

Вторичное использование лингвистических онтологий: изменения в структуре концептуализации *

© Б.В.Добров

Научно-исследовательский
вычислительный центр
МГУ им. М.В.Ломоносова
dobroff@mail.cir.ru

Н.В. Лукашевич

Научно-исследовательский
вычислительный центр
МГУ им. М.В.Ломоносова
louk@mail.cir.ru

Аннотация

В статье рассматриваются типы различий в описаниях одинаковых и близких по смыслу понятий в общезначимой онтологии и предметно-ориентированной онтологии на примере Тезауруса РуТез как общезначимой онтологии и Онтологии по естественным наукам и технологиям как предметно-ориентированной онтологии.

1 Введение

Возможность вторичного использования однажды разработанных онтологий в других областях или других приложениях является важной проблемой в онтологических исследованиях [7,9]. Для поддержки процедуры слияния онтологий и создания на этой основе новой онтологии разработано несколько программных продуктов [11,12].

Отдельное направление исследований составляет использование онтологий верхнего уровня или общезначимых онтологий (онтологий, не ориентированных на конкретную предметную область) для разработки онтологий в конкретных предметных областях. В качестве такой общей онтологии при разработке предметно-ориентированных онтологий для автоматической обработки текстов часто используется лингвистическая онтология WordNet [5,10,13].

Главной характеристикой лингвистических онтологий является то, что их единицы связаны со значениями ("are bound to the semantics") языковых выражений (слов, именных групп и т.п.) [6], что является существенным, когда речь идет о приложениях в области автоматической обработки текстов.

Близкие по смыслу понятия общей и предметно-ориентированной лингвистической онтологии могут

состоять между собой в следующих отношениях [4,7,9]:

- 1) синонимы, то есть понятия двух онтологий могут быть склеены между собой;
- 2) понятие конкретной онтологии является видовым для понятия общей онтологии;
- 3) понятия конкретной онтологии и общей онтологии являются квазисинонимами, то есть одному понятию общей онтологии соответствуют несколько понятий частной онтологии, или одному понятию частной онтологии соответствуют несколько понятий общей онтологии. В случае WordNet наличие в нем двух понятий (синсетов), относящихся к одному понятию предметной онтологии, обычно связано с более детальной трактовкой лингвистических явлений, чем это обычно принято в терминологических ресурсах.

В 2004 году мы выгрузили часть Тезауруса РуТез – лингвистической онтологии [3] в предварительную версию новой онтологии – Онтологии по естественным наукам и технологиям [2]. Таким образом, фрагменты общезначимой онтологии были перемещены в другой контекст – область естественных наук. При этом приложение онтологий является одинаковым – информационно-поисковые задачи такие как индексация и поиск документов, автоматическая рубрикация, поиск ответов на вопросы, поиск похожего документа и т.п.

Решение о разделении онтологий было связано с несколькими серьезными факторами.

Во-первых, обе онтологии достаточно объемны, включают десятки тысяч понятий и отношений, при этом большая часть понятий общей онтологии обычно не используется в текстах естественных наук, и наоборот, научные понятия по большей мере не нужны для тематического анализа таких общезначимых документов, как газетные статьи,

информационные сообщения, законодательные акты.

Во-вторых, разделение онтологий снижает многозначность описанных слов и выражений.

В-третьих, предполагалось, что существует несоответствие, так называемой, «бытовой» картины мира и научной картины мира. То есть отношения, описанные и правильные в рамках одной онтологии, должны быть изменены в рамках другой онтологии.

В течение полутора лет эксперты по знаниям работали над Онтологией по естественным наукам и технологиям, дополняли и изменяли полученные понятийные описания. В настоящее время Онтология по естественным наукам и технологиям включает 25 тысяч понятий и 60 тысяч терминов таких наук как математика, физика, химия, геология и биология. И теперь у нас есть возможность изучить, что изменилось в структуре и отношениях понятий при перемещении их в другие, более специфические предметные области.

Изучение произошедших изменений с описаниями понятий имеет как теоретическое, так и прикладное значение.

Прикладные проблемы при разделении онтологий, связаны с тем, что существуют типы текстов, для анализа которых могут понадобиться обе онтологии, работающие одновременно, и поэтому нужно иметь четкое представление об отражении сходных явлений в разных контекстах.

