

Система поддержки деятельности распределенных экспертных групп по разработке онтологий предметных областей

А.В. Завертайлов*, С.П. Ковалев†

Аннотация. В докладе представлены результаты анализа требований, проектирования и реализации системы поддержки деятельности распределенных экспертных групп по построению онтологий предметных областей. В качестве платформы для организации системы используются технологии Grid.

1. Построение онтологий

В рамках проекта поставлена задача исследовать текущее положение дел в области построения онтологий предметных областей и разработать программную систему, автоматизирующую этот процесс. В качестве технологической базы для такой системы целесообразно использовать архитектуру Grid.

1.1. Онтологии

Тематика онтологий в настоящее время становится все более актуальной среди организаций, занимающихся разработкой информационных систем и web-инфраструктур. Например, значительное внимание уделено онтологиям в рамках проекта Semantic Web.

Четкого определения термина “онтология”, которое устраивало бы все заинтересованные стороны, на данный момент нет. В целом, различные организации, проводящие работы в этой области, сходятся в том, что онтологией называется краткое и наглядное описание предметной области, в котором представлены все ее существенные элементы (объекты, процессы, свойства и т. д.) и отношения между ними. Перечислим ключевые сферы приложения онтологий предметных областей:

- *Системный анализ.*
Онтология предоставляет структурированную и частично формализованную основу для проведения системного анализа предметной области, в том числе при формировании спецификаций различных автоматизированных систем.
- *Понимание и изучение предметной области.*
Онтология облегчает понимание описываемой области, что сокращает время, необходимое для ее изучения, и повышает уровень взаимопонимания между участниками экспертной группы.

*Новосибирский государственный университет.

†Институт вычислительных технологий СО РАН.

- *Автоматический поиск и обработка информации.*

Описание существенных элементов содержания предметной области, структурированное в виде онтологии, расширяет возможности автоматической обработки документов, относящихся к предметной области. Важным частным случаем здесь является поиск, основанный на описании семантики требуемой информации.

1.2. Автоматизация процесса построения онтологий

Онтологии предметной области строятся в процессе совместной деятельности групп экспертов в этой области. В настоящее время такая работа выполняется вручную. Для автоматизации этого процесса требуется создать программную систему, поддерживающую совместную согласованную работу распределенных экспертных коллективов. Такая система должна удовлетворять ряду функциональных требований, а также обеспечивать определенный уровень прозрачности в условиях совместной работы географически распределенного экспертного коллектива над созданием концептуально единого документа – онтологии.

Анализ функциональных требований. С точки зрения использования системы выделяются три ключевые роли:

- *Эксперт.*
Эксперт в предметной области, желающий принять участие в создании ее онтологии, подключается к этой работе (либо инициирует ее). Эксперт является основным пользователем системы. Основная задача системы – автоматизация деятельности эксперта и совместной работы распределенных экспертных коллективов.
- *Пользователь.*
В роли пользователя выступает желающий использовать результаты деятельности экспертных групп, получить в свое распоряжение готовые онтологии. Пользователь может обратиться к системе с запросом на предоставление последней стабильной версии строящейся онтологии.
- *Администратор.*
Администратор выполняет весь спектр административных функций, в том числе обеспечивает совместную работу экспертных групп, участвует в документообороте, обеспечивающем их интеграцию, управляет созданием стабильных версий онтологии для публикации.

С точки зрения представления, создаваемая онтология – это комплект документов определенной структуры (см. следующий пункт). Процесс построения онтологии состоит из итераций по дополнению и изменению этого комплекта документов. Для изменения онтологии эксперт вносит предложение на проведение определенных модификаций и тем самым инициирует процесс прохождения такого предложения через подсистему документооборота (рис. 1).

Администратор назначает внесенному предложению рецензентов из числа экспертов. После возможного корректирования, по решению администратора предложенное изменение может быть слито с рабочей версией онтологии либо отвергнуто. На этом итерация процесса эволюции онтологии завершается.

