ОНТОЛОГИИ - ПРОБЛЕМЫ И РЕШЕНИЯ. ТОЧКА ЗРЕНИЯ РАЗРАБОТЧИКА[1]

ONTOLOGY – PROBLEMS AND SOLUTIONS. THE DESIGNER'S STANDPOINT

Рубашкин В.Ш. *(vrub@mail.nw.ru)*
Санкт-Петербургский государственный университет

В работе обсуждаются актуальные проблемы онтологического моделирования. Представлен авторский опыт разработки универсальной онтологии словарного типа и инструментальной среды, поддерживающей работу с ней.

Представление любой разработки в области онтологического моделирования должно, по нашему мнению, состоять в ответе на вопрос как решаются в представляемом проекте проблемные аспекты, не имеющие на данном этапе однозначного и общепринятого решения. Наше видение актуальных проблем онтологического конструирования определяется следующим перечнем.

1) Определение функциональности онтологии.
2) Выбор и определение базовой модели знаний; ее реализация.
3) Инструментальная среда (онторедакторы) и используемые операционные средства.
4) Пополнение онтологии

Дальнейшее изложение содержит описание и мотивировку тех решений, которые определяют специфику онтологии ***InTez***.

1. Функциональность

Функциональность специализированной онтологии определяется, прежде всего, ее ориентированностью на поддержку приложений определенного типа.[2] Однако все более широкое признание получает мысль, что разработка специализированных онтологий, не пригодных для повторного использования (*reusable*), становится непозволительной роскошью. С этой точки зрения основная перспектива видится в движении по направлению к *универсальной* онтологии. которая, по-видимому, должна иметь двухуровневую архитектуру. Верхний уровень образует достаточно детально разработанная общая часть (*Top-Level Ontology*), являющаяся предметом общего соглашения. К ней могут присоединяться доменно и проблемно специализированные концептуальные подсистемы (с возможностью их раздельного администрирования). Движение к *универсальной* онтологии нам представляется более предпочтительным, чем разработка и поддержание изощренных средств установления соответствий между независимо разрабатываемыми онтологиями. До какого-то момента, очевидно, оба направления могут и должны сосуществовать. Но пренебрежение унификацией может, в конце концов, привести к эффекту Вавилонской башни. Полагаем, что только унификация, обеспечиваемая жестко поддерживаемой всеми участниками движения *Top-Level* онтологией, позволит избежать того печального конца, к которому пришли в свое время попытки разработки интегрированного "многоотраслевого" информационно-поискового тезауруса.

При таком подходе встает вопрос о выработке единой и универсальной точки зрения на функциональность онтологий. Функциональность онтологии (не привязываемая жестко к специфике задачи) должна обеспечить фиксацию и использование в приложениях, во-первых, информации о сочетаемостной специфике концептов; во-вторых, об их собственно логических свойствах, - т.е. о возможностях их использования в тех или иных схемах логического вывода. В одних случаях оба эти аспекта функциональности будут представлены разными функциями, в других – совмещаться в одних и тех же функциях. С содержательной точки зрения здесь имеются в виду: (а) функции, позволяющие представить классификационные различения концептов; (б) функции, представляющие парадигматические отношения между понятиями; (в) функции, представляющие описание валентностей, понимаемое как набор контекстных предсказаний, исходящих от описываемого концепта;

[1] Работа выполнена при финансовой поддержке РФФИ (проект № 06-06-80434)
[2] Для онтологии InTez исходным целевым приложением явилась задача семантического анализа делового текста [1].

(г) набор "служебных" функций, представляющих важные для анализа текста и обработки знаний внутриязыковые связи, не имеющие прямой предметной интерпретации.

В онтологии ***InTez*** функции группы (а) представлены двухуровневой классификационной схемой (*семантическая категория* и *семантический тип* – см. таблицы 1 и 2), плюс частные классификационные признаки. в зависимости от категории и типа.

Группа (б) включает 2 основные функции:

ОБЪЕМНОЕ ОТНОШЕНИЕ (D1, D2) - возвращает значение из множества {*собственно пересечение, несовместимость, включение_12*, *включение_21*} для пары концептов *(D1, D2);*

ПРЕДМЕТНО-АССОЦИАТИВНЫЕ ОТНОШЕНИЯ (D1, D2) - возвращает список значимых в предметной области отношений между заданными концептами.