К числу текстов, требующих, как представляется, использования обеих онтологий относятся:

- анализ соответствий между требованиями технического регулирования и описанием производственных процессов;
- документы вида «заявки/отчеты» о научном исследовании,
- инвестиционные заявки, связанные с промышленным внедрением научных исследований.

Статья посвящена исследованию основных типов изменений описаний понятий, при перемещении из общезначимой онтологии в онтологию по естественным наукам.

Во втором разделе мы опишем основные принципы разработки онтологий и их структуру.

Во третьем разделе описывается процедура выгрузки фрагментов Тезауруса РуТез в новую онтологию.

В четвертом разделе перечисляются основные типы возможных изменений описаний понятий онтологии-прототипа и описываются наиболее интересные явления, которые удалось выявить.

В пятом разделе описывается пример изменения состава отношений в Тезаурусе РуТез и Онтологии по естественным наукам и технологиям.

2 Структура онтологий

Оба ресурса: Тезаурус РуТез и Онтология по естественным наукам и технологиям имеют одинаковую структуру. Они являются онтологиями, поскольку описывают понятия внешнего мира и отношения между ними, которые устанавливаются в соответствии с требованием правомочности расширения запроса по иерархии связей при информационном поиске. Оба ресурса принадлежат к особому классу онтологий, так называемым лингвистическим онтологиям [6,10], поскольку введение понятий в значительной мере мотивируется значениями языковых единиц, относящихся к предметной области ресурса. Далее в этом разделе мы опишем структуру этих ресурсов, ссылаясь на них обобщенным названием РуТез*Онтологии.

В то же время они являются тезаурусами, поскольку каждое понятие связано с набором языковых выражений (слов, терминов, словосочетаний), которыми это понятие может быть выражено в тексте, - такой набор текстовых входов понятий необходим для использования онтологий для автоматической обработки текстов.

Опишем сказанное более подробно.

РуТез*Онтология – это иерархическая сеть понятий. Каждое понятие имеет имя.

Для сопоставления с текстом каждое понятие снабжается набором текстовых выражений («текстовых входов», «терминов»), значения которых соответствует данному понятию. В качестве таких текстовых входов могут выступать однословные существительные, прилагательные, глаголы, именные и глагольные группы. Количество таких текстовых входов понятий может быть достаточно велико, например, превышать 20 единиц. При вводе нового понятия делаются специальные усилия, чтобы максимально подробно перечислить его возможные текстовые входы.

Каждое понятие связывается отношениями (Рис.1) с другими понятиями РуТез*Онтологии. Набор отношений РуТез*Онтологии специально подобран для эффективной работы в информационно-поисковых приложениях. В РуТез*Онтологии имеется четыре основных типа отношения.

Первый тип отношений – родовидовое отношение НИЖЕ-ВЫШЕ, обладает свойством транзитивности и наследования.

Второй тип отношений – отношение ЧАСТЬ-ЦЕЛОЕ. Используется не только для описания физических частей, но и для других внутренних сущностей понятия, таких как свойства или роли для ситуаций. Важным условием при установлении этого отношения является то, что понятия-части должны быть жестко связаны со своим целым, то есть каждый пример понятия-части должен в течение всего времени своего существования являться частью для понятия-целого, и не относиться к чему-либо другому.

