Некоторые проблемы широкого внедрения онтологий в IT и направления их решений

Е.М. Бениаминов[1]

В статье рассматривается современное состояние развития и применения онтологий в информационных технологиях. Особое внимание уделяется проблемам, которые затрудняли широкое внедрение онтологий, внедрение и использование библиотек онтологий. Выделяются направления, которые могут стимулировать массовое внедрение онтотехнологий. К ним относятся применение технологий Web 2.0 при построении библиотек онтологий, использование стандартов Semantic Web и применение открытых языков представлений онтологий, формируемых самими пользователями.

1. Ведение

Системы, основанные на знаниях, – это довольно широкая область в компьютерной науке, в которой складываются собственные методы и принципы и которая оказывает существенное влияние на развитие информационных технологий. Один из существенных принципов, сложившихся в этой области – это разделение декларативных (непроцедурных) и императивных (процедурных) знаний и создание баз декларативных знаний. Тенденция такого разделения в программировании привела к принципам объектно-ориентированного программирования и логического программирования. В базах данных декларативные знания выделяются в виде описании схем баз данных. Особое место базы декларативных знаний получают в связи с развитием Интернет. В 1991 году вводится (Gruber T. R., [1]) термин «онтология» для обозначения связного фрагмента декларативного знания и использования его в информационных технологиях. До этого этот термин использовался в философии.

Много лет назад я занимался алгебраическими моделями баз данных, и мне стало понятно, что схемы баз данных являются элементами особых структур, которые позже были названы онтологиями, и нужны специализированные системы, поддерживающие процессы формирования и отладки многомодульных библиотек онтологий. Я стал заниматься приложением математической теории категорий к моделированию онтологий и разработкой принципов построения системы формирования и отладки онтологий [2]. Десять

[1] *125993, ГСП-3, Москва, Миусская площадь, д. 6, РГГУ, ebeniamin@yandex.ru, http://beniaminov.rsuh.ru*

лет назад Леонид Андреевич Калиниченко любезно указал мне на систему Ontolingua – первую систему в Web для работы с онтологиями. С тех пор я с большим интересом слежу за этой темой [3].

Целью этой статьи является охарактеризовать состояние и развитие систем онтологий в Веб, начиная с 1995 года, постараться определить некоторые причины трудностей внедрения и использования таких систем и определить некоторые направления развития и преодоления трудностей.

2. Общая характеристика онтологий

Определение онтологии было дано Т. Грубером [1]. Здесь я не буду его повторять, а выделю лишь некоторые свойства онтологий, существенные для дальнейшего изложения.

- Онтологии представляют собой спецификации на формальном языке, в которых фиксируются договоренности группы специалистов о том, что как называется в их области и каким свойствам (соотношениям) удовлетворяет.
- На логическом уровне каждой онтологии соответствует некоторая теория (сигнатура+аксиомы), а иногда и некоторая фиксированная модель (множества+операции+отношения). Вопросы к онтологии интерпретируются как запросы к соответствующей ей теории (модели).
- Онтологии, как правило, строятся по модульному принципу: при определении новой онтологии могут использоваться уже ранее построенные онтологии.
- Онтологии должны быть удобны для понимания специалистами и интерпретироваться системами при использовании.

Простейшими, но распространенными и очень важными примерами онтологий, являются системы классификаций. Всем известно значение классификационных систем К. Линнея в биологии и Д. Менделеева в химии. Более того, имеется представление, что в любой области знаний имеется своя классификационная система, которая лежит в ее основе. Естественно, эти классификационные системы довольно устойчивы, но все-таки развиваются по мере развития наук.

Мне посчастливилось быть участником междисциплинарной конференции «Первой Всесоюзной школы-семинара по методологии и теории классификации» 25-31 октября 1979 г. в поселке Борок, где были широко обсуждены проблемы и принципы классификаций в различных науках. Вопросы, поднятые на этой конференции, и энтузиазм ее участников ярко описаны в [4].

