спутниковых ИК измерений // Оптика атмосферы и океана. – 2010. – Т. 23. – № 8. – С. 684–690.
10. Giovanni – The Bridge Between Science and Data. 2007. URL: http://disc.sci.gsfc.nasa.gov/giovanni (дата обращения: 01.12.2010).
11. AERONET Data Synergy Tool – Access Earth Science data sets for AERONET sites. 2007. URL: http://aeronet.gsfc.nasa.gov (дата обращения: 01.12.2010).
12. Google Планета Земля. 2010. URL: http://www.google.com/intl/ru/earth/index.html (дата обращения: 01.12.2010).
13. Atmospheric Correction Parameter Calculator. 2009. URL: http://atmcorr.gsfc.nasa.gov (дата обращения: 01.12.2010).
14. Barsi J.A., Barker J.L., Schott J.R. An Atmospheric Correction Parameter Calculator for a Single Thermal Band Earth-Sensing Instrument // Atmospheric Correction Parameter Calculator. 2009. URL: http://atmcorr.gsfc.nasa.gov/Barsi_IGARSS03.pdf (дата обращения: 01.12.2010).
15. Barsi J.A., Schott J.R., Palluconi F.D., Hook S.J. Validation of a Web-Based Atmospheric Correction Tool for Single Thermal Band Instruments // Atmospheric Correction Parameter Calculator. 2009. URL: http://atmcorr.gsfc.nasa.gov/Barsi_AtmCorr_SPIE05.pdf (дата обращения: 01.12.2010).
16. Django Software Foundation. 2005. URL: http://www.djangoproject.com (дата обращения: 01.12.2010).
17. LAADS FTP site. 2007. URL: ftp://ladsweb.nascom.nasa.gov (дата обращения: 01.12.2010).
18. LAADS Web – Level 1 and Atmosphere Archive and Distributiuon System. 2007. URL: http://ladsweb.nascom.nasa.gov/index.html (дата обращения: 01.12.2010).
19. National Weather Service. 2005. URL: http://www.weather.gov (дата обращения: 01.12.2010).

Поступила 17.12.2010 г.

УДК 002.53:004.89

ОБЕСПЕЧЕНИЕ СОДЕРЖАТЕЛЬНОГО МНОГОЯЗЫЧНОГО ДОСТУПА К ЛИНГВИСТИЧЕСКИМ ИНФОРМАЦИОННЫМ РЕСУРСАМ НА ОСНОВЕ ТЕХНОЛОГИИ ПОРТАЛОВ ЗНАНИЙ

Ю.А. Загорулько, О.И. Боровикова, И.С. Кононенко

Институт систем информатики им. А.П. Ершова СО РАН, г. Новосибирск
Новосибирский государственный университет
E-mail: zagor@iis.nsk.su

Рассматривается интернет-портал знаний, обеспечивающий систематизацию и интеграцию знаний и информационных ресурсов по компьютерной лингвистике на основе онтологии, а также содержательный многоязычный доступ к ним: управляемую онтологией навигацию и поиск информации в терминах предметной области портала знаний.

Ключевые слова:
Портал знаний, компьютерная лингвистика, информационные ресурсы, онтология, содержательный доступ, управляемая онтологией навигация.

Key words:
Knowledge portal, computational linguistics, information resources, ontology, content-based access, ontology-driven navigation.

Введение

В настоящее время наблюдается бурный рост потребности в средствах автоматической обработки документов и естественно-языковых, в том числе речевых, интерфейсах. Это ставит на повестку дня проблему организации эффективного доступа не только к публикациям, описывающим методы и подходы к пониманию текста и речи, но и разного рода словарям, программным компонентам и алгоритмам, обеспечивающим решение различных задач по их обработке. И хотя в Интернете представлен большой объем информационных ресурсов по этой тематике, доступ к ним весьма затруднен, т. к. они плохо систематизированы и рассредоточены по различным Интернет-сайтам, каталогам и электронным архивам.