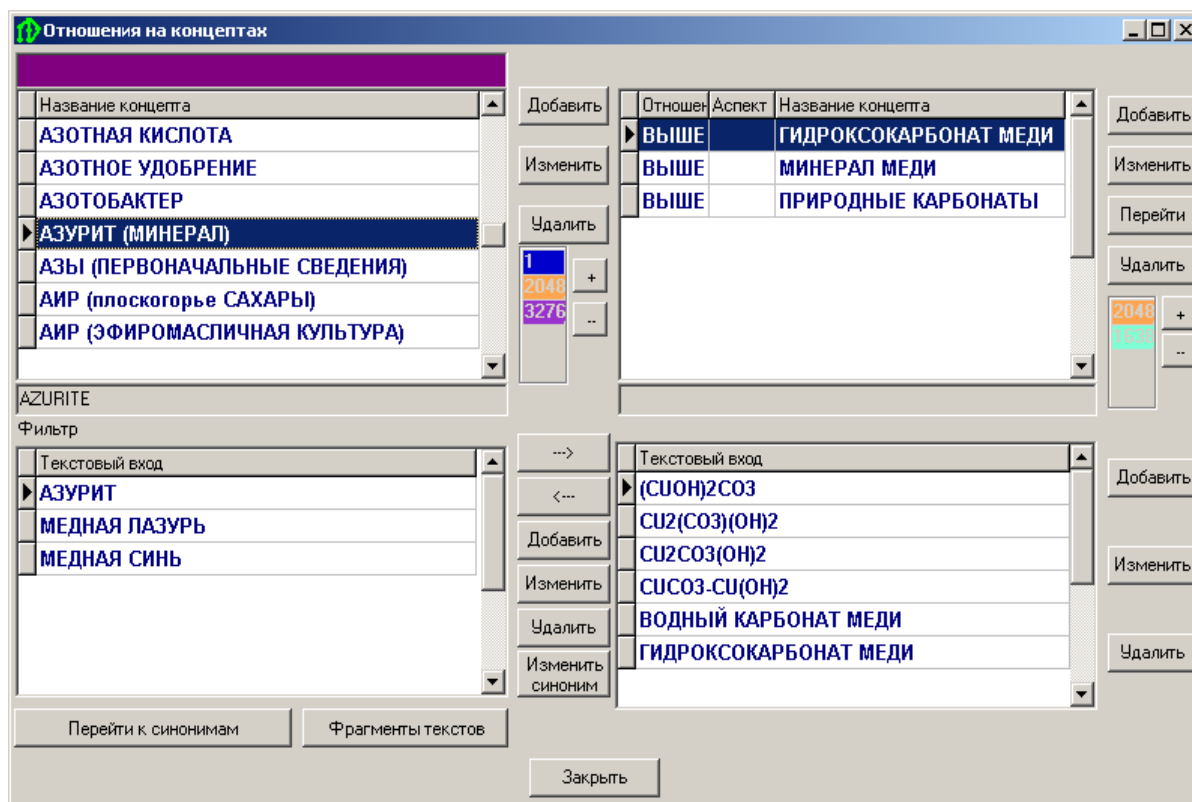


Рис.1. Основная экранная форма редактирования отношений и текстовых входов понятий

Так, например, в RuTез*Онтологиях считается неправильным описывать понятие *ДВИГАТЕЛЬ* частью понятия *АВТОМОБИЛЬ*, поскольку двигатели являются частями различных технических устройств, а не только автомобилей. Мы вводим понятие *АВТОМОБИЛЬНЫЙ ДВИГАТЕЛЬ*, как видовое понятие для понятия *ДВИГАТЕЛЬ*, и затем устанавливаем отношение ЧАСТЬ между понятием *АВТОМОБИЛЬ* и понятием *АВТОМОБИЛЬНЫЙ ДВИГАТЕЛЬ*.

В этих условиях удается выполнить свойство транзитивности введенного таким образом отношения ЧАСТЬ-ЦЕЛОЕ, что очень важно для автоматического вывода в процессе автоматической обработки текстов.

Еще один тип отношения, называемого несимметричной ассоциацией АСЦ2-АСЦ1, связывает два понятия, которые не могут быть связаны выше рассмотренными отношениями, но одно из понятий не существовало бы без существования другого. Например, понятие *ЛЕС* требует существования понятия *ДЕРЕВО* (при этом можно было бы ввести как ЧАСТЬ для понятия *ЛЕС* понятие *ДЕРЕВО В ЛЕСУ*), а понятие *АНТИСТАТИК* требует существования понятия *СТАТИЧЕСКОЕ ЭЛЕКТРИЧЕСТВО*.

Последний тип отношений – симметричная ассоциация связывает, например, понятия очень близкие по смыслу, но которые эксперты не решились склеить в одно понятие.

Отношения НИЖЕ-ВЫШЕ, ЧАСТЬ-ЦЕЛОЕ и несимметричная ассоциация АСЦ2-АСЦ1 являются иерархическими отношениями. Таким образом, на основе свойств иерархичности, транзитивности и наследования для каждого понятия может быть определена совокупность понятий, которые являются для него нижестоящими понятиями по иерархии.

