

Разработка базы знаний в библиотеке КарГТУ с использованием системы проектирования онтологий Protégé-2000

Development of a Knowledge Base Using the «Protégé-2000» System for Ontologies Projecting at Karaganda State Technical University Library

Розробка бази знань в бібліотеці КарДТУ з використанням системи проектування онтологій Protégé-2000

Кочетова О. М.

*Библиотека Карагандинского государственного технического университета,
Караганда, Казахстан*

Olga Kochetova

Karaganda State Technical University Library, Karaganda, Kazakhstan

Кочетова О. М.

*Бібліотека Карагандинського державного технічного університету,
Караганда, Казахстан*

Анализируется ситуация, послужившая отправной точкой для создания базы знаний. Как экспериментальная модель была организована база знаний «Пакет прикладных программ Microsoft Office». Приведены характеристики системы проектирования онтологий Protégé-2000, значительно облегчающие работу в этой среде, как инженерам области знаний, так и операторам. Очерчены перспективы возможности использования данной системы в библиотеке как основы для создания ресурсного центра.

The situation, which became a starting point for a knowledge base creation is analyzed. The knowledge base «Microsoft Office Application Package» was organized as an experimental model. The characteristics of the «Protégé-2000» system for ontologies projecting, which makes the work in this environment easier for both knowledge engineers and operators are presented. The perspectives for using this system at the library as a basis for the Resource Center creation are highlighted.

Аналізується ситуація, яка започаткувала створення бази знань. Як експериментальну модель було організовано базу знань «Пакет прикладних програм Microsoft Office». Наведено характеристики системи проектування онтологій Protégé-2000, що значно полегшують роботу в цій сфері як інженерам області знань, так і операторам. Окреслено перспективи можливості використання даної системи в бібліотеці як основи для створення ресурсного центру

В перечень инструментов современного инженера входит персональный компьютер, и поэтому несколько лет назад в Карагандинском государственном техническом университете (КарГТУ) была выработана образовательная стратегия с широким использованием компьютерной техники. Чтобы обеспечить должный уровень подготовки специалистов в учебном процессе задействовано более 1850 единиц компьютерной техники, большинство из которых подключено в локальную сеть университета.

Требуется отметить, что степень подготовки студентов как пользователей компьютером, находится на разном уровне. Поэтому студентам первого и второго курсов таких специальностей, как строительство, машиностроение, материаловедение, автомобильный транспорт и т.д. преподаются дисциплины «Информатика», «Новые информационные технологии в ...» и т.п. Целью этих, в основном практических, занятий является приобретение студентами навыков свободного оперирования пользовательскими программами.

В репертуар, издаваемой в настоящее время литературы, входит широкий спектр изданий по различным аспектам информационных технологий и средств коммуникации. Эта отрасль является наиболее развивающейся. Изменение в технологии микропроцессоров влечет за собой изменение в программных продуктах, и наоборот. Освоение новых программ требует время на их изучение и внедрение. Для этого издается масса литературы – от инструкций для системных программистов до пособий для пользователей-новичков. Поскольку наш университет выпускает инженеров и бакалавров в области IT, библиотека активно комплектуется изданиями по этой отрасли.

Учебные дисциплины практически ежегодно претерпевают изменения, учебные планы корректируются, меняются списки рекомендуемой литературы. Количество источников, которые студенты должны использовать в течение семестра, неуклонно растёт, и встает вопрос об эффективном распределении времени между процессом разыскания документов и процессом работы с текстами. Современная вычислительная техника, средства телекоммуникации и программное обеспечение позволяют свести к минимуму время первого процесса.

Приходя в библиотеку впервые, большинство студентов не имеют опыта работы с карточными каталогами. Что бы решить эту проблему, с первокурсниками традиционно проводится цикл занятий «Библиотечно-библиографические знания» (по курсу «Введение в специальность»), на которых они получают практические навыки поиска по систематическому и алфавитному каталогам. Учатся они пользоваться и базами данных (БД) электронного каталога, реализованного в среде АБИС ИРБИС.