Сейчас в области информационных технологий, связанных с онтологиями, либо мало говорят о принципах классифицирования (хотя в некоторых работах, связанных с системой Cyc [5], [6] есть, например, упоминание о машинном обучении), оставляя эти вопросы специалистам в предметных областях, либо навязывают конкретные системы классификаций. При этом разрабатываются графические редакторы, делающий процесс построения систем классификаций

более наглядным, и разрабатываются стандарты машиночитаемых документов по классификациям. Теперь, все больше, классификационные системы используются компьютерными программами при решении различных задач.

Большой интерес представляют более сложные онтологии, которые разработчики системы Cyc [5], [6] называют микротеориями. В общем случае, в онтологии задаются имена классов, имена свойств, типы значений свойств, некоторые экземпляры классов, функции (операции) и отношения между классами и элементами, а также аксиомы, связывающие элементы онтологий. Сложные онтологии строятся по модульному принципу и разрабатываются коллективами специалистов в предметных областях. Поэтому в системах, поддерживающих разработку сложных онтологий должна в какой-то степени поддерживаться многоверсионность, тестирование и отладка онтологий. Естественно, при этом возникают библиотеки (базы) онтологий, и необходимы специальные средства для работы с этими библиотеками.

В онтологии представляется фрагмент взгляда на некоторые контексты миров, представляющие интерес для специалистов в данной предметной области. Естественно, что некоторые онтологии из одной библиотеки онтологий могут быть несовместимыми между собой, так как соответствуют разным ситуациям (объединение всех онтологий библиотеки не образует непротиворечивую теорию). Функция библиотеки онтологий – предоставить удобные модули и среду для формирования онтологий конкретных задач или приложений.

Примерами сложных онтологий являются онтологии, описывающие схемы баз данных. Технология работы со сложными онтологиями может использоваться при формировании сложных схем баз данных, согласовании совместной работы нескольких баз данных, при создании распределенной базы данных.

Процесс тестирования и отладки онтологий предполагает, что в системах формирования онтологий должен обеспечиваться некоторый удобный для понимания специалистом способ представления данных об онтологии или следствий введенных в онтологии предположений. В общем случае, это может быть либо язык запросов, либо набор конкретных запросов, ответы на которые программно формируются и представляются пользователям. При этом следует иметь в виду, что теории, соответствующие онтологиям, как правило, неполны, то есть имеются утверждения, истинность или ложность которых не следует из аксиом, введенных в онтологии.

3. Основные примеры серверов онтологий и систем, использующих онтологии в Веб

3.1. Система Cyc

Cyc — это закрытый проект по созданию объёмной онтологической базы знаний, позволяющей программам решать сложные задачи из области искусственного интеллекта. Автор - Дуглас Ленат [7]. Начало разработки - 1984 г. На текущий момент база знаний Cyc содержит 2,2 миллиона утверждений (фактов и правил), описывающих более 250 тысяч термов, включая почти 15000 предикатов [8]. Модули представлены в виде микротеорий. Имеется открытый фрагмент онтологии Cyc (OpenCyc [9]) и его представление в Web [10]. Более подробный обзор этой системы приведен в [11], в котором отражены история проекта Cyc, использование системы, Cyc и концепция Web 3.0, язык CycL и критика системы.

3.2. Система Ontolingua

Web-сервер Ontolingua для хранения онтологий и межмашинного обмена онтологиями разработан в 1995 г. лабораторией KSL Стэнфордского университета. Имеется большая библиотека онтологий в открытом доступе для произвольных пользователей на станице [12]. Интересные демонстрационные примеры применения системы указаны на странице: [13]. В этих примерах показывается, как строятся онтологии задач на основании библиотек онтологий из различных областей знаний.

Об этой системе я подробнее писал в статье [3], поэтому здесь упомяну лишь о существенных изменениях, произошедших за последнее время.

Одна из функций системы Ontolingua – это интеграция с другими системами представления знаний (в частности, с Cyc). Эта функция выполняется подсистемой OKBC (Open Knowledge Base Connectivity) на базе языка межмашинного обмена знаниями KIF (Knowledge Interchange Format). В последнее время в качестве стандарта языка обмена онтологиями в Ontolingua используется и язык OWL, предложенный в проекте Semantic Web.