3 Процедура выгрузки понятийных структур из Тезауруса RuTез в Онтологию по естественным наукам и технологиям

Для каждой области науки [2] из рассматриваемого списка (математика, физика, химия, биология, геология) были сформированы коллекции документов (от 3000 до 8000 документов, от 50 до 90 Мб). Источником коллекций являлись документы, доступные в Интернет, следующих основных типов: материалы школьных уроков, рефераты, университетские лекции, материалы специализированных сайтов.

На собранной коллекции были проведены две различные процедуры извлечения терминоподобных словосочетаний и проведено сопоставление с Тезаурусом RuTез.

Полученные списки терминов-кандидатов были собраны в единый список и упорядочены по частотности, после чего были выданы на проверку экспертам. Задачей экспертов на данном этапе

Таблица концептов

Код концепта: 135790 Ввел: MARIA 17.03.200 Изменил: Olga 17.11.2004

Название концепта: АЗУРИТ (МИНЕРАЛ)

Английский эквивалент: AZURITE

Абстрактность: Спорность:

Комментарий:

Предметная область:

- Общая терминология
- Географическая терминология
- Персоналии
- Общий лексикон
- Выборная терминология
- Изображения
- Информатика
- Государственная Дума
- TREC
- Тест
- candidate
- Наука
- Математика
- Физика
- Химия
- Геология
- Биология
- Медицина

Отношения:

Отношение	Аспект	Связанный концепт
ВЫШЕ		ГИДРОКСОКАРБОНАТ МЕДИ
ВЫШЕ		МИНЕРАЛ МЕДИ
ВЫШЕ		ПРИРОДНЫЕ КАРБОНАТЫ

Добавить
Изменить
Удалить

Готово Отмена

Рис.2. Экранная форма задания понятия и его ассоциирования с предметными областями

являлась очистка списков от ошибок и общеупотребительных сочетаний, а также приписать каждое слово или словосочетание из списка к одной или нескольким наукам.

Термины-кандидаты были загружены в оболочку ведения онтологии как имена понятий онтологии и снабжены отношениями к тем областям наук, которые были проставлены экспертами. Для тех терминов, которые были найдены в Тезаурусе РуТез, были перенесены все вышестоящие понятия, синонимы и отношения. Понятия Тезауруса РуТез были также помечены как понятия-кандидаты.

После этого данные версии были выданы экспертам по знаниям, ответственным за формирование фрагмента новой онтологии по отдельным научным дисциплинам.

Полученный список понятий-кандидатов является одним из важнейших источников информации о предметной области.

В процессе работы над онтологией:

- эксперт может одобрить созданное понятие-кандидат и снять пометку кандидата (Рис.2);
- несколько понятий-кандидатов могут возникнуть на основе синонимичных терминов, в этом случае эксперт склеивает понятия-кандидаты в одно понятие, то есть одно понятие-кандидат исчезает;
- понятия-кандидаты могут быть представлять конкретные использования понятий в текстах и, таким образом, служить источником примеров и контрпримеров для понятийных описаний.