Библиотека располагает доступом к локальным БД и он-лайновым. Например, наши читатели используют БД «Закон», БД «Технорматив», БД «Патенты Республики Казахстан», коллекцию CD/DVD–дисков EBSCO Publishing, Информационно-поисковую систему «Бизнес-карта» и т.д. В он-лайновом режиме – «Медиатека Кирилла и Мефодия», БД EBSCO Publishing, БД журналов издательского дома Springer, портал VASCODA и т.д. У всех этих программ различные интерфейсы, и самый простой, пожалуй, у медиатеки. Следовательно, пользователю требуется время на ознакомление с интерфейсом и на приобретение навыков поиска в конкретной системе. Причем, чем менее опытный пользователь, тем больше ему требуется времени на это.

Из практики видно, что существенным подспорьем для студентов, имеющих недостаточно опыта в работе с ПК, может стать поисковая система, содержащая документы из рекомендательных списков к лабораторным работам и курсовым проектам. Она должна иметь простой интерфейс, ориентированный на запросы читателей категории «начинающий пользователь». Конечным результатом работы системы в ответ на запросы пользователя будет список рекомендуемой литературы, включающий такие элементы, как библиографическое описание отдельной книги, главы из книги, журнальной статьи, ссылки на сайт в Интернет.

Для формирования подхода к разработке такой системы была создана экспериментальная модель на примере узкой темы «Пакет прикладных программ Microsoft Office» в среде Protégé-2000. (Рис. 1) Protégé-2000 является интегрированным инструментальным средством, которое используется разработчиками систем и экспертами по предметным областям для разработки систем, основанных на знаниях. Приложения, разработанные при помощи Protégé-2000, используются при решении задач и принятии решений в конкретной предметной области.

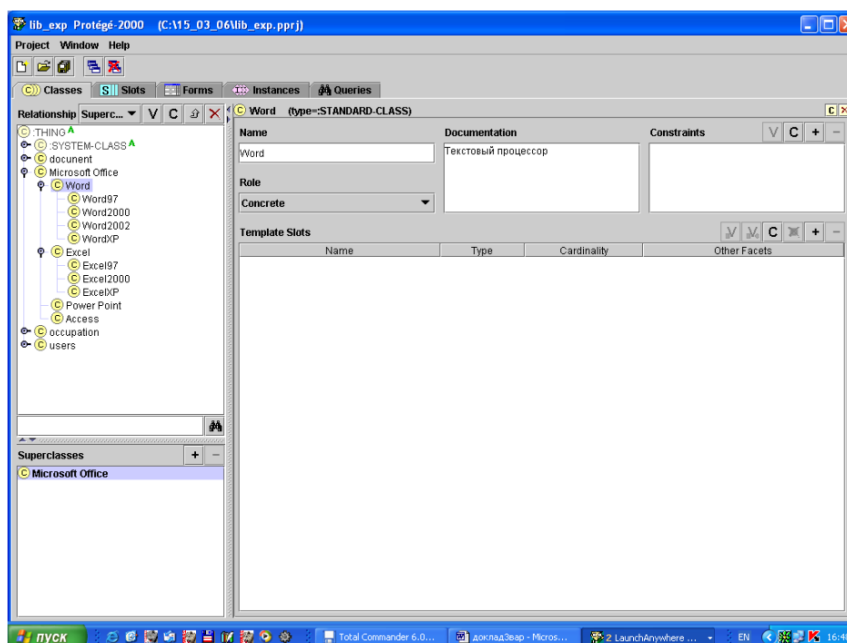


Рис. 1. Интерфейс системы проектирования онтологий Protégé-2000

В действительности, трудно определить, где находятся границы онтологии и база знаний. База знаний образуется из онтологии вместе с набором индивидуальных экземпляров классов. В литературе по искусственному интеллекту содержится много определений понятия онтологии, многие из которых противоречат друг другу. В этой статье онтология – формальное явное описание понятий в рассматриваемой предметной области (классов (иногда их называют понятиями)), свойств каждого понятия, описывающих различные свойства и атрибуты понятия (слотов (иногда их называют ролями или свойствами)), и ограничений, наложенных на слоты (фацетов (иногда их называют ограничениями ролей)).