3.3. Проект Semantic Web (Web 3.0)

Проекту Semantic Web посвящен доклад В.Ф. Хорошевского на данном симпозиуме, поэтому здесь этот проект подробно не обсуждается. Начало проекта относится к 2001 году. Следует отметить, что работы над этим проектом активизировались в последнее время, возможно, в связи с поддержкой DARPA. В рамках этого проекта в развитие принципов языка XML для представления онтологий были разработаны языки RDF, OWL, язык запросов

SPARQL и язык правил SWRL. Эти языки становятся стандартами для межмашинного обмена онтологиями и работы с онтологиями.

3.4. The World FactBook

The World FactBook – пример распределенной базы данных в Web [14], использующей онтологии. Данные в системе The World FactBook формируются Central Intelligence Agency US для правительства США на основании различных источников и баз данных. При интеграции баз данных используются онтологии. В The World FactBook представлена географическая, демографическая, историческая и экономическая информация о странах мира.

3.5. Системы, поддерживаемые DARPA

Многие системы, работающие с онтологиями (включая перечисленные ранее), поддерживаются DARPA и созданы благодаря финансированию из этого ведомства в больших размерах.

Система Cyc в последнее время частично открывается и переводится на коммерческую основу.

Особое внимание DARPA уделяется обеспечению взаимодействия систем в Интернет и стандартам межмашинного взаимодействия (KIF, OWL).

3.6. Другие примеры разработок онтологий

Онтологии верхнего уровня [15]. Примеры:

- Descriptive Ontology for Linguistic and Cognitive Engineering (DOLCE);
- General Formal Ontology (BFO);
- WordNet;
- Suggested Upper Merged Ontology (SUMO).

Онтологии верхнего уровня – это особая тема. Под онтологией верхнего уровня специалистами, занимающейся этой темой, понимается самая общая онтология, которая является общей для всех областей знаний. Считается, что такая онтология существует, и разные группы специалистов предъявляют свои варианты. Общие онтологии (онтологии верхнего уровня) связаны с мировоззрением, и, естественно, эта область близка к исследованиям в философии и лингвистике. Возникновение нескольких онтологий верхнего уровня и конкуренция в этой области означает, что создание онтологий верхнего уровня не простой вопрос, и требуются удобная (модульная) организация онтологий, ясное представление онтологий, специальные средства, поддерживающие процесс согласования онтологий. Становится очевидным, что разработка онтологий верхнего уровня – это не разовая акция, а постоянный процесс, и нужно иметь

инструменты, поддерживающие этот процесс и коллективную работу над онтологиями.

Специализированные онтологии. Примеры:

Список ссылок на онтологии, разработанные с помощью Protégé, приведен в [16]. Имеется большое количество отологий и их применений в биологии и медицине.

Онтологии в корпоративных системах [17].

В больших (распределенных) корпоративных системах онтологии могут использоваться в трех целях:

- для унификации ведущихся в корпорации документов и сбора на их основе данных для ввода в базу данных корпорации;
- для представления и организации метаинформации в системах типа «хранилища данных» с целью использования ее при формировании запросов для экономического анализа данных работы корпорации;
- для ведения, поиска и организации нормативно-справочной информации.

В некоторых больших российских корпорациях, например, «Интегра», «Татнефть», «Норникель», «Сибур», ТНК-ВР уже созданы и используются онтологии для нормативно-справочной информации (фирма проектировщик: НЦИТ ИНТЕРТЕХ, система ONTOLOGIC).

Онтологии и СУБД.

В некоторых системах управления базами данных, например ORACLE [18], вводятся специальные средства для работы с онтологиями и для их использования в языке запросов.

3.7. Инструменты для работы с онтологиями

В этом подразделе хотелось бы обсудить инструменты, с помощью которых строятся онтологии, в частности, в виде OWL файлов. Таких инструментальных систем становится все больше с разными претензиями на удобство и универсальность.

Одна из первых разработок в этой области – это система Protégé [19] с большим опытом применения. Судя по тому, как строятся многие онтологии в виде файлов, Protégé является наиболее распространенным инструментом. Система Protégé разработана в лаборатории KSL Стэнфордского университета. Первоначально она разрабатывалась как программное инструментальное средство для формирования словарей в области медицины, но оказалась полезной для применений и в других областях. Protégé 2000 разработана уже для работы в Web-браузерах. В настоящее время с ее помощью читаются и формируются OWL-файлы. На конференции в Будапеште (июль 2007г.) [20] определены проблемы и некоторые направления развития Protégé.