Для решения этой проблемы разрабатываются различные интернет-ресурсы. Самым известным из них является англоязычный каталог LINGUIST List [1], созданный для обмена знаниями между лингвистами и содержащий информацию о публикациях, персоналиях, научных учреждениях, грантах, конкурсах, проектах, фондах, конференциях и семинарах лингвистической тематики.

Российским аналогом LINGUIST List является портал «Лингвистика в России: ресурсы для исследователей» [2], организованный в виде иерархического каталога ссылок, тематические категории которого представлены разделами по компьютерной, теоретической и прикладной лингвистике и их приложениям.

Из других разработок стоит отметить созданный в Германском Исследовательском Центре Искусственного Интеллекта информационный портал «Language Technology World» [3]. Тематические разделы этого портала содержат информацию о лингвистических технологиях, продуктах и информационных системах в области обработки естественного языка, а также о проектах, организациях и персонах. В основу портала положена онтология [4, 5], благодаря чему возможно установление связей между его разделами. К сожалению, на этом

портале отсутствует информация о российских исследователях и исследованиях, проводимых в России.

Существуют также Интернет-ресурсы, посвященные отдельным направлениям компьютерной лингвистики. В качестве примера можно привести российский сайт «Речевые технологии» [6], на котором представлена информация о прикладных аспектах развития данного направления (технологии, программные средства, конкретные системы и т. п.), а также каталог систем генерации текстов [7], содержащий информацию практически обо всех известных системах генерации текстов (на момент написания статьи каталог включал описания около 400 систем).

Практически все известные интернет-ресурсы по компьютерной лингвистике либо направлены на информационную поддержку лингвистических сообществ и представление общеязыковой лингвистической информации, либо имеют узкую тематическую направленность. При этом ни один из них не ориентирован одновременно и на интеграцию информационных ресурсов по компьютерной лингвистике и на обеспечение к ним содержательного доступа широкому кругу пользователей. Для решения этой проблемы нами разработан специализированный интернет-портал – портал знаний по компьютерной лингвистике (КЛ).

1. Информационная модель портала знаний по компьютерной лингвистике

Чтобы портал знаний мог предоставлять пользователям описанные выше возможности, он должен не только иметь гибкие средства представления разнородной информации по компьютерной лингвистике и содержательного доступа к ней, но и обеспечивать оперативное управление своим информационным наполнением (контентом). Этим целям служит информационная модель портала знаний [8], которая объединяет модели его предметной и проблемной областей, а также описывает типы представляемой в его контенте информации.

Формально информационная модель портала M_p описывается двойкой $M_P=<O_P,IC_P>$, где O_P – онтология портала, а IC_P – информационное содержание (контент) портала.

Онтология O_P является ядром, базовым компонентом информационной модели портала. Она не только описывает систему знаний портала, но и задает формальные структуры для представления его контента IC_P.

Для представления онтологии портала предложен следующий формализм:

$$O=<C,R,T,D,A,F,Ax>,$$

где $C=\{C_1,\ldots, C_n\}$ – конечное непустое множество классов, описывающих понятия некоторой предметной или проблемной области; $R=\{R_1,\ldots,R_m\}$, $R_i \subseteq C\times C$, $R=\{R_T\}\cup\{R_P\}\cup R_A$ – конечное множество бинарных отношений, заданных на классах (понятиях); здесь R_T – антисимметричное, транзитивное, нерефлексивное бинарное отношение наследования, задающее частичный порядок на множестве понятий C, R_P – бинарное транзитивное отношение включения («часть-целое»), R_A – конечное множество ассоциативных отношений; T – множество стандартных типов; $D=\{d_1,\ldots, d_r\}$ – множество доменов $d_i=\{s_1,\ldots, s_k\}$, где s_i – значение стандартного типа T; A – конечное множество атрибутов, описывающих свойства понятий C и отношений R_A; F – множество ограничений на значения атрибутов понятий и отношений; Ax – множество аксиом, определяющих семантику классов и отношений онтологии.