В центре большинства онтологий находятся классы. Классы описывают понятия предметной области. Класс может иметь подклассы, которые представляют более конкретные понятия, чем надкласс. Например, класс Microsoft Office представляет все программные продукты, входящие в этот пакет. Конкретные версии программ этого пакета – экземпляры этого класса. Текстовый процессор Word 97 – это экземпляр подкласса Word надкласса Microsoft Office. Слоты описывают свойства классов и экземпляров. В качестве слотов выступают заглавие, фамилия автора, место издания, название издательства, название журнала, заглавие статьи из журнала или главы из книги, количество страниц и т.д.

Интегрированная среда Protégé-2000 содержит модель знаний, состоящую из классов информации, слотов, экземпляров и приложений. Средство Protégé-2000 получает доступ ко всем этим частям при помощи унифицированного графического интерфейса пользователя, верхний уровень которого включает перекрывающиеся вкладки для компактного представления частей и удобного их совместного редактирования. Такой дизайн верхнего уровня с вкладками позволяет интеграцию (1) моделирования онтологии классов, описывающей определенную дисциплину, (2) создания средства приобретения знаний для сбора информации, (3) ввода отдельных экземпляров данных и создание базы знаний и (4) выполнения приложений. Онтология определяет набор понятий и их отношения. Средство приобретения знаний разрабатывается специально для предметной области, позволяя экспертам по предметной области легко и свободно вводить свои знания в предметной области. Затем конечная база знаний может быть использована вместе с методом поиска решения задач для ответа на вопросы и решения задач в предметной области. В итоге, приложение является конечным продуктом, созданным, когда база знаний используется для решения задачи конечного пользователя с применением подходящих методов поиска решения задач, методов «эксперт-система» или методов поддержки принятия решений. В нашем случае это поиск по фамилии автора, началу заглавия книги и т.п. в рамках области изданий о продуктах пакета Microsoft Office.

В настоящее время система Protégé-2000 используется в клинической медицине и биометрических науках, хотя ее можно использовать в любой области, в которой понятия можно представить в виде иерархии классов. В нашем случае была сделана попытка использования Protégé-2000 в области библиотечного дела.

На практике разработка онтологии включает:

- определение классов в онтологии;
- расположение классов в таксономическую иерархию (подкласс – надкласс);
- определение слотов и описание допускаемых значений этих слотов;
- заполнение значений слотов экземпляров.

Пройдя все эти этапы, мы создали базу знаний, определив отдельные экземпляры этих классов, введя в определенный слот значение и дополнительные ограничения для слота.

Не существует единственного «правильного» способа или методологии разработки онтологий. Однако имеются общие моменты, которые нужно учитывать, и мы выбрали один из возможных способов разработки онтологии. Подход к разработке онтологии является итеративным. Для начала производим первый черновой просмотр онтологии, затем – проверку и уточнение получаемой онтологии, добавление деталей. Параллельно происходит обсуждение решений, касающихся моделирования, которые должен принять разработчик, а также «за» и «против» и результаты принятия различных решений.

Рекомендуется придерживаться следующих правил разработки онтологии, которые помогут принять проектные решения.

1. Не существует единственного правильного способа моделирования предметной области – всегда существуют жизнеспособные альтернативы. Лучшее решение почти всегда зависит от предполагаемого приложения и ожидаемых расширений.
2. Разработка онтологии – это обязательно итеративный процесс.
3. Понятия в онтологии должны быть близки к объектам (физическим или логическим) и отношениям в предметной области. Скорее всего, это существительные (объекты) или глаголы (отношения) в предложениях, которые описывают предметную область.

То есть, знание того, для чего будет использоваться онтология и насколько детальной или общей она будет, повлияет на многие решения, касающиеся моделирования. Среди нескольких жизнеспособных альтернатив нужно определить, какая поможет лучше решить поставленную задачу и будет более наглядной, более расширяемой и более простой в обслуживании. Нужно помнить, что онтология – это модель реального мира и понятия в онтологии должны отражать эту реальность. После того, как определена начальная версия онтологии, можно оценить и отладить ее, используя ее в приложениях или в методах решения задач и/или обсудив ее с экспертами предметной области. В результате почти наверняка нужно будет пересмотреть начальную онтологию. Этот процесс итеративного проектирования, вероятно, будет продолжаться в течение всего жизненного цикла онтологии.

