ПОСТРОЕНИЕ ПОРТАЛОВ НАУЧНЫХ ЗНАНИЙ НА ОСНОВЕ ОНТОЛОГИЙ*

Ю. А. Загорулько
Институт систем информатики СО РАН, Новосибирск, Россия
e-mail: zagor@iis.nsk.su

This paper presents an approach to development of specialized Internet portals, which provide content-based access to knowledge and information resources relating to the given field of knowledge. The information basis of the portal is formed by an ontology. This allows us to provide uniform representation of heterogeneous knowledge and data and their connectivity. The ontology serves as a basis for the portal's content management, the semantic search and the knowledge-driven navigation in the information space of the portal.

Введение

В настоящее время накоплен большой объем данных и информационных ресурсов по различным областям научных знаний. Однако эти данные слабо структурированы, плохо систематизированы, рассредоточены по различным Интернет-сайтам, библиотекам и архивам, что существенно ограничивает к ним доступ. Для решения задачи интеграции накопленных знаний и информационных ресурсов заданной области и обеспечения к ним содержательного доступа нами был предложен подход к построению специализированных Интернет-порталов знаний [1].

Как информационный ресурс портал знаний обеспечивает следующие возможности:

— представление информации в заданной научной области знаний, ее составляющих и участников научной деятельности (персоналий исследователей, групп, сообществ и других организаций, включенных в процесс исследования), выполняемой в ее рамках;

— интеграцию доступных информационных ресурсов по данной научной тематике в единое информационное пространство;

— содержательный доступ к систематизированным знаниям и данным, относящимся к определенной научной тематике, т. е. возможность поиска и получения информации в терминах данной области знаний, а также удобную навигацию по его информационному пространству;

— персонификацию пользовательского интерфейса (способа и степени подробности предоставления информации, поиска и навигации по порталу).

В качестве концептуальной основы портала знаний выбрана онтология, содержащая наряду с описанием определенной области научных знаний и выполняемой в ее

*Работа выполняется при финансовой поддержке РГНФ (проект № 07-04-12149).

рамках научной деятельности соотнесенное с ним описание соответствующих сетевых ресурсов. Использование онтологии позволяет обеспечить унифицированное представление разнородной информации и содержательный доступ к ней, делает портал знаний настраиваемым на любую область знаний, облегчает управление его контентом.

1. Онтология портала

Одной из целей онтологии является описание и изучение сущностей, которые имеются в реальном мире и/или сознании человека. Для систем информатики и искусственного интеллекта, в частности порталов знаний, существует только то, что уже в них представлено или может быть представлено, поэтому мы придерживаемся определения онтологии, данного в работе [2], согласно которому онтология является точной спецификацией концептуализации. Здесь под концептуализацией понимается некоторая абстракция, т. е. упрощенное представление мира, построенное для определенной цели. Концептуализация включает объекты, понятия и другие сущности, имеющиеся в рассматриваемой области знаний, а также отношения между ними.

В работах [3, 4] подчеркивается, что онтология — это спецификация концептуализации, но только в той ее части, которая зависит от определенной области интересов. В работе [5] делается упор на то, что онтологии должны помочь в решении проблем, возникающих из-за существования различных интерпретаций одних и тех же терминов. В этой связи онтология рассматривается как соглашение о некоторой области интересов для достижения определенных целей. Для установления соглашения о знаниях, представленных на некотором, в частности логическом, языке, по мнению N. Guariano [6], онтология должна характеризовать концептуализацию, ограничивая возможные значения предикатов и функций.

Таким образом, можно сказать, что онтология представляет собой точное подробное описание (модель) некоторой части мира применительно к конкретной области интересов. Именно такой интерпретации мы и придерживались при построении онтологии портала знаний.