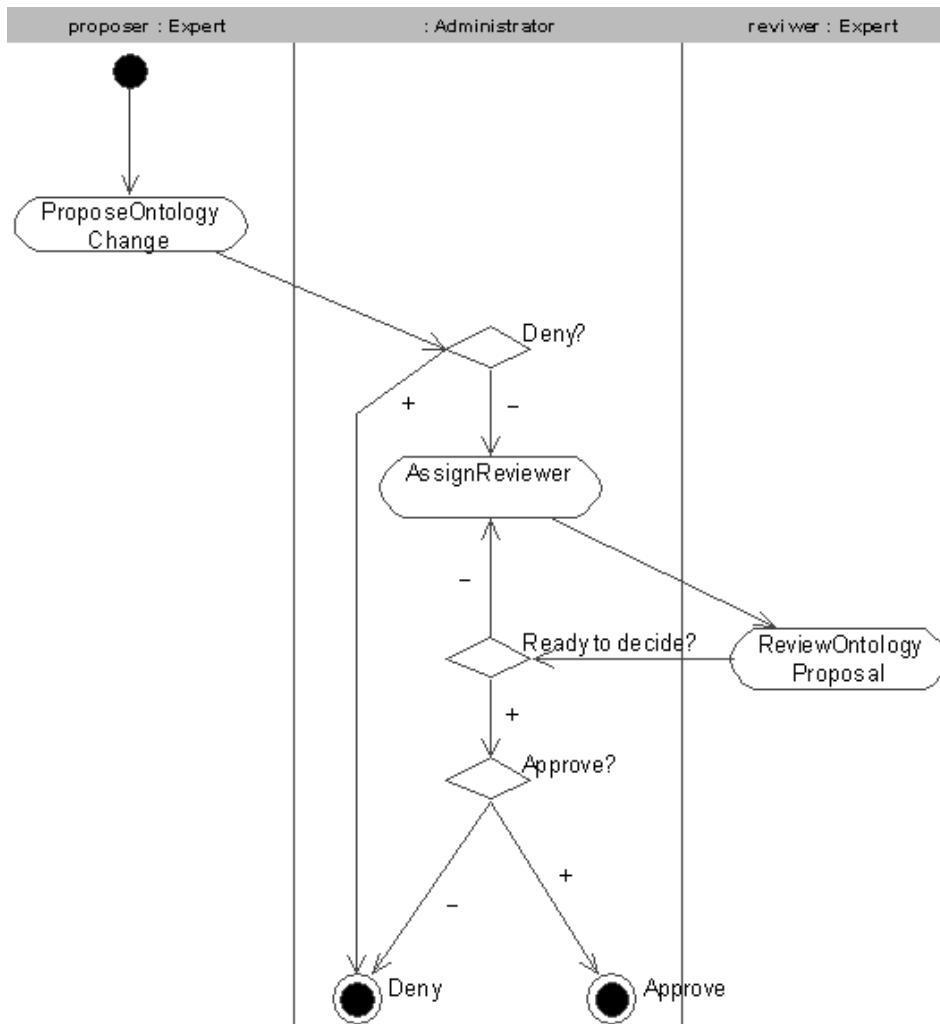


Рис. 1. Документооборот итерации изменения онтологии

По результатам проведения ряда итераций изменения онтологии администратор принимает решение о завершении очередного этапа самого процесса построения онтологии. Результатом такого решения является публикация очередной стабильной версии онтологии. Процесс эволюции онтологии имеет много общего с процессом издания книги – ключевыми точками в этом процессе являются очередные издания, полученные в результате совокупности изменений, внесенных в предыдущее издание. В свою очередь, эти изменения проходят процедуру рецензирования со стороны компетентных специалистов.