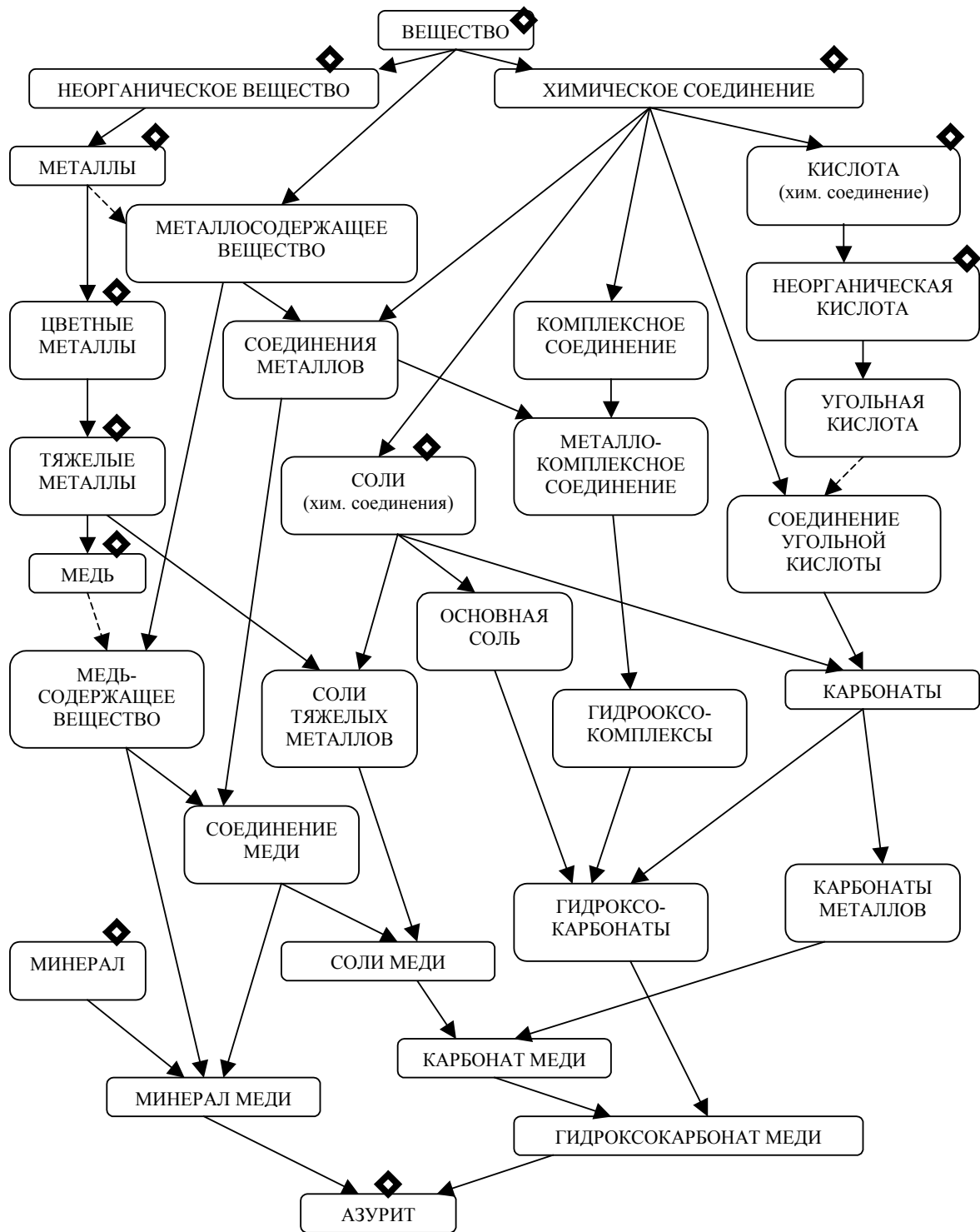


Рис.3. Фрагмент Онтологии по естественным наукам и технологиям

4 Изменения в описаниях понятий, полученных из Тезауруса РуТез

Для изучения описаний понятий, перенесенных из Тезауруса РуТез (далее онтология-прототип), мы образовали список таких понятий, которые эксперты одобрили для включения в Онтологию по

естественным наукам и технологиям, то есть сняли пометку «понятие-кандидат» (Рис.2). Таких понятий оказалось 4540.

С описаниями понятий могли произойти следующие типы изменений:

- 1) Изменение названия понятия;

- 2) Изменение набора текстовых входов понятия:
 - а. Удаление текстовых входов понятия;
 - б. Добавление текстовых входов понятия;
- 3) Изменение отношений между понятиями онтологии-прототипа:
 - а. Исчезновение отношений между понятиями онтологии-прототипа;
 - б. Появление новых отношений между понятиями онтологии-прототипа;
- 4) Введение отношений понятий онтологии-прототипа с новыми понятиями:
 - а. Введение отношений вверх по иерархии;
 - б. Введение отношений вниз по иерархии.

В следующих подразделах рассмотрим наиболее интересные явления, которые удалось выявить.

4.1. Удаление текстовых входов понятия

Изменения набора текстовых входов понятия связано в основном с двумя причинами.

Во-первых, от понятия отсоединяются текстовые входы, носящие метафорический, образный характер, свойственные газетным текстам и неупотребляемые в научной речи, например, *ВЕРБЛЮД – корабль пустыни*.

