежского полуострова (101)
- 5. Современное состояние макрообиотоса озер Заонежского полуострова (92)
- 6. Деградация поздневалдайского оледенения в бассейне Онежского озера (91)
- 7. О некоторых закономерностях в распределении карбонатных пород среди лидитов шунгит-доломит-лидитового комплекса в старом карьере п. Шуньга (37)

Рис. 2 Поиск научных публикаций по онтологии географических названий региона.

тель имеет возможность подготовки набора из картографического материала хранящегося в соответствующей базе данных и состоящего из векторных (с атрибутивной информацией) и растровых электронных карт, а также космических снимков. Картографический материал систематизирован как по тематике (с выделением отдельных тем), так и по масштабу. Основная страница картографического сервиса содержит стандартные инструменты просмотра картографической информации: масштабирование и сдвиг карты, окно обзора, установление уровня прозрачности тематических слоев, получение информации об объектах и выбор объектов в области. Легенда имеет древовидную структуру и оформлена в виде всплывающего (Рис. 1) окна.

На начальной стадии разработки географическая (коллекции тематических научных карт) и «негеографическая» (коллекции электронных научных публикаций) компоненты системы функционировали практически независимо друг от друга. Необходимость более «тесной» интеграции этих компонент послужило для нас стимулом для поиска эффективного решения этой проблемы.

Проведенный анализ достаточно представительного массива электронных научных публикаций (http://www.krc.karelia.ru/rio_stats.php) по ресурсоведческим направлениям исследований ученых КарНЦ РАН (биология, лесное хозяйство, геология, водные ресурсы) показал, что в подавляющем большинстве текстов публикаций имеются ссылки на названия географических объектов (т.е. названия населенных пунктов, болот, озер, рек, месторождений и т.п.), с которыми связаны результаты проведенных исследований. Это обстоятельство дало нам основания для применения разработанной технологии систематизации и поиска электронных

публикаций для индексации этих электронных публикаций по онтологии географических названий региона. В результате применения процедуры индексации к документам коллекции электронных публикаций, каждой публикации, дополнительно к ее «текстовому» индексу, приписывается еще и «географический» индекс. В настоящее время при построении онтологии географических названий региона использовались атрибутивные данные карты Карелии масштаба 1:1000 000 (СевЗапАэроГеодезия, 2000–2001 гг.) по административным районам Карелии, озерам, населенным пунктам и др. В качестве эксперимента, по сформированному нами фрагменту онтологии (города, поселки и озера Карелии) были проиндексированы все научные публикации электронной библиотеки.

На рис. 2 представлен пример поиска научных публикаций, проиндексированных по выделенному фрагменту онтологии географических названий региона.

Онтология географических названий региона имеет иерархическую (древовидную) структуру (т.е. соответствует административно-территориальному делению региона с включением названий природно-культурных и др. объектов, расположенных на данной территории, а также их синонимов, с помощью которых на наш взгляд успешно решается вопрос, связанный с историческим переименованием географических объектов).

Следует отметить, что в нашем случае «географический» индекс электронной публикации формируется по той же схеме, что и «текстовый» индекс публикации. Общая схема организации географического поиска электронных научных публикаций в ИАС «Природные ресурсы Карелии» с использованием географического индекса представлена на рис. 3.