Другая система, Chimaera [20], (также разработка подразделения KSL Стэнфордского университета) предназначена для программной поддержки процесса объединения больших онтологий. Это графический редактор, который выделяет сомнительные места в объединенной онтологии и позволяет редактировать онтологию.

4. Проблемы формирования и использования библиотек онтологий

Пятнадцать лет – большой срок в области информационных технологий, и за эти пятнадцать лет технологии построения и использования онтологий становятся яснее. Однако темпы внедрения онтотехнологий все-таки медленны. Главная причина такого замедления является то, что онтологии должны строиться высоко квалифицированными специалистами в своей области, а языки представления онтологий являются сложными, техничными и далекими от этих областей знаний.

Для формирования простейших онтологий в виде классификаций были построены графические редакторы, которые упрощают работу с такими онтологиями и делают их более наглядными. Это определило активность построения классификационных онтологий во многих областях знаний. В свою очередь, когда специалисты стали активно строить классификационные онтологии, у некоторых специалистов стали появляться потребности и в более сложных онтологиях, в библиотеках онтологий, в новых методологиях их построения. Так, например, появились работы [22] об использовании алгебраического языка спецификаций CASL в проекте DOLCE.

Другая причина отсутствия массового использования онтотехнологий в Веб в настоящее время состоит в том, что массовый пользователь не видит непосредственного эффекта от использования онтологий, а от него эти технологии требуют больших усилий по семантической разметке той информации, которую он выставляет в Веб. Поэтому для преодоления этого барьера нужно разработать Веб-среды и инструменты, в которых пользователи смогут создавать собственные семантически размеченные страницы и языки запросов к ним с тем, чтобы пользователи сами могли создавать системы с новой функциональностью в виде информационных систем.

Движение в этом направлении предложено в проекте Semantic Wiki [23].

Понимание проблем приходит с опытом. Ниже перечисляются некоторые, как мне кажется, важные положения современного состояния разработки библиотек онтологий и использования онтологий.

Проблемы формирования и использования библиотек онтологий:

1. Так как онтология есть фиксация в формальном виде договоренностей группы специалистов в определенной области о системе используемых ими понятий, их свойствах и аксиомах, то каждая система онтологий имеет смысл только для группы людей, принимающих эти

договоренности (социальный характер онтологий). Поэтому должна быть обеспечена возможность формирования онтологий для различных групп специалистов.

2. Так как науки и представления в областях знаний меняются, то в компьютерных системах онтологий требуются средства поддержки целостности и версионности онтологий при изменениях и постепенном накоплении онтологий.
3. Так как в онтологиях фиксируются договоренности специалистов, представлять онтологии должны специалисты в предметных областях. Поэтому язык представления онтологий должен быть удобен для этих специалистов. Заметим, что в каждой области знания при формировании понятий этой области формируются и специализированные языки. Поэтому язык представления онтологий должен быть открытым для пользователей. При этом внутреннее представление онтологий для компьютерного использования и межмашинного обмена должно быть стандартизованным.

Проблемы реализации:

1. Большие онтологии и большие библиотеки онтологий требуют разработки специальных средств их ведения.
2. Формирование сложных систем онтологий требует соответствующих средств опробования и отладки онтологий.
3. Для сложных онтологий полностью отделить непроцедурные и процедурные знания не удается (эффективность использования онтологий, прагматика).
4. Модульность построения библиотек онтологий. При этом следует учитывать контекстность онтологий (их возможную взаимную противоречивость), целевое создание и многоцелевое, многоразовое использование.
5. Проблема интеграции онтологий, представленных на разных языках в разных логиках и моделях.