Данный формализм обеспечивает описание понятий проблемной и предметной областей портала и разнообразных семантических связей между ними, а также выстраивание понятий в иерархию «общее-частное» (с помощью отношения R_T) и поддержку наследования свойств по этой иерархии. Его особенностью является то, что при наследовании от родительского класса его классу-потомку передаются не только все его атрибуты, но и отношения. Другая особенность предложенного формализма – он позволяет задавать для ассоциативных отношений R_A атрибуты, специализирующие связи между их аргументами (объектами).

С содержательной точки зрения онтология портала служит для представления понятий, необходимых для описания как научной деятельности и научного знания в целом, так и конкретной научной дисциплины, в частности. В связи с этим онтология портала включает универсальные онтологии научной деятельности и научного знания [8], а также онтологию научной дисциплины «компьютерная лингвистика».

Онтология научной деятельности является онтологией верхнего уровня и включает базовые понятия, относящиеся к организации научно-исследовательской деятельности, такие как *Персона*, *Организация*, *Событие*, *Деятельность*, *Публикация*, используемые для описания участников научной деятельности, мероприятий, научных программ и проектов, различного типа публикаций. В эту онтологию также включено понятие *Информационный ресурс*, которое служит для описания информационных ресурсов, представленных в сети Интернет.

Онтология научного знания, по своей сути, является метаонтологией. Она содержит метапонятия, задающие структуры для описания предметной области (области знаний) портала, такие как *Раздел науки*, *Предмет исследования*, *Объект исследования*, *Метод исследования*, *Научный результат*, позволяющие выделить в данной науке значимые разделы и подразделы, задать типизацию предметов, объектов и методов исследования, описать результаты научной деятельности.

Свойства каждого понятия описываются с помощью атрибутов и ограничений, наложенных на область их значений. Понятия базовых онтологий связаны между собой ассоциативными отношениями, выбор которых осуществлялся не только исходя из полноты представления проблемной

и предметной областей портала, но и из удобства навигации по его информационному пространству и поиска информации.

Понятия **онтологии научной дисциплины «компьютерная лингвистика»** являются реализациями метапонятий онтологии научного знания и организованы в 5 иерархий «общее-частное», каждая из которых соответствует одному из метапонятий, представленных в этой онтологии. Все эти иерархии связаны между собой посредством ассоциативных отношений, часть которых наследуется из базовых онтологий, а часть отражает специфику данной предметной области.

Предметом исследования в КЛ являются *Процессы и задачи*, связанные с функционированием языковых единиц в коммуникации (*Морфологический анализ*, *Моделирование звуков речи* и т. п.), и *Прикладные процессы*, отвечающие определенному социальному запросу (*Классификация документов*, *Распознавание голосовых команд* и т. п.). Иерархия предметов исследования связана ассоциативным отношением «Аспект» с иерархией объектов исследования и отношением «Предмет изучения» с иерархий разделов науки.

В качестве базовых объектов исследования КЛ рассматриваются *Невербальные коммуникации*, *Речевые произведения* (РП), как объективная форма существования и использования ЕЯ, и *Структурные языковые единицы*, соответствующие различным языковым уровням: предложения, словосочетания, слова, морфемы, звуки и пр. Класс понятий РП представлен в иерархии двумя подклассами: *Текст* и *Звучащая речь*. Для представления связи между целостными РП и их структурными единицами используется отношение «Включение».

Иерархия методов исследования служит для систематизированного описания инструментов исследования, применяемых в КЛ. В этой иерархии выделены следующие подклассы: *Методы обработки текста*, *Методы обработки речи*, *Методы теоретической лингвистики*, *Математические модели и методы* и др.