Формально онтология портала представляет собой пятерку вида

$$O =< C, R, A, TD, F >,$$

где C — множество классов, описывающих понятия некоторой предметной или проблемной области; R — множество отношений, заданных на понятиях (классах понятий); A — множество атрибутов, описывающих свойства понятий и отношений; $TD = T \cup D$ — множество типов значений атрибутов, включающее три стандартных типа данных $T = \{\text{string}, \text{integer}, \text{date}\}$ и множество доменов $D = \{D_1, D_2, \ldots, D_n\}$, где D_i — именованный набор элементарных (строковых) значений; F — множество ограничений на значения атрибутов.

С содержательной точки зрения онтология портала служит для представления понятий, необходимых для описания как научной деятельности и научного знания в целом, так и описания конкретной области знаний в частности.

Для упрощения настройки портала на выбранную область знаний в онтологии портала выделены базовые онтологии, независимые от предметной области портала, и предметная онтология, описывающая определенную область знаний. Базовыми онтологиями являются онтология научной деятельности и онтология научного знания.

Онтология научной деятельности является развитием онтологии, предложенной в [7], и включает базовые классы понятий, относящихся к организации научной и исследовательской деятельности, такие как *персона*, *организация*, *событие*, *деятельность*, *проект*, *публикация*, *информационный ресурс*.

Класс *персона* служит для представления субъектов научной деятельности: исследователей, сотрудников, членов организаций и т. п.

Класс *организация* включает понятия, которые описывают различные организации, научные сообщества, институты, исследовательские группы и другие объединения.

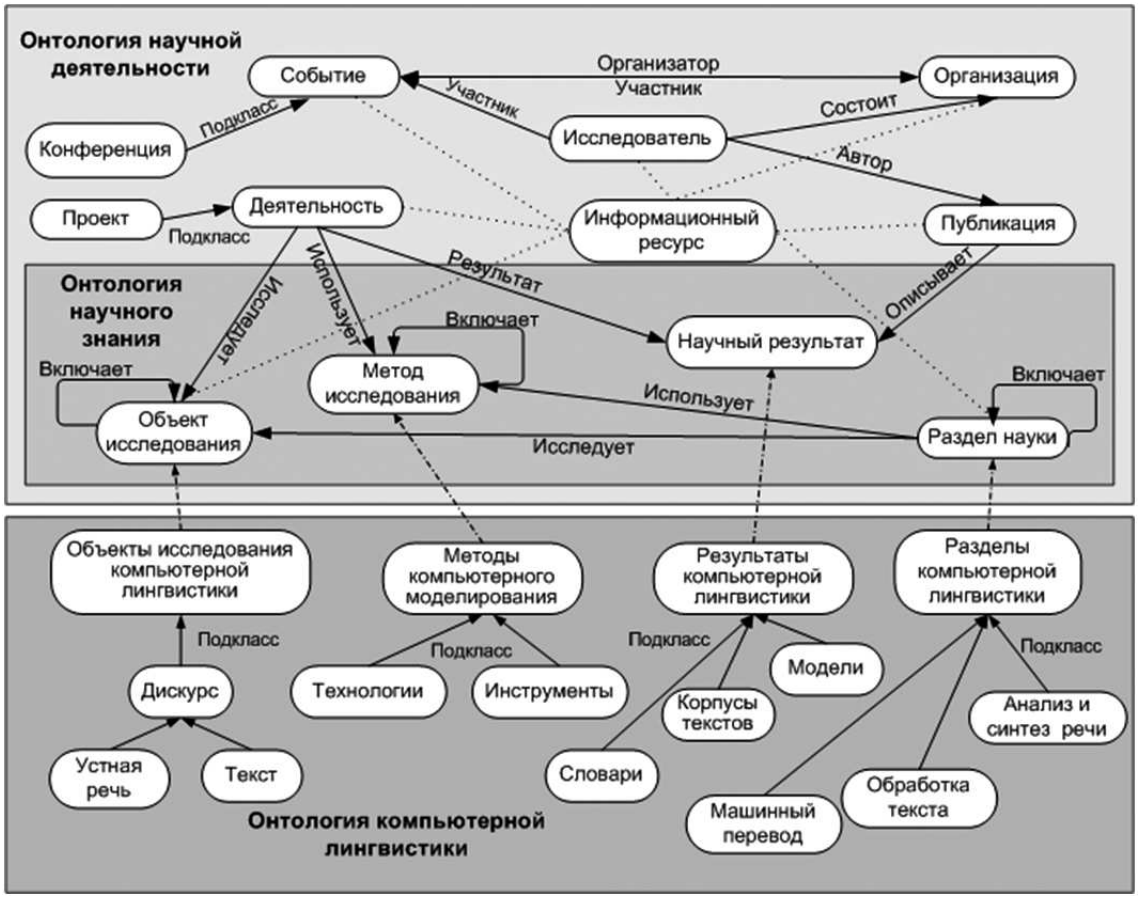
В класс *событие* входят понятия, описывающие такие научные мероприятия, как семинары, конференции, выставки и т. п.