Направления преодоления трудностей формирования и использования больших библиотек онтологий:

1. Использование Web 2.0-технологий в Web для создания социальных сетей и сред в Web, наполняемых самими пользователями (яркий пример – Wikipedia).
2. Построение систем с открытым языком пользователя и стандартным языком внутреннего представления онтологий.
3. Предоставление пользователям Web, при формировании своих страниц, удобных средств модульного (с использованием чужих модулей) формирования внутреннего (семантического) представления данных своих страниц и **языка запросов к странице** (новые сервисы).
4. Алгебраический подход к моделированию онтологий, как путь выработки принципов интеграции разнородных онтологий.

5. Wiki-технологии для формирования и использования библиотек онтологий

Wikipedia – это грандиозное достижение современности. Ее развитие происходило в русле основных принципов Интернет и Веб. Основной принцип Интернет – это распределенность ресурсов и отсутствие единого центра управления. Основной принцип Веб – наполнение содержания сети происходит самими пользователями сети Интернет. Оба эти принципа выдерживаются в Wikipedia и дополняются принципами демократической самоорганизации людей, наполняющих Wikipedia содержанием, а также возможностью использования любой страницы Wikipedia, как шаблона страницы для вашей статьи.

В чем достоинства технологий Wikipedia для создания и использования библиотек онтологий? Первый и очевидный – это социальный характер Wikipedia. Второй, и очень важный, – это возможность семантически разметить только страницы-шаблоны, освобождая остальных пользователей от тяжкой работы семантической разметки своих страниц. Например, можно создать шаблон для страницы «Person» с соответствующими полями и конкретную страницу некоторой персоны. Другие пользователи, используя эту страницу и редактируя ее, заводят информацию о многих других персонах. Кто-то может завести страницу о конкретном правителе России, добавив на страницу дополнительные семантически размеченные поля (например, годы правления и тип правления). Кто-то, уже используя эту страницу, заведет страницу о конкретном царе из династии Романовых, введя дополнительное поле «династия» и ее значение. Далее, на основе этого шаблона могут быть заведены страницы других Романовых.

Так в Wikipedia могут появиться семантически размеченные страницы. Но эта разметка бессмысленна, если нет языка запросов, для ответов на которые используется данная семантическая разметка.

В качестве языка запросов низкого уровня может использоваться язык SPARQL Query Language for RDF [24], предлагаемый в проекте Semantic Web. На основе этого языка уже может формироваться язык более высокого уровня для пользователей. Пример такого языка имеется в Semantic MediaWiki [25].

Более того, на основе онтологии «Person» в Wikipedia можно создать страницу «Родственные отношения», в которой будут введены шаблоны запросов «Брат», «Сестра», «Дядя», «Теща» и т.д. с соответствующими формулами запросов. Аналогично, можно создать страницу «Престолонаследник» с соответствующим шаблоном и формулой запроса. Наконец, мы можем создать страницу «Династия Романовых», в которой будет присутствовать стандартная текстовая и графическая информация, а часть страницы будет представляться ответом на запрос, в виде таблицы (или дерева) обо всех персонах, страницы которых есть в Wikipedia, и относящихся к роду Романовых. При этом может быть загружены онтологии страниц не только из этих страниц, но и онтологии страниц «Родственные отношения», «Престолонаследник». Так будет сформирована онтология страницы «Династия Романовых», и на ней будут доступны шаблоны запросов страниц «Person», «Родственные отношения», «Престолонаследник», комбинируя которые

пользователь может строить сложные содержательные запросы к странице «Династия Романовых». На этой странице могут быть приведены примеры сложных и простых запросов, по аналогии с которыми пользователь может сформулировать свои запросы.

6. Выводы

- Онтосистемы и онтопроекты создаются и развиваются уже более 10 лет. Успех и значимость этого направления очевидны.
- Однако, темп внедрения онтотехнологий все еще невелик. Пока практические успехи получены при финансовой поддержке государственных органов, либо внутри больших корпораций.
- Для широкого внедрения онтотехнологий предлагается строить онтосистемы с использованием следующих трех принципов.

Три принципа построения новых баз онтологий

1. Онтологии строятся в стиле Wikipedia с поддержкой модульности, коллективной работы, версий и системы согласований.
2. В системе поддерживается среда открытого языка работы с онтологиями, который формируется самими пользователями, по мере пополнения базы онтологий.
3. Вместе с текстом онтологии в системе формируется внутреннее представление онтологии, которое используется при семантическом анализе выражений языка, при формировании ответов на запросы к онтологии и ее отладке, при межмашинном обмене онтологиями в некотором стандарте и при использовании онтологий в приложениях.