В основе иерархии разделов КЛ лежит классификация базовых теоретических и прикладных направлений компьютерной лингвистики. В качестве главных разделов КЛ выделены *Моделирование языка и языковой деятельности* и *Создание прикладных систем*. Первый из них включает подразделы *Автоматическая обработка текста* (АОТ), *Речевые технологии* (РТ) и др. Другой раздел включает подразделы *Создание прикладных систем АОТ*, *Создание прикладных систем РТ*, *Машинный перевод*, *Вопросно-ответные системы* и др.

В иерархии научных результатов выделены такие классы, как *Технологии и программные продукты*, *Прикладные системы*, *Лингвистические ресурсы*. Последний класс включает такие подклассы, как *Словари и тезаурусы*, *Лексические онтологии*, *Корпуса* (текстовые и речевые) и *Лингвистические БД*.

Вводя формальные описания понятий области знаний портала в виде классов объектов и отношений между ними, онтология задает структуры для представления реальных объектов и связей между ними. В соответствии с этим данные на портале представлены как множество разнотипных информационных объектов и связей между ними, которые в совокупности и образуют контент портала.

2. Контент портала знаний

Контент портала включает как знания общего характера, представленные в онтологии, так и конкретные знания о реальных объектах и информационных ресурсах, систематизированные в соответствии с онтологией портала.

В первую очередь, портал содержит знания об основных разделах компьютерной лингвистики, о ее предметах и объектах исследования, используемых в ней моделях, методах и алгоритмах. Кроме того, пользователи портала могут получить представление не только о КЛ как научной дисциплине, но и найти информацию о выполняемой в этой области научной и производственной деятельности.

В деятельности организаций и исследователей особое место занимают научные и коммерческие проекты, в рамках которых, большей частью, и создаются лингвистические знания и ресурсы. Портал знаний обеспечивает доступ к результатам такой деятельности, отраженной как в публикациях – монографиях, статьях, материалах конференций, отчетах и других текстовых ресурсах, так и в виде технологий, программных продуктов, прикладных систем, словарей, корпусов текстов и лингвистических БД. Для организации более эффективного доступа к ресурсам, описывающим эти результаты, в контенте портала представлена информация о различных аспектах их разработки (получения): организациях, персонах и проектах, с которыми связано их появление, а также о таких их содержательных характеристиках, как отнесенность к разделу науки, объекту, предмету или методу исследования. Эта информация связывает эти ресурсы с остальными данными и знаниями, представленными в контенте портала, что позволяет пользователю выделить группы ресурсов, созданные, например, в ходе осуществления некоторого проекта или с использованием определенного метода исследования.

Важным компонентом контента портала является описание интернет-ресурсов, систематизированных в соответствии с онтологией портала. К таким ресурсам относятся сайты организаций, конференций, проектов, порталы и каталоги, посвященные компьютерной лингвистике, а также отдельные страницы с материалами графического, мультимедийного или текстового типа. Набор атрибутов и связей ресурса основан на стандарте Dublin Core [9]. Само описание ресурса хранится в контенте портала и включает экземпляр понятия *Информационный ресурс* и набор экземпляров отношений, связывающих данный ресурс с другими объектами, представляющими организации, исследователей, публикации, события и др.

3. Настройка и управление контентом портала знаний

Для настройки портала знаний на предметную область и предпочтения пользователя, а также управления его контентом используются специализированные редакторы, реализованные как web-приложения и доступные зарегистрированным пользователям через Интернет, а также коллекционер онтологической информации о ресурсах, рис. 1.

Настройка портала на предметную область осуществляется с помощью редактора онтологии, который позволяет создавать, редактировать и удалять любые элементы онтологии (классы понятий, отношения и др.), а также задавать и модифицировать иерархии понятий.

Этот редактор проектировался таким образом, чтобы он был прост и удобен в использовании для экспертов, не являющихся специалистами в области информатики и математики. В частности, из-за этих требований мы отказались от такого популярного средства построения онтологий как редактор Protégé [10].

Для более удобного представления информации пользователю портала в редактор онтологий включены средства настройки визуализации знаний и данных, которые позволяют для каждого класса онтологии задать шаблон визуализации объектов этого класса и шаблон визуализации ссылок на эти объекты.