АССОЦИИРОВАН (D, R) - возвращает список концептов? присоединенных к концепту *D отношением R.*

Набор функционально оформленных словарных характеристик для категории *Наименование признака* приведен в таблице 3.

Семантическая категория
Наименование признака
Имя объекта
Служебные термы
Термины свободного употребления
Процессные термины
Статические отношения

Таблица 1. Семантические категории

Наименования признаков - семантические типы	**Примеры**
Классификационный	цвет, пол, должность
Бинарный	истинность, исправность
Целочисленный	число полюсов, этажность
Количественный	вес, скорость, стоимость
Строковый	марка, фамилия
Наименование группы признаков	Параметры, химические свойства

Таблица 2. Семантические типы для категории Наименование признака

2. Базовая модель знаний

Необходимость логического обоснования – через логическую интерпретированность всех элементов описания концептов в онтологии — никем не подвергается сомнению.[3] Проблему мы видим в том, что этот принцип редко проводится последовательно, а логические языки, которые для этого используются, далеко не всегда представляются подходящими для этой цели. Базовым языком, мотивирующим и логически интерпретирующим систему словарных описаний в онтологии ***InTez,*** является логический язык ИНФОЛ [3]. В отличие от ряда других авторов мы исходим из того, что для представления профессиональных знаний не требуется конструировать принципиально новых формальных языков – достаточна логика предикатов. В числе языковых явлений, которые якобы не могут быть удовлетворительным образом формализованы в классической логике предикатов, чаще всего указывают интуитивно очевидные ограничения на смысловую сочетаемость терминов (понятие осмысленности языкового выражения – ср. такие конструкции как **идеи спят, *жидкая пирамида, *медь смертна, *металлическая медь* и т.п.); отношения и различия, существенные для установления вопрос-ответных соответствий (ср. выражения *цвет шара красный* и **цвет шара тяжелый*); представление числовых характеристик в логическом языке и др. В указанной выше работе нами показано, что решение всех этих проблем следует искать не на пути отказа от классической логики, а на пути поиска и обоснования адекватной нотации и дополнительных аксиом, позволяющих, в частности, полностью определить систему объемных отношений между концептами.

[3] Ср. [2], гл. гл. 1 и 2.

Элемент словарной статьии реализующая функция	**Дополнит. условие применимости**	**Область значений**	**Примечание**
Применимость к объектам **P_Obj (D)**	-	{ДА, НЕТ}	
Применимость к процессам **P_Proc (D)**	-	{ДА, НЕТ}	
Сочетаемость с числовым значением **P_Num (D)**	ST# = 4 Or ST# = 3	{ДА, НЕТ}	
Условие применимости признака к объектам **UP_Obj (D)**	P_Obj = ДА	DSCR#: 3.*	
Условие применимости признака к процессам **UP_Proc (D)**	P_Proc = ДА	DSCR#: 6.2	
Оцениваемость признака **P_Appraise (D)**	ST# = 4 Or ST# = 3	{0,1,2}	0 – нейтральный 1 – позитивный 2 - негативный
Небазовый для числовых характеристик **NoBasP (D)**	ST# = 4 And P_Num = ДА	{ДА, НЕТ}	Базовый - признак, восстанавливаемый по числовому параметру
Ссылка на базовый признак **BasP_P (D)**	ST# = 4 And NoBasP = ДА	DSCR#: 1.4	
Признак индивидуализирующий **P_Ind (D)**	ST# = 5 Or ST# = 1	{ДА, НЕТ}	Для ST=1 означает, что признак экземплярообразующий
Отсылка к обобщающему признаку **GeneralaizeAttrTerm (D)**	-	DSCR#: 1.9	

Таблица 3. Словарные признаки для семантической категории Наименование признака

Логический язык в данном, так и в большинстве других проектов используется как средство теоретического обзора и логической интерпретации элементов рабочего языка, на котором формулируются описания концептов. Сами же эти описания обычно представлены на другом языке – это может быть, например БНФ или табличный язык. В нашем случае используется табличный язык, поскольку он позволяет в обозримой форме представить не только состав словарных описаний, но и связь элементов словарного описания по условиям применимости (ср. таблицу 3).