Понятия класса *деятельность* служат для представления научно-организационной и научно-исследовательской деятельности. Они являются связующим звеном между методом и объектом исследования и полученным научным результатом. Класс описывает такие понятия, как *проект*, *программа исследований* и т. п.

Класс *публикация* служит для описания различных типов публикаций и материалов, представленных в печатном или электронном формате (монографии, статьи, отчеты, труды конференций, периодические издания, фото- и видеоматериалы и др.).

Класс *информационный ресурс* служит для описания различных информационных ресурсов, представленных в сети Интернет.

Онтология научного знания фактически является метаонтологией и содержит метапонятия, задающие структуры для описания рассматриваемой области знаний, та-



Фрагмент онтологии портала знаний по компьютерной лингвистике

кие как *раздел науки*, *метод исследования*, *объект исследования*, *научный результат*, позволяющие выделить в данной области знаний значимые разделы и подразделы, задать типизацию методов и объектов исследования, описать результаты научной деятельности.

Понятия предметной онтологии являются реализациями метапонятий онтологии научного знания и могут быть упорядочены в иерархию общее — частное или часть — целое. Так, разработанная нами онтология компьютерной лингвистики [8] включает четыре базовых иерархии: иерархию разделов компьютерной лингвистики как научной дисциплины, иерархию объектов исследования, иерархию методов исследования и иерархию научных результатов (см. рисунок).

Иерархия разделов науки определяет важнейшие направления компьютерной лингвистики, например, такие как *машинный перевод*, *обработка текста*, *анализ* и *синтез речи*. Эти общие направления подразделяются на более частные. Так, машинный перевод включает *автоматический* и *автоматизированный машинный перевод*.

Иерархия методов исследования служит для систематизированного описания инструментов исследования, применяемых в компьютерной лингвистике. Здесь выделяются такие подразделы, как *методы*, *технологии*, *системы*.

Иерархия объектов исследования задает типизацию объектов исследования и структуры для их описания. В качестве базового объекта исследования рассматривается *дискурс* как форма существования и использования языка (языковых единиц различных уровней в их системной взаимосвязи). Здесь учитываются такие формы дискурса, как *текст* и *устная речь*.

Иерархия научных результатов служит для типизации и описания результатов научной деятельности. Она включает такие типы результатов, как *модели*, *словари*, *корпусы текстов*.

Понятия онтологии, а следовательно, и описанные иерархии связаны между собой различными ассоциативными отношениями, выбор которых осуществлялся не только исходя из полноты представления области знаний портала, но и с учетом удобства навигации по его информационному пространству. Наиболее важными из них являются:

— “научное направление” — связывает события, публикации, организации, исследователей и информационные ресурсы с разделами науки;

— “описывает” — задает связь публикации с научным результатом, объектом или методом исследования;

— “использует” — связывает метод исследования с деятельностью, исследователем или разделом науки;

— “исследует” — сопоставляет какую-либо деятельность или раздел науки с объектом исследования;

— “результат деятельности” — связывает научный результат с деятельностью;

— “ресурс” — связывает информационный ресурс с событиями, публикациями, исследователями, методами и объектами исследования.