Технологические требования. Grid. Описанная задача характеризуется набором специфических требований, определяющих ее технологическую организацию. К таким требованиям относятся:

- *Масштабность.*
К созданию онтологии привлекаются многочисленные экспертные коллективы.
- *Распределенность.*
В разработке концептуально единого комплекта документов онтологии могут принимать участие географически распределенные коллективы экспертов.
- *Динамичность.*
Топология задействованных узлов меняется по мере подключения новых коллективов или прекращения работы старых.
- *Неоднородность.*
Участие коллектива экспертов не должно ограничиваться предъявлением жестких требований к архитектуре его рабочих узлов, их аппаратным платформам и операционным системам.

Адекватную основу для построения систем, удовлетворяющих таким требованиям, предоставляют вычислительные технологии, известные под общим названием Grid [6].

2. Представление онтологии

При разработке представления структуры онтологии были рассмотрены различные существующие на сегодняшний день подходы и стандарты. Как наиболее основательный и перспективный был принят стандарт OWL (Web Ontology Language), разработанный и рекомендованный консорциумом W3C.

2.1. Web Ontology Language

Язык OWL предназначен для такого представления информации, которое отвечает двум, в некотором смысле противоположным, требованиям. С одной стороны, эта информация содержит **знания**, а не только **представление**. С другой стороны, она предназначена для автоматической обработки **компьютерными** программами, в противоположность использованию непосредственно **человеком**.

OWL позволяет:

- формализовать предметную область путем определения классов и отношений;
- определить отдельных представителей выделенных классов и декларировать значения их свойств;
- строить утверждения о классах и представителях.

Именно такая информационная структура – явно выраженные значения терминов словаря предметной области и отношения между описанными терминами – в OWL называется онтологией [3].

OWL обладает большей выразительной силой для достижения указанных целей, чем такие структурные языки как XML, RDF и RDF-S, и может быть представлен в их форме. Например, XML-схема определяет формат представления информации в XML-документе, так что появляется возможность учитывать информационную структуру конкретной предметной области при работе в ней. В то же время OWL-документ позволяет, используя лежащую в основе OWL дескриптивную логику [2], выводить такие факты о сущностях предметной области, которые не содержатся непосредственно в этом документе. Таким образом, OWL-онтология является теорией – совокупностью формальных утверждений, замкнутой относительно выводимости. Для OWL возможно (теоретически) создать приложения, которые будут обрабатывать информацию, представленную в документах, независимо от предметной области, к которой эта информация относится.

Концептуально, документ OWL представляет собой множество декларативных утверждений о сущностях словаря предметной области. Это множество может быть распределено между набором физических документов. OWL делает предположение “открытого мира” [1], в соответствии с которым применимость описаний предметной области, помещенных в конкретном физическом документе, не ограничивается лишь рамками этого документа – содержание онтологии может быть использовано и дополнено другими документами, добавляющими новые факты о тех же сущностях или описывающих другую предметную область в терминах данной.

2.2. Трехуровневая структура онтологии

Исходя из требований к системе поддержки деятельности распределенных экспертных групп, была разработана структура представления онтологии средствами языка OWL, схематически показанная на рис. 2.

Онтологическая структура разделена на три уровня. На базовом уровне находится общая часть структуры любой онтологии. При всем разнообразии подходов к представлению онтологий выделяется некоторая минимальная базовая часть, общая для различных подходов к пониманию онтологий и методик их представления. Это ядро состоит из одного класса и нескольких отношений. Абстрактным базовым классом для всех сущностей предметной области в будущей онтологии является термин (Term). Отношение самого верхнего уровня *Relate* представляет произвольное отношение между терминами. Конкретизированные отношения между наследниками класса *Term* являются наследниками отношения *Relate*, уточняющими его семантику.

На этом уровне ядра онтологической структуры целесообразно выделить также более семантически конкретизированные отношения, широко употребляющиеся при построении моделей любых предметных областей. Исходя из опыта, накопленного специалистами в области объектно-ориентированного проектирования, в качестве таких распространенных отношений выделены агрегация и использование [5].

