ТЕХНОЛОГИЯ ПОСТРОЕНИЯ ПОРТАЛОВ НАУЧНЫХ ЗНАНИЙ: ОПЫТ ПРИМЕНЕНИЯ, ПРОБЛЕМЫ И ПЕРСПЕКТИВЫ

Загорулько Ю. А.
Институт систем информатики имени А. П. Ершова
Сибирского отделения Российской академии наук
проспект Академика Лаврентьева, 6, г. Новосибирск, 630090, Россия
тел.: +7-383-3328359, e-mail: zagor@iis.nsk.su

Аннотация — В докладе рассматривается опыт применения технологии построения порталов знаний, обеспечивающих систематизацию и интеграцию научных знаний и информационных ресурсов, а также содержательный доступ к ним. Особенностью данной технологии является использование для этих целей онтологического подхода. Рассматриваемая технология была применена при создании научных Интернет-порталов по компьютерной лингвистике и археологии, а также прототипа русско-английского тезауруса по компьютерной лингвистики.

I. Введение

В настоящее время в сети Интернет накоплены огромные объемы информации по различным отраслям научных знаний. Причем эти объемы неконтролируемо растет, что делает задачу эффективного информационного обеспечения научных и производственных процессов все более актуальной.

Решение указанной задачи осложняется сложившимися особенностями представления научных знаний и информационных ресурсов в Интернет. Во-первых, эти знания и ресурсы большей частью представлены в виде текстовых документов в электронных архивах научных организаций или специализированных интернет-каталогах и порталах, а иногда и просто размещены на отдельных сайтах как общенаучной, так и технической направленности, что значительно затрудняет их поиск и использование. Во-вторых, проблема организации и поиска необходимой информации усугубляется тем, что для большинства научных знаний, размещенных в Интернет, характерна слабая степень формализации, а большинство научных информационных ресурсов либо вообще не систематизированы, либо их систематизация носит случайный или односторонний характер.

Для решения проблемы эффективного доступа к научным знаниям и ресурсам была предложена концепция специализированного портала знаний [1], основанного на онтологии [2, 3], которая представляет собой согласованную систему понятий, принятых в определенной области знаний. Такие порталы позволяют не только поддерживать систематизацию знаний и информационных ресурсов моделируемых областей научных знаний, но и обеспечивать содержательный доступ к ним. Эта концепция положена в основу развиваемой нами технологии создания и сопровождения порталов научных. Рассмотрению этой технологии и опыта ее применения посвящен данный доклад.

II. Информационная модель портала знаний

Для обеспечения унифицированного представления разнородных знаний и данных, учета их связанности, а также поддержки функциональности портала знаний предложена информационная модель. Такая модель объединяет модели проблемной и предметной (области знаний) областей портала, а также описывает структуру представляемой в его контенте информации. На основе этой модели строятся внутренние хранилища данных системы, организуется его информационное наполнение, навигация и поиск.

Формально информационная модель портала M_P описывается двойкой $M_P = \langle O_P, IC_P \rangle$, 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 формализм, представляющий собой метаонтологию вида: $O = \langle C, R, T, D, A, F, Ax \rangle$,

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;

$D = \{d_1,\ldots,d_l\}$ – множество доменов $d_i = \{s_1,\ldots,s_k\}$, где s_i – значение стандартного типа *string*;

$TD = T \cup D$ – обобщенный тип данных, включающий множество стандартных типов и множество доменов;

A – конечное множество атрибутов, описывающих свойства понятий C и отношений R_A и принимающих значения из T и D;

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 области и области знаний портала и разнообразных семантических связей между ними. Он обеспечивает выстраивание понятий предметной области (ПО) в иерархию «общее-частное» (с помощью отношения R_T) и поддержку наследования свойств по этой иерархии. Особенностью данного формализма является то, что при наследовании от родительского класса его классу-потомку передаются не только все его атрибуты, но и отношения. Другой особенностью предложенного формализма является возможность задания для ассоциативных отношений R_A атрибутов, специализирующих связи между аргументами (объектами).