Последовательность шагов, определяющих состав словарных описаний можно представить следующим образом:

- строится базовая логическая модель представления знаний;
- определяются и обосновываются ограничения, накладываемые на выразительные средства базового логического языка и его аксиоматику;
- определяются и обосновываются ограничения на реализуемые в онтологии схемы логического вывода.

Далее строится “терминологическая” модель, учитывающая особенности выразительных средств естественного языка; в ее рамках определяются основные элементы “рабочих” описаний концептов:

- категоризация терминов;
- состав словарной статьи для каждой категории;
- набор связей между понятиями (опять таки, с учетом их категориальной принадлежности.)

Системообразующим компонентом всей конструкции является *дерево признаков* [3], определяющее набор основных семантических примитивов.

Ограниченный логический вывод предусматривает использование трех типов правил вывода: по иерархии признаков в дереве признаков; по дополнительным импликативным связям; частные схемы вывода, действительные для конкретных групп концептов. Реализована схема прямого вывода, в которой базовой процедурой является построения *развертки концепта* [3] – множества всех концептов со значением более широким, чем значение данного. Эта же конструкция используется для поддержки механизма наследования свойств.

---

[4] Запись DSCR#: M.N следует понимать как обозначение множества концептов, отнесенных к семантической категории M и семантическому типу N. Т.е., в таких случаях признак реализуется указанием (именованной связи между концептами).

[5] Признак P будем называть индивидуализирующим, если он удовлетворяет следующему условию: $\forall x\ \forall y\ \forall v_1 \forall v_2 (\ (P(x, v_1)\ \&\ P(y, v_2)\ \&\ v_1 = v_2) \rightarrow x \equiv y\ )$

3. Инструментальная среда

Для пополнения и редактирования онтологии используется **специализированный онторедактор**. Специфику онторедактора ***InTez*** в достаточно обширном ряду инструментов такого рода[6] можно определить следующим образом.

Прежде всего, функционально редактор неразрывно связан со специфичной для данной онтологии моделью знаний и вытекающей из нее схемой классификации лексики.

Редактор имеет графический интерфейс, обеспечивающий визуальный режим редактирования. Графический интерфейс реализован на основе стандартного программного объекта *TreeView*, с существенным добавлением дополнительной функциональности – главным образом, в части поиска, ввода и логического контроля.

Операционной средой, обеспечивающей хранение и непосредственную манипуляцию данными в рассматриваемой онтологии, является среда реляционной СУБД.

Функциональность онторедактора можно определить следующим перечнем.

- Броузинг и поиск: поддерживается просмотр как в табличном, так и в графическом представлении, стандартные виды поиска по термину, его части или по коду концепта, выбор для просмотра терминов определенной категории и типа, выбор по параметрам администрирования (редактор и дата), выбор всех связей данного концепта, всех синонимов, всех значений слова, представленного в толковом словаре и т.п.
- Редактирование (ввод, корректировка, удаление): помимо индивидуализированных операций редактирования поддерживается ввод группы однотипных концептов списком из текстового файла.
- Логический контроль при вводе: технология ввода практически полностью исключает нарушения заданных схем описания.
- Тестирование функциональности (в текущей версии реализовано частично).
- Взаимодействие с другими онтологиями (импорт – экспорт, обычно с использованием коммуникативных форматов представления).[7]

Специфической проблемой практически всех инструментов управления словарями является **переменный состав словарной статьи**: набор релевантных описываемому объекту (слову, концепту) признаков должен динамически формироваться (уточняться) в самом процессе построения словарного описания. С этим связаны очевидные сложности, касающиеся как организации диалогового ввода, так и хранения данных. В редакторе ***InTez*** первая проблема решается посредством организации регламентированного диалога, управляемого вводимыми данными; вторая – путем использования в таблицах БД полей с переменной семантикой. В рассматриваемом редакторе специальная процедура, обеспечивающая поддержку регламентированного диалога, после добавления к словарной статье каждого очередного признака перевычисляет набор признаков, релевантных текущей ситуации ввода. При этом описание условий применимости словарных признаков вынесено в данные и реализовано в виде таблиц БД, так что механизм реализации регламентированного диалога не привязан к какой-либо конкретной системе словарных признаков. С точки зрения пользователя дело обстоит так, что ему всегда предъявляются в процессе ввода только релевантные признаки, и, соответственно, только релевантные наборы их значений. Логический контроль ввода оказывается побочным результатом построенной таким образом процедуры ввода словарных описаний.