Ядро онтологической структуры предоставляет основу для построения шаблонов онтологий, отражающих общую терминологическую структуру отдельных классов предметных областей. В шаблоне выделяются категории терми-

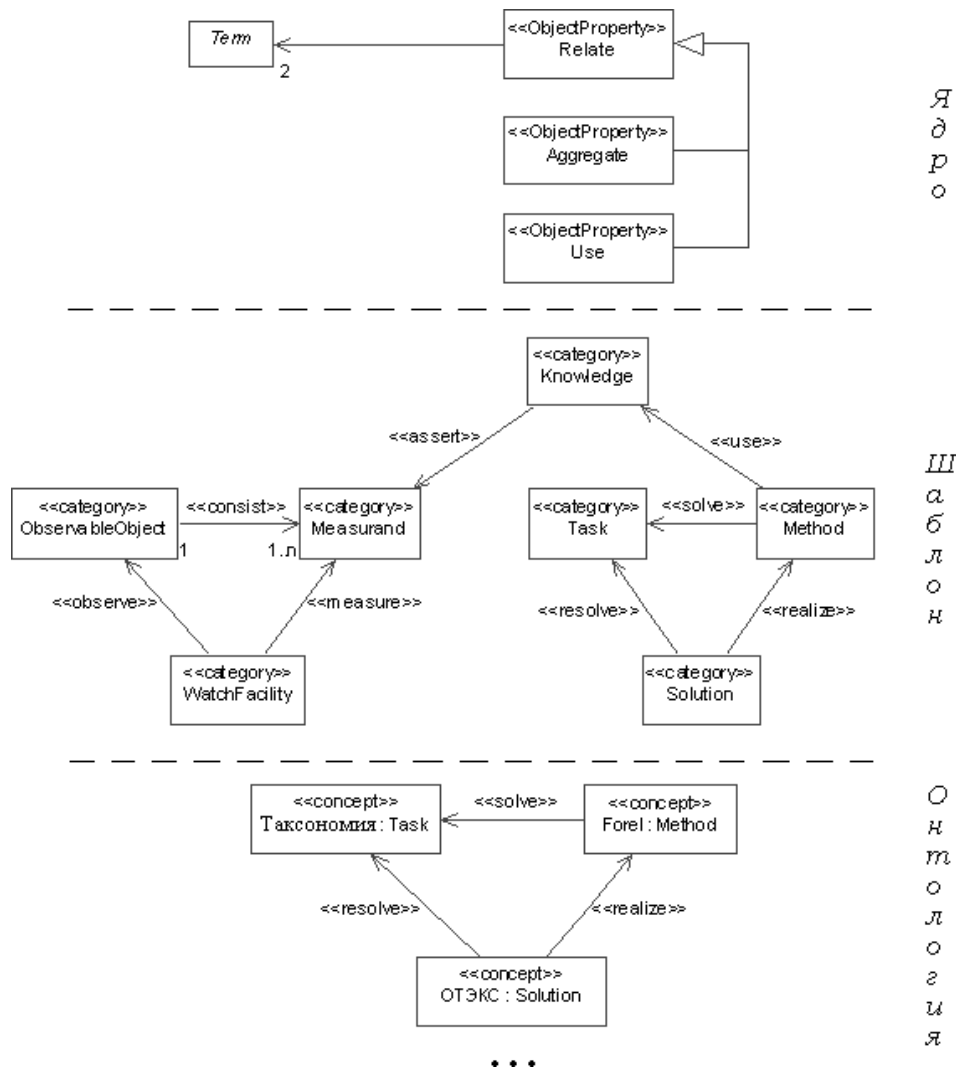


Рис. 2. Структура онтологии

нов, встречающиеся в предметных областях данного класса, и семантически детализированные отношения между этими категориями. При формировании шаблона можно учесть специфику понимания проблематики онтологий, присущую конкретному коллективу экспертов. На рис. 2 представлен шаблон онтологии, разработанный для построения онтологий научно-технических предметных областей, которые связаны с процессами анализа, синтеза и преобразования информации о произвольных фрагментах реального мира. К числу таких процессов относятся измерение и накопление данных, обнаружение закономерностей (знаний), хранение, обработка и передача данных и знаний, использование знаний для прогнозирования и синтеза.