Вводя формальные описания понятий некоторой области знаний в виде классов объектов и отноше-

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine
 ISBN: 978-966-335-351-7. IEEE Catalog Number: CFP11788

ний между ними, онтология портала задает структуры для представления реальных объектов и связей между ними. В соответствии с этим данные в контенте портала представлены как множество разнотипных информационных объектов (ИО) – экземпляров классов онтологии, связанных между собой отношениями, заданными в онтологии.

На основе предложенной информационной модели не только строится контент портала знаний, но и организуется содержательный доступ к систематизированным знаниям и информационным ресурсам моделируемой области знаний.

III. Методология построения онтологии портала

Как было сказано выше, онтология является ядром информационной модели портала знаний, она служит для представления понятий, необходимых для описания как проблемной области портала, так и его области знаний. В связи с этим построение онтологии является ключевым моментом разработки портала знаний. Поэтому очень важно иметь хорошую и практичную методологию построения онтологии, которая бы облегчала настройку портала знаний и его сопровождение.

Согласно предложенной нами методологии (Рис.1) онтология портала создается на основе базовых онтологий путем их достройки и развития [4].

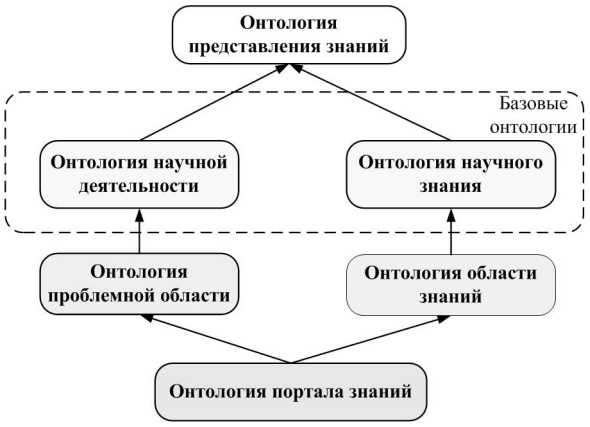

Рис. 1. Система онтологий портала научных знаний.

Fig. 1. System of ontology for a knowledge portal

В качестве базовых были выбраны онтология научной деятельности, которая составляет базис онтологии проблемной области портала, и онтология научного знания, на основе которой строится онтология области знаний портала.

Онтология научной деятельности, выступающая в качестве основы (базиса) онтологии проблемной области портала знаний, фактически, является онтологией верхнего уровня. Она включает базовые классы понятий, относящиеся к организации научной и исследовательской деятельности, такие как *Исследователь, Организация, Событие, Деятельность (Проект), Публикация* и др. В эту онтологию также включен класс *Информационный ресурс*, который служит для описания релевантных моделируемой области знаний информационных ресурсов, представленных в сети Интернет. Этот класс был веден в онтологию научной деятельности, потому что описание информационных ресурсов является весьма важным компонентом контента портала. Набор атрибутов и связей информационного ресурса основан на стандарте Dublin Core [5]. Его атрибутами являются: название, Интернет-ссылка (URL), язык, тип доступа и т.п. Описание ресурса включает экземпляр класса *Информационный ресурс* и набор экземпляров отношений, связывающих его с экземплярами других классов онтологии, представляющих в контенте портала организации, исследователей, публикации, события, разделы области знаний и т.п.

Онтология научного знания фиксирует основные содержательные структуры, используемые для построения предметных онтологий. В частности, эта онтология содержит мета-понятия, задающие структуры для описания понятий конкретной области знаний, такие как *Раздел области знаний (науки), Метод исследования, Объект исследования, Предмет исследования, Научный результат*.

Понятия базовых онтологий связаны между собой ассоциативными отношениями, выбор которых осуществляется не только исходя из полноты представления проблемной области и области знаний портала, но и с учетом удобства навигации по его контенту и поиска информации.

IV. Технология построения порталов научных знаний

На основе рассмотренной выше методологии разработана технология создания порталов научных знаний (см. Рис. 2), ориентированная на экспертов в конкретных областях знаний. Она позволяет им собрать и систематизировать в рамках единого информационного пространства обширные знания и данные в определенной области знаний.