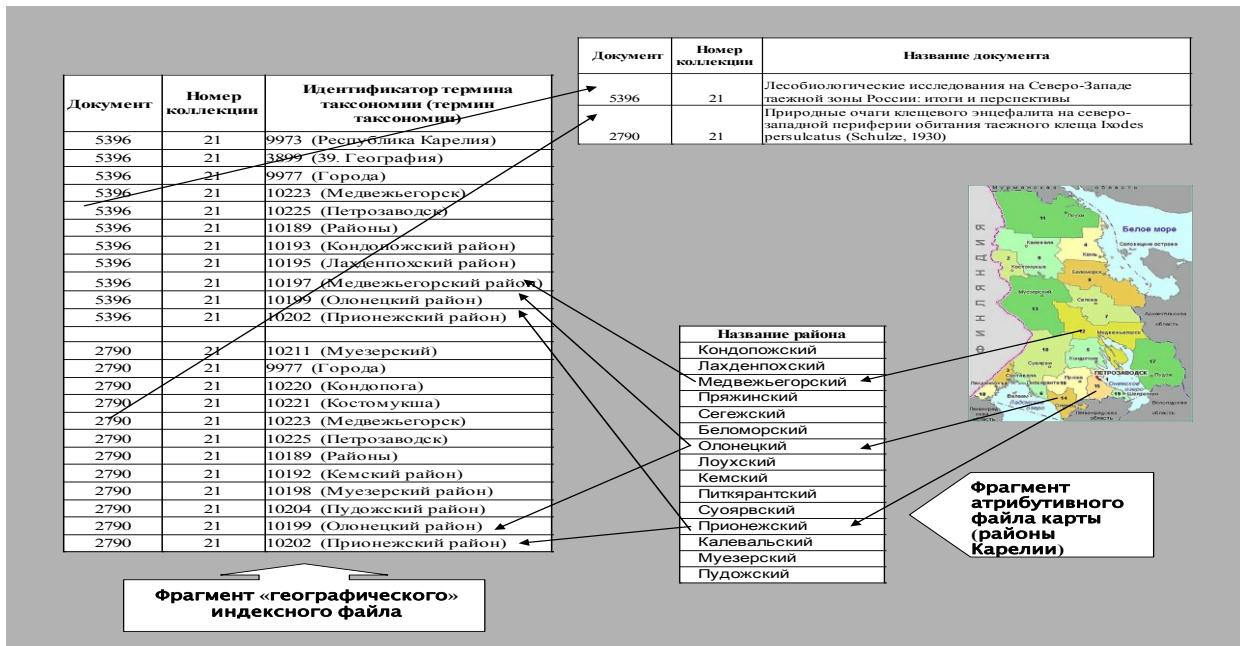


Рис. 3 Общая схема организации географического поиска электронных научных публикаций

Такой подход к решению задачи обеспечения более «тесной» интеграции картографической компоненты ИАС с коллекциями электронных научных публикаций позволяет реализовать для пользователя следующую схему доступа, как к ГИС-данным, так и к электронным научным публикациям.

- Географический поиск.** Пользователь с помощью специального сервиса доступа к ГИС-данным выделяет на тематической карте интересующую его область (например, в виде прямоугольника). Система осуществляет **обратное геокодирование** (т.е. реализуется процедура запроса всех названий объектов по указанным координатам с использованием онтологии географических названий). Далее пользователь ИАС может осуществить поиск электронных публикаций, семантически связанных с выделенными названиями географических объектов, двумя следующими способами:

1. Запустить процедуру поиска электронных публикаций, в географических индексах которых присутствуют названия указанных географических объектов.
2. Обратиться к сервису доступа к электронным публикациям и с его помощью получить нужные публикации, как по интересующей его тематике, так и с учетом их связи с выделенными географическими объектами (в этом случае при формировании логического условия отбора данных система автоматически добавит в условие отбора дизъюнкцию названий соответствующих географических объектов, указанных в запросе).

- Текстовый поиск.** Пользователь с помощью сервиса доступа к электронным публикациям находит нужные ему по тематике публикации и,

при необходимости, может запросить те участки карты, с которыми найденные публикации семантически связаны (в этом случае система осуществляет **прямое геокодирование** с использованием онтологии географических названий региона и по географическому индексу найдет соответствующие участки тематической карты).

4 Заключение

В настоящее время данная технология используется нами при разработке и развитии сервисов информационно-аналитической системы «Природные ресурсы Карелии» – <http://ias.krc.karelia.ru>. Реализация предлагаемого подхода позволит пользователям ИАС по запросу находить нужную научную информацию, которая относится к конкретному географическому участку региона, используя при этом как текстовую, так и географическую формы поиска.

Авторы приносят свои благодарности Ю.В. Чирковой, Н.Б. Луговой и В.Г. Старковой за плодотворное обсуждение рассматриваемых вопросов и реализацию исследовательского прототипа технологии.