Последний уровень онтологической структуры представляет собой собственно онтологию предметной области, которая строится на базе соответствующей

шего шаблона. В онтологии выделяются конкретные представители категорий (понятия) и задаются отношения между ними. На рис. 2 представлен фрагмент прототипа онтологии предметной области “Распознавание образов и Анализ Данных” [7], разработанной на основе шаблона онтологий научно-технических предметных областей.

Приведем в качестве иллюстрации небольшой фрагмент представления шаблона онтологий научно-технических предметных областей на языке OWL.

```
<owlx:Class owlx:name="Task" owlx:complete="true">
  <owlx:Class owlx:name="Term"/>

  <owlx:Annotation>
    <owlx:Label xml:lang="ru">
      Задача
    </owlx:Label>
    <owlx:Documentation>
      Представляет общие задачи предметной области.
    </owlx:Documentation>
  </owlx:Annotation>
</owlx:Class>

<owlx:Class owlx:name="Method" owlx:complete="true">
  <owlx:Class owlx:name="Term"/>

  <owlx:Annotation>
    <owlx:Label xml:lang="ru">
      Метод
    </owlx:Label>
    <owlx:Documentation>
      Представляет методы решения задач предметной области.
    </owlx:Documentation>
  </owlx:Annotation>
</owlx:Class>

<owlx:ObjectProperty owlx:name="RSolve">
  <owlx:superProperty owlx:name="Relate"/>
  <owlx:range owlx:class="\&Method;"/>
  <owlx:domain owlx:class="\&Task;"/>
</owlx:ObjectProperty>
```

2.3. Представление метаинформации

В дополнение к содержанию онтологии в системе поддерживается формирование и хранение метаинформации об ее элементах. Содержание метаинформации определяется в соответствии со стандартами Dublin Core. Согласно этим стандартам, метаинформация обеспечивает элементы онтологии такими данными, как реквизиты автора и соавторов, время создания и публикации, источники информации и т. д. В качестве представления метаинформации ис-

пользуется XML-представление Dublin Core [4]. Помимо непосредственных преимуществ, получаемых от строгого структурирования метаинформации, такое представление обеспечивает единство представления содержания и метаинформации онтологии.

3. Архитектура системы

Как указывалось выше, требования к системе автоматизации работы экспертной группы по построению онтологий предметных областей могут быть адекватно удовлетворены за счет применения Grid-технологий. Рабочая среда формируется географически распределенной группой экспертных коллективов, ведущих совместную работу над совокупностью физических документов, образующих онтологию. Должна быть обеспечена возможность параллельной работы экспертных коллективов при поддержании целостного состояния и управляемой эволюции создаваемой информационной структуры.

Наиболее развитая открытая архитектура разработки Grid-систем на сегодняшний день предлагается консорциумом The Globus Alliance. К числу последних разработок этого консорциума относится архитектура OGSA (Open Grid Services Architecture) [8], основанная на концепции веб-сервисов.

3.1. Архитектура Grid-сервисов

В системе выделяются следующие ключевые сервисы:

- *Центральный репозиторий.*
Координирует функционирование распределенной сети рабочих сервисов:
 - обеспечивает разрешение ссылок на физические документы, составляющие онтологию;
 - управляет документооборотом, сопровождающим модификацию онтологии;
 - отвечает за публикацию стабильных версий онтологии.
- *Комплект рабочих сервисов.*
Обеспечивает работу локального экспертного коллектива:
 - предоставляет различные виды интерфейса, как для экспертов, так и для специализированных программных средств;
 - поддерживает локальные процессы документооборота;
 - осуществляет хранение рабочих версий частей онтологии, разрабатываемых данным коллективом.
- *Архив.*
Выполняет архивирование устаревших версий онтологии:
 - отвечает за сохранение предыдущих стабильных версий онтологии;
 - управляет процессом назначения версий архивным документам.