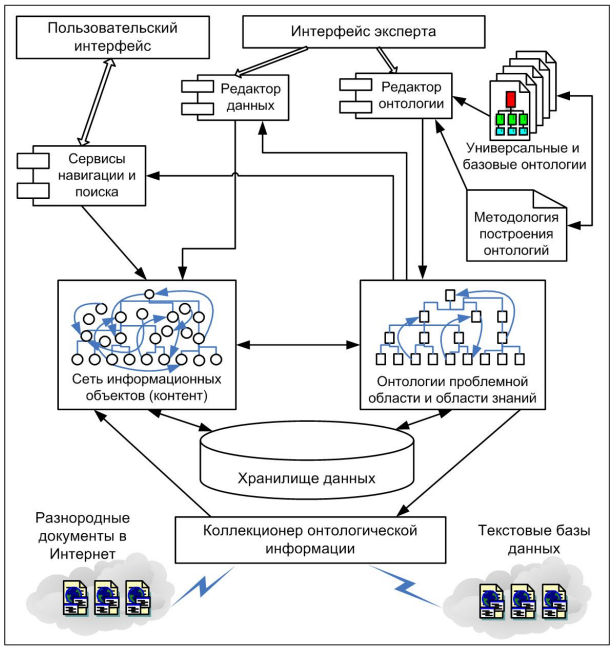

Рис. 2. Технология построения портала научных знаний.

Fig. 2. Technology for construction of knowledge portals

Основными элементами данной технологии являются следующие «знаниевые» и программные компоненты: методология построения онтологий вместе с рассмотренным выше набором базовых онтологий, интерфейс эксперта, обеспечивающий доступ к программным средствам, поддерживающим построение онтологий и управление контентом пор-

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine

тала, коллекционер онтологической информации о ресурсах, а также пользовательский интерфейс, позволяющий осуществлять поиск и навигацию по контенту портала. Предоставляемое данной технологией хранилище данных обеспечивает универсальные структуры для согласованного хранения онтологии и контента портала.

Для построения онтологий и управления ими служит редактор онтологий, реализованный как web-приложение и доступный зарегистрированным пользователям через Интернет. 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 этого требования мы отказались от такого популярного средства построения онтологий как редактор Protégé [6].

Управление информационным контентом портала осуществляется с помощью редактора данных, который позволяет создавать, редактировать и удалять информационные объекты и связи между ними. Формы для ввода конкретных ИО и связей между ними автоматически генерируются на основе онтологии.

Создание контента портала научных знаний – довольно трудоемкая задача. Для ее автоматизации разработана специальная подсистема – коллекционер онтологической информации о ресурсах, который, фактически, выполняет функцию извлечения знаний и данных из сети Интернет. Коллекционер выполняет поиск в Интернете релевантных области знаний портала ресурсов и фиксирование информации о них как об экземплярах понятия *Информационный ресурс* в контенте портала. Последнее состоит в определении значений атрибутов ресурса и задании связей с другими понятиями онтологии портала (организациями, публикациями, событиями и т.д.).

Кроме того, коллекционер решает задачу автоматического наполнения контента портала знаний библиографическими сведениями о научных статьях, собранных в текстовые коллекции или базы данных. На первом этапе для каждой научной статьи строится ее формальное описание, включающее ее основные атрибуты («название», «авторы», «название журнала», «год издания» и т.п.) и список содержащихся в ней библиографических ссылок. На втором этапе полученное описание добавляется в контент портала; при этом обеспечивается согласованность и связанность библиографических сведений о статье с ранее введенными данными.

Содержательный доступ к систематизированным знаниям и информационным ресурсам заданной области знаний обеспечивается с помощью предоставляемых порталом развитых средств навигации и поиска, функционирование которых также базируется на онтологии. Благодаря этому основной сценарий работы пользователя с порталом состоит из выбора либо с помощью средств визуализации онтологии, либо с помощью механизма поиска объектов определенного класса, их просмотра, навигации по их связям (реализациям отношений онтологии) и фильтрации списков таких объектов.

Информация о конкретном объекте и его связях отображается в виде html-страницы (см. Рис.3), формат и содержание которой зависят от класса объекта и заданных для него отношений. При этом объекты, связанные с данным объектом, представляются в виде содержательных гиперссылок, по которым можно перейти к их детальному описанию.