Для настройки портала знаний на различные языки в его состав включен многоязычный тезаурус [11], который строится как лингвистическое дополнение онтологии и включает термины проблемной и предметной области портала, т. е. слова и словосочетания на нескольких естественных языках, с помощью которых понятия онтологии представляются в текстах и пользовательских запросах. Наличие отношений, связывающих термины тезауруса с понятиями онтологии, обеспечивает возможность визуализации онтологии и представленной в контенте портала информации на разных языках (по выбору пользователя), а также поддерживает навигацию по его контенту и формулирование запросов с использованием удобного для пользователя естественного языка.

Управление контентом портала осуществляется с помощью редактора данных, который позволяет создавать, редактировать и удалять информационные объекты (экземпляры классов онтологии) и связи между ними. Формы для ввода конкретных информационных объектов и связей между ними автоматически генерируются на основе онтологии.

Для автоматизации пополнения контента портала знаний релевантными информационными ресурсами, был разработан коллекционер онтологической информации [12], который осуществляет сбор, анализ, оценку релевантности, автоматическое индексирование и классификацию интернет-ресурсов.

4. Обеспечение содержательного доступа к лингвистическим ресурсам

Содержательный доступ к систематизированным знаниям и информационным ресурсам по компьютерной лингвистике, представленным в портале знаний, организуется на основе рассмотренной выше информационной модели и осуществляется путем навигации по контенту портала, а также через развитые средства содержательного поиска.