Экспертная группа, желающая принять участие в работе над созданием онтологии некоторой предметной области, устанавливает комплект рабочих

сервисов в своем локальном окружении и включается в разработку онтологии, формируя предложения по ее модификации. Хранение рабочих версий создаваемых документов обеспечивается на основе системы CVS (Concurrent Versions System). При интенсивной работе одних экспертных групп над документами, созданными другими коллективами, эти документы могут перемещаться между хранилищами экспертных групп под управлением центрального репозитория. Разрешение ссылок на документы онтологии осуществляется центральным репозиторием, что обеспечивает прозрачный доступ к частям онтологии со стороны рабочих сервисов экспертных групп. Так происходит динамическое формирование целостной распределенной информационно-вычислительной среды, топология которой отражает структуру глобальных потоков информации, порождаемых в ходе исследования рассматриваемой предметной области.

По решению администратора, сервисы центрального репозитория осуществляют публикацию последней целостной версии онтологии в качестве очередной стабильной версии. При этом собирается комплект целостных версий документов, находящихся в локальных хранилищах рабочих сервисов экспертных групп. Предыдущая стабильная версия передается архивному сервису для помещения на хранение в архив.

Таким образом, обеспечивается процесс эволюции онтологии, основанный на управляемых итерациях, с публикацией стабильных версий для использования.

4. Реализация

Описанная структура представления информации и архитектура ключевых Grid-сервисов были успешно апробированы путем создания прототипа. В настоящее время ведутся работы по дальнейшей реализации системы поддержки деятельности распределенных экспертных групп по разработке онтологий предметных областей. К числу приоритетных задач этого этапа относятся:

- Построение полномасштабной онтологии предметной области “Распознавание Образов и Анализ Данных” с привлечением представителей различных школ распознавания образов.
- Завершение создания полнофункционального набора сервисов, поддерживающих процесс построения онтологий.
- Подключение специализированных компонентов, реализующих различные алгоритмы анализа текстов, формирования семантических сетей, семантически осмысленного поиска и др.

Список литературы

- [1] Smith M.K., Welty C., McGuinness D.L. OWL Guide. – W3 Consortium, 2004. – <http://www.w3.org/TR/owl-guide/>.
- [2] Patel-Schneider P.F., Hayes P., Horrocks I. OWL semantics and abstract syntax. – W3 Consortium, 2004. – <http://www.w3.org/TR/owl-absyn/>.
- [3] McGuinness D.L., van Harmelen F. OWL Overview. – W3 Consortium, 2004. – <http://www.w3.org/TR/owl-features/>.

- [4] Powell A., Johnston P. Guidelines for implementing Dublin Core in XML. – DCMI, 2003. – <http://dublincore.org/documents/dc-xml-guidelines/>.
- [5] Буч Г. Объектно-ориентированный анализ и проектирование с примерами приложений на C++: 2-е изд. – М.: Бином; СПб.: Невский диалект, 2000.
- [6] Воеводин В.В., Воеводин Вл.В. Параллельные вычисления. – СПб.: БХВ-Петербург, 2002.
- [7] Загоруйко Н.Г. Прикладные методы анализа данных и знаний. – Новосибирск: Ин-т математики СО РАН, 1999.
- [8] Foster I., Kesselman C., Nick J.M., Tuecke S. The physiology of the Grid. An open Grid services architecture for distributed systems integration // Grid Computing: Making the Global Infrastructure a Reality. – New York: Wiley & Sons, 2003. – P. 217–250.