Специфичным для данного онторедактора является также возможность тестирования функциональности - как на случайных выборках, так и для задаваемых администратором концептов. В процессе тестирования строится развертка понятия, для пары понятий вычисляются объемное отношение и ассоциированные отношения и т.д.

4. Пополнение онтологии

Здесь существенны, на наш взгляд, следующие аспекты:

- технологичность “ручного” ввода;
- поддержание корректности и целостности концептуальной модели;
- переход к автоматизированным методам пополнения онтологии.

Технологичность “ручного” ввода, на наш взгляд, определяется:

- минимизацией клавиатурного ввода;.

[6] См. [4], [5].

[7] В представляемой версии данная функция пока не реализована. В дальнейшем планируется также разработка функционального модуля, обеспечивающего генерацию схем БД и их концептуальную интерпретацию. Этот же модуль, совместно с системой анализа текста должен поддерживать семантический доступ к БД.

• простой навигацией и обозримостью словаря;
• наличием средств графического редактирования;
• наличием, тотальностью и "ненавязчивостью" средств логического контроля.

Поддержание корректности и целостности в онтологиях значительно более сложная и более критичная проблема, чем в менее сложно устроенных словарных системах. Требуется, в частности, жесткий и весьма скрупулезный логический контроль при вводе, о чем уже сказано выше.

В отношении методов пополнения онтологий большинство специалистов полагает, что методы "ручного" ("интеллектуального") ввода не могут решить проблемы создания больших концептуальных словарей, и в этой связи обсуждаются разные подходы и технологии использования разного рода информационных ресурсов: текстовых корпусов, "традиционных" информационно-поисковых тезаурусов, энциклопедических словарей.[8] Из перечисленного наиболее значимы и полезны, как нам представляется, те интеллектуальные ресурсы, которые накоплены в форме профессиональных энциклопедических словарей. Дело за тем, чтобы разработать эффективные методы автоматизированного структурирования и "перекачки" этих ресурсов в онтологические форматы.

Список литературы

1. Рубашкин В. Ш. Семантический компонент в системах понимания текста // КИИ-2006. Десятая национальная конференция по искусственному интеллекту с международным участием. Труды конференции. – М.: Физматлит, 2006.
2. Staab Steffen, Studer Rudi (eds) Handbook on Ontologies. – Berlin—Heidelberg: Springer—Verlag, 2004.
3. Рубашкин В. Ш. Представление и анализ смысла в интеллектуальных информационных системах. — М.: Наука, 1989.
4. Mizoguchi Riichiro. Ontology Engineering Environment // Staab Steffen, Studer Rudi (eds). Handbook on Ontologies. – Berlin—Heidelberg: Springer—Verlag, 2004. P. 275 – 295.
5. Овдей О. М., Проскудина Г. Ю. Инструменты инженерии онтологий // [Электронный ресурс]. (http://www.elbib.ru/index.phtml?page=elbib/rus/journal/2004/part4/op).
6. Maedche A., Staab S. Ontology Learning // Staab Steffen, Studer Rudi (eds). Handbook on Ontologies. – Berlin—Heidelberg: Springer—Verlag, 2004. P. 173 – 190.
7. Gomez, F., Hull R., and Segami C. Acquiring knowledge from encyclopedic texts. In *Proc. of the ACL's 4th Conference on Applied Natural Language Processing, ANLP94*, pages 84-90, Stuttgart, Germany, 1994.
8. Nirenburg S., Raskin V. Ontological Semantics. – Cambridge, MA: MIT Press, 2004

[8] В этой связи даже введен в употребление специальный термин **Resource Processing.** См., например, [6]. См. также [7] и[8].