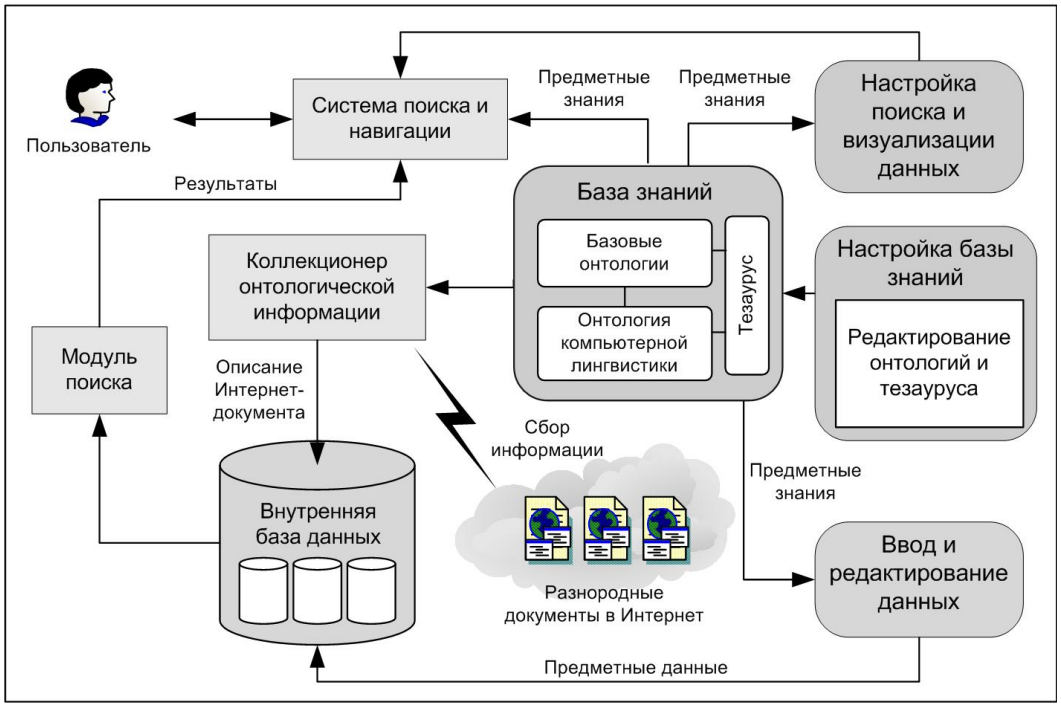

Рис. 1. *Общая схема настройки и доступа к данным и знаниям портала*

Навигация по контенту портала начинается с выбора некоторого класса C в дереве понятий онтологии, построенном на основе отношения R_T. При этом пользователю выдается список информационных объектов выбранного класса, который отображается в виде html-страницы, содержащей набор ссылок на эти объекты. Следует заметить, что при формировании такого списка выполняется транзитивное замыкание по отношению R_T, благодаря чему результирующий список будет включать как объекты искомого класса C, так и объекты его классов-наследников.

Информация о конкретном объекте и его связях также отображается в виде html-страницы (рис. 2), формат и наполнение которой зависят от класса данного объекта и заданных для него отношений и шаблона визуализации. При этом объекты, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию.

Дальнейшая навигация по порталу представляет собой процесс перехода от одних информационных объектов к другим по заданным между ними связям – экземплярам ассоциативных отношений R_A. Например, при просмотре информации о конкретном проекте (например, AGILE) мы можем видеть значения его атрибутов и его связи с другими объектами, рис. 2. Используя представленные связи в качестве элементов навигации, можно перейти к просмотру подробной информации как по прямым связям (об объекте исследования, об используемых методах и научных результатах, полученных в ходе выполнения проекта), так и по обратным (об участниках проекта, публикациях о проекте, информационном ресурсе, описывающем данный проект).

При переходе по конкретной связи любого информационного объекта мы можем получить достаточно большой список объектов (например, публикаций). В связи с этим был введен механизм фильт-

Просмотр объекта

Проект	
Название деятельности	Проект AGILE
Описание деятельности	Automatic Generation of Instructions in Languages of Eastern Europe
Дата начала	1 января 1998
Дата окончания	31 декабря 2000

Связи объекта

Результат-Деятельности

Научный Результат
Система AGILE

Направление деятельности

Раздел Науки
Генерация текста

Ссылки на объект

Персона-Участник-Деятельности

Персона	Роль Участника Деятельности
Bateman, J.A.	исполнитель
Hana, J.	исполнитель
Hartley, T.	исполнитель
Kruijff, G.-J.	исполнитель
Kruijff-Korbayová, I.	исполнитель

(Всего: 10)

Организация-Участник-Деятельности

Организация
Information Technology Research Institute, University of Brighton, ITRI
Institute for Applied Linguistics, University of the Saarland
Institute of Formal and Applied Linguistics(Charles University), ÚFAL
Institute of Information Technology, Bulgarian Academy of Sciences
РосНИИ искусственного интеллекта, РосНИИ ИИ

Публикация о Деятельности

Публикация
Bateman, J.A., Hana, J., Hartley, T., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Staykova, K., Teich, E., Соколова, Е.Г., Шаров, С.А., A multilingual system for text generation in three slavic languages, 2000, статья
Bateman, J.A., Kruijff, G.-J., Kruijff-Korbayová, I., Skoumalová, H., Teich, E., Шаров, С.А., Resources for multilingual text generation in three Slavic languages, 2000, статья

Ресурс-Деятельности

Ресурс
Сайт проекта AGILE

Рис. 2. *Представление информационного объекта и его связей*

рации списков, который позволяет, например, отфильтровать множество публикаций как по дате публикации (условия на атрибут), так и по описываемому научному результату или объекту исследования (условия на связанный объект).

При поиске информации пользователю предоставляется возможность формулирования запроса в терминах предметной области портала. При этом он должен выбрать понятие, к которому относятся искомые информационные объекты, и определить ограничения, которым должны удовлетворять атрибуты выбранного понятия и его связи с другими понятиями.

Поисковые запросы задаются через специальный графический интерфейс, управляемый онтологией портала знаний. При выборе пользователем понятия автоматически генерируется поисковая форма, в которой можно задать ограничения на значения атрибутов объектов выбранного понятия, а также на значения атрибутов объектов, связанных с данным объектом ассоциативными отношениями.