Литература

- [1] Christopher B. Jones, Harith Alani and Douglas Tudhope (2001) Geographical Information Retrieval with Ontologies of Place. In Proceedings of the International Conference on Spatial Information Theory: Foundations of Geographic Information Science (COSIT), Morro Bay, CA, USA, 322–335.
- [2] David S Batista, Mário J Silva, Francisco M Couto, Bibek Behera Geographic Signatures for Semantic Retrieval //GIR'10 18-19th Feb. 2010, Zurich, Switzerland.

- [3] Nieves R. Brisaboa, Miguel R. Luaces, Ángeles S. Places, Diego Seco Exploiting geo-graphic references of documents in a geographical information retrieval system using an ontology-based index //Geoinformatica (2010) 14:307–331 DOI 10.1007/s10707-010-0106-3.
- [4] О.Л. Жижимов, Н.А. Мазов, Проблемы географической привязки цифровых объектов в электронных библиотеках // Труды 12й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2010, Казань, Россия, 2010 г. С. 207–214.
- [5] Д.М. Скачков, О.Л. Жижимов Об использовании ретроспективного геокодирования для географического поиска в электронных библиотеках // Труды 13й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» – RCDL'2011, Воронеж, Россия, 19–22 октября 2011 г. С. 51–58.
- [6] Вдовицын В.Т., Лебедев В.А. Ранжирование документов в системе поиска, основанной на применении онтологии // Труды XIV Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». RCDL'2012, Переславль-Залесский. 15–18 октября 2012 г. (принята к печати).
- [7] В. Вдовицын, В. Лебедев Технологии систематизации и поиска электронной научной информации с применением онтологий // Информационные ресурсы России. – 2010. – № 5. – С. 6–10.
- [8] В. Вдовицын, В. Лебедев. Технологии информационного обеспечения научных исследований в ИАС «Природные ресурсы Карелии» // Информационные ресурсы России. № 1. 2012. С. 7–12.
- [9] Титов А.Ф., Вдовицын В.Т., Лебедев В.А., Полин А.К. Информационно-аналитическая система поддержки и сопровождения исследований природных ресурсов региона //Труды XII Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». RCDL'2010, Казань. 13–16 октября 2010 г. С. 529–534.
- [10] В.Т. Вдовицын, А.К. Полин Разработка и развитие картографического сервиса информационно-аналитической системы «Природные ресурсы Карелии» //Интернет и современное общество. Сборник научных статей. Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». Санкт-Петербург, 12–14 октября 2011 г. С. 36–39.
- [11] Miriam Fernandez, Iván Cantador, Vanesa Lypez, David Vallet, Pablo Castells, Enrico Motta Semantically enhanced Information Retrieval: An ontology-based approach /Web Semantics: Science, Services and Agents on the World Wide Web 9 (2011) 434–452.
- [12] Raquel Trillo, Laura Po, Sergio Ilarri, Sonia Bergamaschi, Eduardo Mena Using semantic techniques to access web data //Information Systems. 36 (2011). P. 117–133.
- [13] Mauro Dragoni, Célia da Costa Pereira, Andrea G.B. Tettamanzi An Ontological Representation of Documents and Queries for Information Retrieval Systems /Proceedings of the 1-st Italian Information Retrieval Workshop (IIR'10), January 27–28, 2010, Padua, Italy, <http://ceur-ws.org/Vol-560/paper18.pdf> (дата обращения: 18.04.2012).

Ontology-oriented approach towards integration of geographical information system with digital collections of scientific publications

Vladimir Vdovitsyn, Aleksandr Polin

We suggest an ontology-oriented approach towards integration of geographical information system with collections of digital publications. The approach rests on the idea of supplying each publication with generated «geographical» index to complement its «textual» index. The geographical index will be generated automatically from the ontology of the region place names applying the authors' technology for systematization and retrieval of digital publications. Through this approach users will be able to automatically access the digital documents semantically related to the areas they highlight in the map. Vice versa, they can be automatically redirected from the digital publications found after their inquiry to the map areas corresponding to the geographical objects relevant for the publications.