В январе 2008 года мной сформирован открытый проект http://ezop-project.wiki.sourceforge.net/ по созданию макета такой системы на основе перечисленных выше принципов построения. Приглашаю всех желающих к обсуждению проекта и к участию в нем.

Литература

1. Gruber T. R. The role of common ontology in achieving sharable, reusable knowledge bases. In J. A. Allen, R.Fikes, and E. Sandewell, editors, Principles of Knowledge Representation and Reasoning – Proceedings of the Second International Conference, pp. 601-602. Morgan Kaufmann (1991)
2. Бениаминов Е.М. Основания категорного подхода к представлению знаний. Категорные средства. // Изв. АН СССР Техн. кибернетика, №2, 21-33 (1988)
3. Бениаминов Е.М., Болдина Д.М. Система представления знаний Ontolingua - принципы и перспективы // НТИ. Сер.2. № 10 (1999)

4. Кожара В.Л. Классификационное движение. // Институт биологии внутренних вод им. И.Д. Папанина РАН, Борок, 2006, http://ibiw.ru/win/kd2.pdf
5. Matuszek C., Cabral J, Witbrock M., DeOliveira J. An Introduction to the Syntax and Content of Cyc, http://www.cyc.com/doc/white_papers/AAAI06SS-SyntaxAndContentOfCyc.pdf
6. Википедия о системе Cyc, http://en.wikipedia.org/wiki/Cyc
7. Википедия о Дугласе Ленате (авторе проекта Cyc), http://en.wikipedia.org/wiki/Douglas_Lenat
8. Официальный сайт компании Cycorp, http://cyc.com/
9. OpenCyc – открытый фрагмент онтологии Cyc, http://www.opencyc.org/
10. Представление онтологии OpenCyc в Web, http://www.cycfoundation.org/concepts
11. Алексеева М. В. Обзор системы Cyc. М.:РГГУ (2008), http://ezop-project.wiki.sourceforge.net/Alekseeva_Cyc
12. Сервер онтологий Ontolingua, http://www.ksl.stanford.edu/software/ontolingua/
13. Примеры использования системы Ontolingua, http://www.ksl.stanford.edu/htw/htw-demos.html
14. The World FactBook, https://www.cia.gov/library/publications/the-world-factbook/index.html
15. Википедия об онтологиях верхнего уровня, http://en.wikipedia.org/wiki/Upper_ontology_(computer_science)
16.Список ссылок на онтологии, разработанные с помощью Protégé, http://protegewiki.stanford.edu/index.php/Protege_Ontology_Library
17. Гладун А.Я., Рогушина Ю.В. Онтологии в корпоративных системах. // "Корпоративные системы" , №1, С. 41-47 (2006), http://www.management.com.ua/ims/ims115.html, http://www.management.com.ua/ims/ims116.html
18. Oracle® Database Semantic Technologies Developer's Guide 11g Release 1 (11.1) Part Number B28397-02, http://download-uk.oracle.com/docs/cd/B28359_01/appdev.111/b28397/toc.htm
19. Protégé, http://protege.stanford.edu/,
20.Конференция в Будапеште (июль 2007г.), http://protege.stanford.edu/conference/2007/schedule.html
21. Chimaera, http://www.ksl.stanford.edu/software/chimaera/
22. Luettich, K.; Masolo, C.; Borgo, S. Development of Modular Ontologies in CASL.// In Proceedings of International Workshop on Modular Ontologies (WoMO), Athens (Georgia, USA), 05 November 2006, http://www.loa-cnr.it/Publications.html
23. Википедия о Semantic Wiki, http://en.wikipedia.org/wiki/Semantic_wiki
24. SPARQL Query Language for RDF, http://www.w3.org/TR/rdf-sparql-query/
25. Semantic MediaWiki, http://semantic-mediawiki.org/wiki/Semantic_MediaWiki
26. Пример системы с запросами на естественном языке, http://www.trueknowledge.com/