2. Информационное содержание портала

В соответствии с описанной выше онтологией данные на портале представлены как множество разнотипных информационных объектов и связей, которые в совокупности образуют информационное содержание или контент портала.

Информационный объект (ИО) — это структурированная совокупность данных, представляющая описание некоторого объекта выбранной области знаний. Каждый ИО соответствует некоторому классу онтологии и имеет заданную этим классом структуру.

Между конкретными информационными объектами могут существовать связи, семантика которых определяется отношениями, заданными между соответствующими классами онтологии.

Таким образом, информационное содержание портала включает как знания общего характера (представлены в онтологии), так и конкретные знания о реальных объектах и событиях (такие знания мы называем данными).

Важным компонентом информационного наполнения портала является описание информационных ресурсов. Каждый ресурс соответствует такому понятию онтологии, как *информационный ресурс*, а его описание хранится в базе данных (БД) и включает экземпляр данного понятия и набор экземпляров отношений, связывающих это понятие с другими понятиями онтологии.

Набор атрибутов и связей ресурса основан на стандарте Dublin Core [9]. Его атрибутами являются: *название*, *Интернет-ссылка (URL)*, *язык*, *тип доступа* и др. Ресурс может быть связан ассоциативными отношениями с организациями, учеными, публикациями, событиями, разделами науки и т. д.

3. Настройка портала на область знаний

Настройка портала на заданную область знаний и его информационное наполнение осуществляются с помощью специализированных редакторов (редактора онтологий и редактора данных), реализованных как web-приложения и доступных зарегистрированным пользователям через Интернет.

Редактор онтологии позволяет создавать, модифицировать и удалять любые элементы онтологии (классы понятий, отношения, домены), а также задавать способы визуализации информационных объектов.

При создании класса понятий указываются его уникальное имя и набор атрибутов, служащих для задания различных свойств понятий, фактически описывающих структуру объектов данного класса. Для класса может быть выбран родитель из ранее созданных классов, при этом от родительского класса наследуются не только все атрибуты, но и отношения, а сам родитель связывается с новым классом отношением класс — подкласс.

Для каждого атрибута класса задаются имя, область допустимых значений (тип или домен), количество возможных значений (одно или множество), а также указывается обязательность заполнения.

Домен имеет название и множество элементарных (строковых) значений. Для каждого значения домена может быть указан язык, на котором оно было введено.

Хотя отношения онтологии являются бинарными, они могут иметь собственные атрибуты, уточняющие связь между аргументами, т. е. имеют вид

$$R(\mathrm{Arg1}, \mathrm{Arg2}, \mathrm{Matr}),$$

где R — имя отношения; Arg1 и Arg2 — аргументы отношения (классы); Matr — множество атрибутов, описывающих дополнительные свойства отношения.

Для более удобного представления информации пользователям портала в редакторе онтологии задаются шаблоны визуализации объектов для каждого класса онтологии, а также шаблоны визуализации ссылок на такие объекты.

Наполнение портала осуществляется с помощью редактора данных, который позволяет создавать, редактировать и удалять информационные объекты (объекты введенных в онтологии классов) и связи между ними.

Функционирование редактора данных основано на онтологии портала, поэтому при создании нового информационного объекта прежде всего выбирается соответствующий класс онтологии. Затем по представленному в онтологии описанию класса автоматически создается форма для ввода данных, включающая поля для ввода значений атрибутов объекта. Если атрибут принимает значение из домена, то выводится список его возможных значений.

Одновременно с созданием объекта можно задать его связи с другими объектами, уже существующими во внутренней базе данных портала. Эти связи и их атрибуты определяются соответствующими отношениями онтологии, а форма для их задания автоматически генерируется на основе описаний этих отношений.