Начать разработку онтологии можно с определения ее области и масштаба. То есть следует ответить на несколько основных вопросов:

1. Какую область будет охватывать онтология?
2. Для чего мы собираемся использовать онтологию?
3. На какие типы вопросов должна давать ответы информация в онтологии?
4. Кто будет использовать и поддерживать онтологию?

Ответы на эти вопросы могут измениться во время процесса проектирования онтологии, но в любой заданный момент времени они помогают ограничить масштаб модели.

Рассмотрим онтологию, которая будет использоваться для помощи студентам и преподавателям в подборе документов в области «Программы пакета Microsoft Office». В эту онтологию будут включены понятия уровня владения компьютером (начинающий, опытный, профессионал) и профессиональный статус пользователя (студент, бакалавр, магистр, инженер, преподаватель).

Почти всегда стоит учесть, что сделал кто-то еще, и проверить, можем ли мы улучшить и расширить существующие источники для нашей конкретной предметной области и задачи. Повторное использование существующих онтологий может быть необходимым, если нашей системе нужно взаимодействовать с другими приложениями, которые уже вошли в отдельные онтологии или контролируемые словари. Многие онтологии уже доступны в электронном виде и могут быть импортированы в используемую среду проектирования онтологии. В литературе и всемирной паутине существуют библиотеки повторно используемых онтологий. Однако, в нашем случае мы посчитали, что соответствующих онтологий еще не существует, и начали разработку с нуля.

Проект Protégé-2000 включает классы и экземпляры. Вкладка Classes используется экспертами по предметной области и инженерами по знаниям для создания онтологий, то есть, для разработки набора классов, описывающих данную предметную область. Вкладка Instances используется для получения экземпляров, и обычно ей пользуются люди, которые не являются ни экспертами по предметной области, ни инженерами по знаниям. Например, команда специалистов в области прикладного программного обеспечения может тщательным образом построить онтологию. Затем операторы могут ввести экземпляры для описания каждого документа. Несомненный плюс данного проекта в том, что введением данных по каждому отдельному экземпляру можно поручить человеку несведущему в данной области знаний.

Как конечные пользователи вводят экземпляры? Они заполняют формы. Protégé-2000 генерирует исходную форму для каждого класса. Protégé попытается создать начальный, полезный набор форм, основываясь на своих знаниях онтологии. Исходные интерфейсы пользователя, тем не менее, не всегда удобны. Управление формами решает эту проблему. Используя вкладку Forms, разработчик может изменить внешний вид и удобство форм. Формы, которые создаются для своих классов на вкладке Forms, затем используются для ввода экземпляров на вкладке Instances (Рис.2).

Вкладка Forms также позволяет создать несколько интерфейсов пользователя для одной и той же онтологии – формы приобретения знаний могут быть изменены для каждой группы пользователей.