Во-вторых, (и таких удаленных текстовых входов большинство) часть текстовых входов исходного одного понятия перешло как текстовые входы к новообразованному понятию, то есть практически понятие расщепилось на два (или более) понятий. Например, были разделены в отдельные понятия бывшие синонимы (текстовые входы одного и того же понятия): *ХИМИЧЕСКАЯ РЕАКЦИЯ* и *ХИМИЧЕСКИЙ ПРОЦЕСС*, *СУДОРОГА* и *СПАЗМ*, *СОЛИ ФОСФОРНЫХ КИСЛОТ* и *ФОСФАТЫ* и т.п.

4.2. Замена отношений между понятиями онтологии-прототипа на более длинные цепочки отношений

Авторы [5,10], работавшие с двумя онтологиями, одна из которых более общая, а вторая относится к конкретной предметной области, предполагали, что набор вышестоящих отношений более общей онтологии не подвергается изменениям.

Однако наше сопоставление показало значимое число удаленных родовидовых отношений между понятиями онтологии-прототипа. Более тщательный анализ показал, что достаточно часто удаленное отношение заменяется на более длинную цепочку отношений, состоящую из двух или трех отношений, то есть между понятиями, перешедшими из более общей онтологии, вклиниваются одно-два понятия из предметной онтологии.

Например, в Тезаурусе РуТез для понятия *АДСОРБЕНТ* было установлено родовидовое отношение к понятию *ВЕЩЕСТВО*, а в новой онтологии создана цепочка понятий *АДСОРБЕНТ – СОРБЕНТ – ВЕЩЕСТВО*.

Отношение между понятиями *БОКСИТ – ГОРНАЯ ПОРОДА* заменилось на цепочку *БОКСИТ – БИОГЕННАЯ ГОРНАЯ ПОРОДА – ОСАДОЧНАЯ ГОРНАЯ ПОРОДА – ГОРНАЯ ПОРОДА*.

Отношение между понятиями *БУЙВОЛ – ЖВАЧНОЕ ЖИВОТНОЕ* заменилось на цепочку *БУЙВОЛ – ПОЛОРОГИЕ – ЖВАЧНОЕ ЖИВОТНОЕ* и т.д.

Количество таких замен одного отношения на цепочку отношений оценивается на текущий момент как более 1000 единиц, что для множества рассматриваемых понятий онтологии-прототипа (4540) представляется значительной величиной.

Важно отметить, что часть из нововведенных отношений может быть перенесена и в исходную онтологию, послужить для уточнения исходных описаний. Вместе с тем значительная часть нововведений не подлежит переносу в онтологию-прототип (см. примеры выше), поскольку введенные понятия соответствуют исключительно научной терминологии и практически не используются в общезначимых текстах.

4.3. Несоответствие наивной, бытовой картины мира и научной картины мира

Тезаурус РуТез предназначен для обработки общезначимых документов: информационных сообщений, нормативных документов, газетных статей. Поэтому он должен отражать знания о мире, которыми обладают авторы и читатели такого вида документов. Картина мира, представленная в тезаурусе, может отличаться от картины мира, излагаемой в рамках естественных наук.

Хрестоматийным примером отличия бытовой картины мира и научной картины мира является знание о том, что кит является млекопитающим, а не рыбой [1]. Однако этому вопросу уделяется достаточное внимание в курсе зоологии средней школы. В частности, не удалось найти ни одного такого текста в текстовой коллекции Университетской информационной системы РОССИЯ (www.cir.ru, более миллиона документов), в котором бы автор считал, что кит – это рыба. Тезаурус РуТез также описывает китов как морских млекопитающих.

Однако удалось выявить ряд несоответствий наивной картины мира, зафиксированной в Тезаурусе РуТез, и научной картиной мира.

Здесь можно выделить два типа различий. Первый тип различий состоит в том, что, то, что в наивной картине мира кажется связанным простым отношением (например, родовидовым), в научной картине мира напрямую не связано. Второй тип различий – то, что представляется несвязанным в наивной картине мира, непосредственно связано между собой в научной картине мира.