Проекты	
Название деятельности	АОТ
Описание деятельности	Автоматическая обработка текста
Стадия проекта	эксплуатация

Связи объекта

Результат-Деятельности
Научные результаты и продукты
Морфологический анализатор системы Диалинг
Русский морфологический словарь Диалинг

Направление деятельности
Раздел Науки
Автоматическая обработка текста

Ссылки на объект

Персона-Участник-Деятельности	
Персоны	**Роль Участника Деятельности**
Сокирко, А.В.	руководитель

Публикация о Деятельности
Публикации
Сокирко, А.В., DDC - программа поиска по морфологически и синтаксически размеченному массиву, 2003, статья
Сокирко, А.В., Диссертация А.Сокирко «Семантические словари в автоматической обработке текста», монография
Сокирко, А.В., Морфологические модули на сайте www.aot.ru, 2004, статья

Ресурс-Деятельности
Интернет-ресурсы
Сайт Рабочей группы Aot.ru

Рис. 3. Представление информационного объекта.

Fig. 3. Representation of an information object

V. Применение технологии построения порталов научных знаний

Рассмотренная технология была успешно использована в рамках проектов по созданию научных Интернет-порталов по археологии [7] и компьютерной лингвистике [8]. В настоящее время она используется при разработке прототипа русско-английского тезауруса по компьютерной лингвистике.

Разработанный совместно с Институтом археологии и этнографии СО РАН археологический портал знаний был создан для решения задачи систематизации и интеграции накопленных знаний и информационных ресурсов по археологии, а также обеспечения содержательного доступа к ним. Этот портал рассчитан на широкий круг пользователей – от научных работников и преподавателей до студентов и школьников, интересующихся достижениями археологической науки.

Портал знаний по компьютерной лингвистике разработан для организации эффективного доступа к лингвистическим ресурсам. Пользователями такого портала являются как научные работники, преподаватели и студенты, имеющие отношение к этой дисциплине, так и специалисты, разрабатывающие программные системы, предназначенные для обработки текстов, анализа и синтеза речи. В контенте этого портала представлены знания об основных разделах компьютерной лингвистики, о ее предметах и объектах исследования, используемых моделях и методах, а также богатая информация о ресурсах компьютерной лингвистики, к которым относятся технологии, программные продукты, прикладные системы, словари, корпуса и лингвистические БД. В контенте портала также представлена информация о различных аспектах разработки этих ресурсов: организациях, персонах и проектах, с которыми связано их появление, а также о таких их содержательных характеристиках, как отнесенность к разделу науки, объекту или предмету исследования, методам исследования.

VI. Заключение

Применение описанной выше технологии создания порталов научных знаний показало ее продуктивность. Поддерживаемая ею методология построения

 ISBN: 978-966-335-351-7. IEEE Catalog Number: CFP11788

онтологий, главным принципом которой является построение требуемой онтологии на основе базовых онтологий путем их достройки и развития, значительно упрощает создание онтологий порталов знаний и их дальнейшее сопровождение. Предложенный формализм спецификации онтологий является достаточно гибким средством представления различных типов знаний, а построенный на его основе редактор онтологий – весьма удобным и интуитивно понятным для экспертов инструментом представления и систематизации научных знаний.

Способ представления информации в виде сети знаний и данных, построенной на основе онтологии, является более удобным и информативным для пользователя, чем гипертекст или каталог.

При использовании технологии выявились следующие проблемы: сложность подбора экспертов и организации взаимодействия между ними; отсутствие средств содержательного визуального анализа онтологии и контента портала; сложность визуализации иерархий понятий и объектов, построенных по разным основаниям классификации; отсутствие плодотворной обратной связи с пользователями портала; проблема авторских прав.

В связи с этим развитие данной технологии будет проходить в следующих направлениях:

- реализация средств поддержки визуализации иерархий понятий и объектов, построенных по разным основаниям классификации;
- подключение развитых средств графической визуализации, позволяющих представлять в виде графа не только иерархии понятий онтологии, но и содержимое (контент) портала знаний;
- обеспечение возможности компетентным в моделируемой области знаний пользователям расширять контент портала новыми знаниями и ресурсами.

Работа выполнена при финансовой поддержке РФФИ (проект № 09-07-00400).

VII. Список литературы

[1] *Загорулько Ю. А. и др.* Подход к построению порталов научных знаний // Автометрия. 2008. Т. 44. № 1. С. 100—110.