Заключение

Рассмотренный портал знаний обеспечивает систематизацию и интеграцию знаний и доступных информационных ресурсов, относящихся к компьютерной лингвистике, в единое информационное пространство, а также содержательный доступ к ним. Благодаря тому, что систематизация и структуризация таких знаний и информационных ресурсов выполнена на основе онтологии, доступ к ним осуществляется путем навигации по дереву понятий онтологии и контенту портала, а также через средства поиска в терминах его предметной области.

Портал знаний по компьютерной лингвистике доступен по адресу [13]. Его пользователями могут стать как научные работники, преподаватели и студенты, исследующие, преподающие и изучающие эту дисциплину, так и специалисты, разрабатывающие программные системы, предназначенные для обработки текстов, анализа и синтеза речи.

При создании портала использовались программные средства, методология и технология разработки порталов научных знаний предложенные в [11, 12, 14].

Ближайшими целями авторов является доработка и развитие онтологии компьютерной лингвистики и связанного с ней многоязычного тезауруса, а также пополнение контента портала новыми лингвистическими ресурсами. Кроме того, планируется подключение к порталу знаний развитых средств графической визуализации, что позволит представлять в виде графа не только иерархии понятий онтологии, но и весь контент.

Работа выполняется при финансовой поддержке РФФИ (проект № 09–07–00400).

СПИСОК ЛИТЕРАТУРЫ

1. The LINGUIST List. Дата обновления: 27.01.2011. URL: http://linguistlist.org/ (дата обращения: 27.01.2011).
2. Научно-образовательный портал «Лингвистика в России: ресурсы для исследователей». 2004. URL: http://uisrussia.msu.ru/linguist/index.jsp (дата обращения: 27.01.2011).
3. Language Technology World. Дата обновления: 23.01.2010. URL: http://www.lt-world.org/ (дата обращения: 10.01.2011).
4. Gruber T. Toward Principles for the Design of Ontologies Used for Knowledge Sharing // International Journal of Human-Computer Studies. – 1995. – V. 43. – Iss. 5–6. – P. 907–928.
5. Guarino N. Formal Ontology in Information Systems // Formal Ontology in Information Systems. Proc. of FOIS'98, Trento, Italy, 6–8 June 1998. – Amsterdam: IOS Press, 1998. – P. 3–15.
6. Портал «Речевые технологии». 2011. URL: http://speech-soft.ru/ (дата обращения: 10.01.2011).
7. NLG Systems Wiki. Дата обновления: 23.02.2010. URL: http://www.nlg-wiki.org/systems/ (дата обращения: 10.01.2011).
8. Загорулько Ю.А., Боровикова О.И. Информационная модель портала научных знаний // Информационные технологии. – 2009. – № 12. – С. 2–7.
9. Using Dublin Core. [1995–2011]. 2011. URL: http://dublincore.org/documents/usageguide/ (дата обращения: 10.01.2011).
10. Protégé. 2010. URL: http://protege.stanford.edu/ (дата обращения: 10.01.2011).
11. Загорулько Ю.А. Подход к обеспечению многоязычного доступа к систематизированным знаниям и информационным ресурсам заданной предметной области // Известия Томского политехнического университета. – 2009. – Т. 314. – № 5. – С. 161–165.
12. Загорулько Ю.А. Автоматизация сбора онтологической информации об Интернет-ресурсах для портала научных знаний // Известия Томского политехнического университета. – 2008. – Т. 312. – № 5. – С. 114–119.
13. Портал знаний по компьютерной лингвистике. 2008. URL: http://uniserv.iis.nsk.su/cl (дата обращения: 10.01.2011).
14. Загорулько Ю.А., Боровикова О.И. Подход к построению порталов научных знаний // Автометрия. – 2008. – Т. 44. – № 1. – С. 100–110.

Поступила 27.01.2011 г.