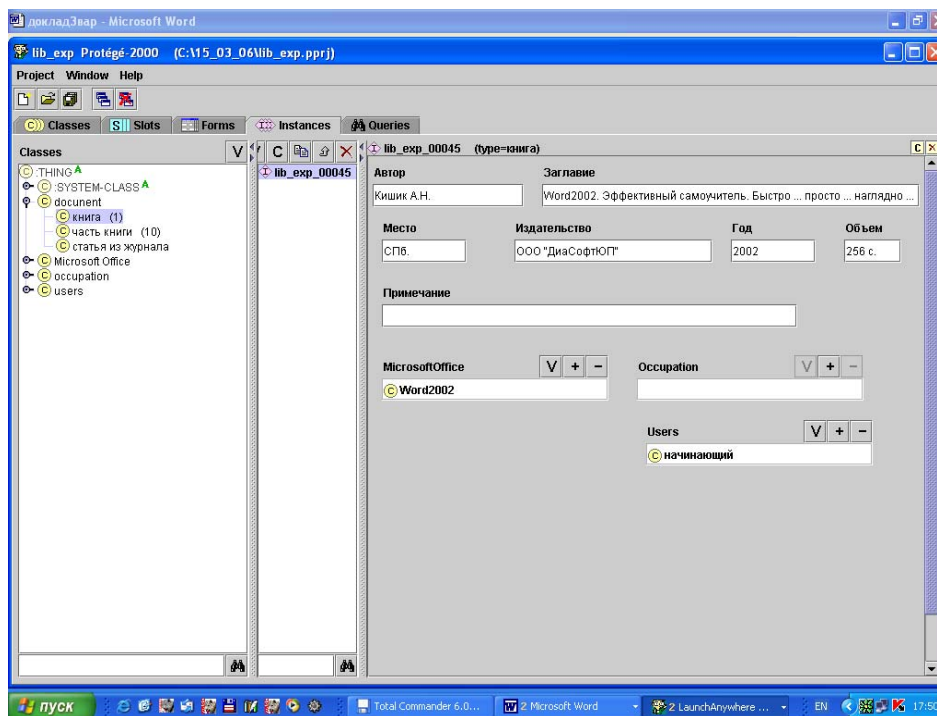


Рис. 2. Заполненные оператором слоты, соответствующие определенному экземпляру

Вкладка Queries позволяет создавать, делать и сохранять запросы, которые позволяют выбрать экземпляры из базы проекта по значению одного или более слотов. Запросы не являются частью базы знаний, а являются способом определения экземпляров в проекте, основываясь на классах и свойствах слотов.

В процессе разработки данной модели и обсуждения её с коллегами возникли идеи создания баз знаний по определенным прикладным задачам. Например, год от года в нашем университете всё большее количество студентов обучается на родном казахском языке. Однако, все они в течение 1-2 курсов изучают русский язык. Выполняя задания, они обращаются к разнообразным источникам за примерами различных стилей речи. Поскольку запросы повторяются согласно программе обучения, целесообразно создать базу знаний, содержащую цитаты из разных литературных произведений, официальных документов и т.д. И постоянно пополнять её новыми экземплярами. Другую базу знаний можно создать, взяв за основу еще одну группу повторяющихся запросов студентов, – это пословицы и поговорки на разных языках по различным темам.

Основное предположение в Protégé-2000 – это то, что системы баз знаний обычно очень дорого строить и поддерживать. Например, предполагается, что разработкой системы, основанной на знаниях, занимается команда, включая как разработчиков, так и экспертов по предметной области, которые могут быть в меньшей степени знакомы с компьютерным программным обеспечением. Protégé-2000 предназначено для того, чтобы вести разработчиков и экспертов по предметной области в процессе разработки системы. Оно позволяет разработчикам повторно использовать онтологии предметных областей и методы поиска решения задач, таким образом уменьшая время, необходимое для разработки и поддержки программы. Несколько приложений могут использовать одну и ту же онтологию предметной области для решения различных задач; один и тот же метод поиска решения задач может быть использован с различными онтологиями.

Опыт построения модели «Программы пакета Microsoft Office» показал, что работать с интерфейсом Protégé-2000 удобно и просто. Если на втором и третьем этапе возникла необходимость вернуться к определению классов и слотов, программа предоставляла возможность их редакции.

Простота в построении запросов, в использовании ранее выполненных запросов и в доступе к результатам предполагает легкое приобретение студентами навыков работы с данной системой.

Итогом реализации высказанных выше идей в стенах библиотеки КарГТУ может стать ресурсный центр, объединяющий различные базы знаний в среде Protégé-2000. Конечными элементами этих баз знаний могут быть не только библиографические описания книг, частей книг и журнальных статей, но и цитаты из изданий разного типа, тексты глав книг и статьи в формате HTML-файлов. На базе Protégé-2000 целесообразно создавать базы знаний, ориентированные на решение прикладных задач и отвечающие запросам, возникающим в течение учебного процесса.