[2] *Gruber T.* Toward Principles for the Design of Ontologies Used for Knowledge Sharing // International Journal of Human-Computer Studies. November 1995. Vol. 43. Issues 5–6. P. 907—928.

[3] *Guarino N.* Formal Ontology in Information Systems // Proceedings of FOIS'98 (Trento, Italy, 1998). Amsterdam: IOS Press, 1998. P. 3—15.

[4] *Загорулько Ю. А. и др.* Технология построения онтологий для порталов научных знаний // Вестник НГУ. Серия: Информационные технологии. 2007. Т. 5. Вып. 2. С. 42—52.

[5] *Using* Dublin Core. URL: http://dublincore.org/documents/usageguide/ (дата обращения: 20.05.2011).

[6] *Protégé*. Web site. URL: http://protege.stanford.edu/ (дата обращения: 20.05.2011).

[7] *Андреева О. А. и др.* Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Труды 10-й национальной конференции по искусственному интеллекту с международным участием (КИИ–2006). Москва: Физматлит, 2006. Т. 3. С. 832—840.

[8] *Боровикова О. И. и др.* Разработка портала знаний по компьютерной лингвистике // Труды 11-ой национальной конференции по искусственному интеллекту с международным участием (КИИ–2008). М.: ЛЕНАНД, 2008. Т. 3. С. 380—388.

TECHNOLOGY FOR CONSTRUCTION OF KNOWLEDGE PORTALS: EXPERIENCE OF APPLICATION, PROBLEMS AND PROSPECT

Zagorulko Yu. A.
A. P. Ershov Institute of Informatics Systems
Siberian Branch of the Russian Academy of Sciences
6, Lavrentiev Ave., Novosibirsk, 630090, Russia
Ph.: +7-383-3328359, e-mail: zagor@iis.nsk.su

Abstract — The paper concerns the experience of application of technology for construction of knowledge portals providing the systematization and integration of scientific knowledge and information resources, as well as the content-based access to them. The main feature of this technology is usage of ontology-based approach for these purposes.

I. Introduction

Recently, a great amount of scientific knowledge and information resources relating to various areas of knowledge has been accumulated in the Internet. However, the access to this knowledge and resources is rather complicated as they are disembodied and ill-structured, or distributed over various Internet catalogues and sites, electronic libraries and archives.

To solve a problem of effective access to scientific information, we have suggested the concept and architecture of a specialized knowledge portal based on ontology which is a consistent system of concepts accepted in a certain field of knowledge. Such portals allow one to provide systematization and integration of knowledge and information resources related to the modeled area of knowledge, as well as the content-based access to them. The paper presents the technology of building knowledge portals and the experience of its application.

II, III, IV. V. Main Part

To support a unified representation of heterogeneous knowledge and data and their connectivity, as well as to provide a required functionality of knowledge portals, the information model was suggested. This model joins the subject domain model (model of area of knowledge) with the problem domain model of the knowledge portal and describes the types of information presented in its content. Based on the information model, the portal internal data base is built and filling of the portal content, navigation through and search in the portal content are organized. The ontology is a core of the information model. The methodology of ontology building, the main principle of which is to build the ontology of the knowledge portal by means of completion and evolution of the basic ontologies was suggested. It considerably simplifies creation and maintenance of such portals.

Based on the methodology considered above, the technology of building knowledge portals oriented to experts in a subject domain was developed. The main components of the technology are a set of the basic ontologies, the methodology of ontology building, expert interface providing access to software tools supporting the ontology building and content management, the ontology information collector which performs extension of the portal content in automatic mode, as well as a user interface allowing one to perform search and navigation through the portal content.

The considered technology was applied for development of the scientific knowledge Internet portals on archeology and computational linguistics as well as prototype of a Russian-English thesaurus on computational linguistics.

VI. Conclusion

The experience of constructing a set of scientific knowledge portals demonstrates the soundness and productivity of the proposed technology. The ontologies developed by means of the technology are a good basis for systematization and integration of the knowledge and information resources, providing convenient navigation through them and content-based search in terms of the modeled area of knowledge.

2011 21st Int. Crimean Conference "Microwave & Telecommunication Technology" (CriMiCo'2011). 12-16 September, Sevastopol, Crimea, Ukraine
 ISBN: 978-966-335-351-7. IEEE Catalog Number: CFP11788