Особенность предложенного подхода состоит в том, что портал знаний обеспечивает доступ не только к собственным информационным ресурсам, но и поддерживает навигацию по заранее размеченным (проиндексированным) ресурсам, размещенным в сети Интернет. При этом информация о ресурсах может задаваться вручную экспертами или накапливаться коллекционером онтологической информации [10], т. е. специальной подсистемой портала, осуществляющей сбор, анализ, оценку релевантности, автоматическое индексирование и классификацию Интернет-ресурсов.

4. Организация навигации и поиска на портале

Содержательный доступ к систематизированным знаниям и информационным ресурсам заданной области знаний обеспечивается с помощью предоставляемых порталом развитых средств навигации и поиска.

Для конечного пользователя данные на портале представлены в виде множества связанных информационных объектов. Вся информация о конкретном объекте и его связях отображается в виде HTML-страницы, формат и наполнение которой зависят от класса данного объекта и заданного для него шаблона визуализации. При этом объекты, связанные с данным объектом, представляются на его странице в виде гиперссылок, по которым можно перейти к их детальному описанию. Навигация по данным портала представляет собой процесс перехода от одних информационных объектов к другим по заданным между ними связям. Например, при просмотре информации о конкретной публикации мы можем видеть значения ее атрибутов и ее связи с другими объектами.

Используя представленные связи в качестве элементов навигации, можно перейти к просмотру подробной информации как по прямым связям (об авторах, об описываемом объекте исследования), так и по обратным (об информационном ресурсе, описывающем данную публикацию).

При переходе по конкретной связи любого информационного объекта пользователь может получить достаточно большой список связанных с ним объектов (например, список людей, работающих в некоторой организации). В связи с этим был введен механизм фильтрации списков.

Фильтрация есть способ выборки подмножества информационных объектов из списка путем наложения на него ограничений, т. е. задания фильтра. Фильтр является набором условий, которые определяют допустимые значения атрибутов ИО и требования к существованию связей с другими информационными объектами. Этот метод 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 задания запроса в терминах области знаний портала. Поисковые запросы можно сформулировать через специальный графический интерфейс, генерируемый на основе онтологии портала знаний. При выборе класса искомых информационных объектов автоматически генерируется поисковая форма, в которой можно задать ограничения на значения атрибутов объектов выбранного класса, а также на значения атрибутов объектов, связанных с данным объектом ассоциативными отношениями.

Например, запрос "Найти публикации Юрия Апресяна по машинному переводу за период с 1980 по 1991 г." формально будет выглядеть следующим образом:

```
Класс ‘‘Публикация’’:
       Атрибут ‘‘Дата публикации’’: (>= 1980) & (<= 1991)
Отношение "Автор’’:
           Класс ‘‘Исследователь’’
                  Атрибут ‘‘Фамилия’’ = ‘‘Апресян’’
                  Атрибут ‘‘Имя’’ = ‘‘Юрий’’
Отношение ‘‘Научное направление’’:
           Класс ‘‘Раздел науки’’
                  Атрибут ‘‘Название раздела’’ = ‘‘Машинный перевод’’
```

Заключение

Описан подход к организации содержательного доступа к систематизированным знаниям и информационным ресурсам заданной области знаний с помощью Интернет-портала знаний. Важным преимуществом этого подхода является то, что портал знаний предоставляет пользователю доступ не просто к каталогу ресурсов по данной тематике, а обеспечивает удобную навигацию по сети знаний и данных.

Использование онтологии в качестве концептуальной основы портала позволяет достичь гибкого и целостного представления области знаний и выполняемой в ее рамках научной деятельности. Разделение онтологии портала на предметно-независимые и предметные онтологии делает портал настраиваемым на любую область научных знаний.

На основе онтологии портала автоматически выполняются следующие функции:

— строится схема базы данных портала;

— создаются формы для заполнения БД портала данными;

— генерируются формы поисковых запросов (по классам и отношениям онтологии);

— определяется схема навигации по информационному пространству портала (по отношениям онтологии).

Использование онтологий и других элементов технологии Semantic Web при построении портала знаний позволяет отнести его к разновидности семантических порталов

(Semantic Web Portal) [11], активно разрабатываемых за рубежом. Однако порталы, построенные в рамках предлагаемого нами подхода, имеют ряд особенностей, делающих их действительно порталами знаний. В частности, они имеют управляемые знаниями средства навигации и поиска и, в отличие от обычных семантических порталов, обеспечивают содержательный доступ не только к информационным ресурсам, релевантным заданной области знаний, но и к систематизированным знаниям как о ней самой, так и об организации научной деятельности в ее рамках.

На основе предложенного подхода совместно с Институтом археологии и этнографии СО РАН разработан археологический портал знаний [12], доступный по адресу (`http://www.sati.archaeology.nsc.ru/classarch2/`). Основой для построения предметной онтологии портала послужила предложенная в [13] системная классификация археологической науки, включающая более 500 понятий. В настоящее время археологический портал знаний содержит более 4 тысяч информационных объектов, связанных примерно 15 тысячами онтологических отношений, и продолжает пополняться новыми знаниями и данными.

Список литературы

[1] Боровикова О.И., Загорулько Ю.А. Организация порталов знаний на основе онтологий // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. сем. “Диалог-2002”, Протвино, 6–11 июня 2002 г. М.: Наука, 2002. Т. 2. С. 76–82.

[2] Gruber T.R. Towards Principles for the Design of Ontologies Used for Knowledge Sharing // Intern. Workshop on Formal Ontology, March, Padova, Italy, 1993.

[3] Ushold M., Gruninger M. Ontologies: principles, methods and applications // Knowledge Eng. Review. 1996. Vol. 11, N 2.

[4] Ushold M., King M. Towards a methodology for building ontologies // IJCAI-95, Workshop on Basic Ontologica Issues in Knowledge Sharing, 1995.

[5] Takeda H., Takaai M., Nishida T. Collaborative development and use of ontologies for design // Proc. of the Tenth Intern. IFIP WG 5.2/5.3 Conf. PROLAMAT-98, Sept. 9–12, Trento, Italy, 1998.

[6] Guariano N., Giaretta P. Ontologies and knowledge bases. Towards a terminological clarification // Towards Very Large Knowledge Bases: Knowledge Building and Knowledge Sharing. Amsterdam: IOS Press, 1995. P. 25–32.

[7] Benjamins V.R., Fensel D. et. al. Community is knowledge! in KA2 // Proc. of the KAW-98. Banff, Canada, 1998.

[8] Загорулько Ю.А., Боровикова О.И., Кононенко И.С. и др. Подход к построению предметной онтологии для портала знаний по компьютерной лингвистике // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. “Диалог-2006”, Бекасово, 31 мая–4 июня 2006 г. М.: Изд-во РГГУ, 2006. С. 148–151.

[9] Using Dublin Core. `http://dublincore.org/documents/usageguide/`

[10] Боровикова О.И., Загорулько Ю.А., Сидорова Е.А. Подход к автоматизации сбора онтологической информации для интернет-портала знаний // Компьютерная лингвистика и интеллектуальные технологии: Тр. Междунар. конф. "Диалог-2005". Звенигород, 1–5 июня 2005 г. М.: Наука, 2005. С. 65–70.

[11] Lausen H., Stollberg M., Hernández R.L. et al. Semantic Web Portals — State of the Art Survey. Technical Rep. TR-2004-04-03, DERI (`www.deri.org`), 2004.

[12] Андреева О.А., Боровикова О.И., Загорулько Ю.А. и др. Археологический портал знаний: содержательный доступ к знаниям и информационным ресурсам по археологии // Тр. X Нац. конф. по искусственному интеллекту с международным участием "КИИ-2006". М.: Физматлит, 2006. Т. 3. С. 832–840.

[13] Холюшкин Ю.П., Гражданников Е.Д. Системная классификация археологической науки (элементарное введение в археологическое науковедение). Новосибирск: Изд-во ИДМИ Минобразования, 2000. 58 с.

Поступила в редакцию 11 мая 2007 г.