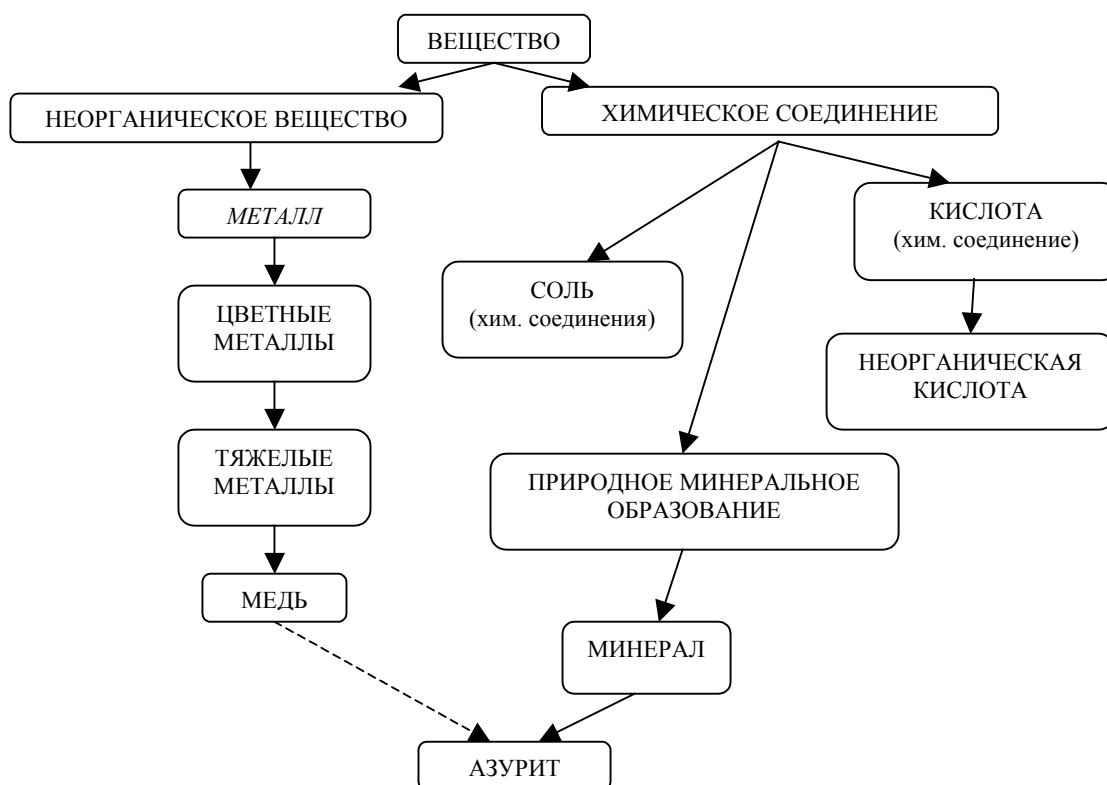


Рис.4. Аналогичный фрагмент Тезауруса РуТез

Большинство примеров несоответствий находится в сфере биологии. Так, птица эму, которую часто называют *страус эму*, по биологической классификации не является *страусом*.

С другой стороны, по биологической классификации *бледная поганка* относится к *мухоморам*, а *горчица* и *брюква* к *роду капуста*.

Наиболее запутанной ситуацией является ситуация с употреблением слова *орех*. Биологическая наука рассматривает орех как особый вид плода, к которым, например, не относятся грецкие орехи. Одновременно существует «хозяйственный» (по выражению Большой Советской энциклопедии) взгляд на орехи – плоды деревьев и кустарников, «состоящие из сухой деревянистой оболочки и заключённого в ней съедобного и питательного ядра».

Кроме того, существует еще более отличающееся от научного употребление слова *орех*, которое включает в *орехи* – *арахис*, *земляной орех*. Это растение по биологической классификации относится к бобовым культурам и не является деревом или кустарником.

Работа с такими несоответствиями связана с двумя видами деятельности: изменение отношений между понятиями на более научно-мотивированные (в том числе и в онтологии-прототипе) и/или ввод разных понятий для разного употребления того или иного слова и описание такого слова как многозначного. Так, видимо, целесообразно иметь

два понятия для плода орех – орех как плод ореховых культур (биологическая картина мира) и орех как плод орехоплодных культур («хозяйственная» картина мира).

4.4. Смена антропоцентрической картины мира на естественнонаучную картину мира

Наивная картина мира отличается еще и тем, что она ставит в свой центр человека и общество, то есть является антропоцентрической. При переходе к естественнонаучной картине мира эта антропоцентричность пропадает, что находит отражение в отношениях онтологии.

Мы заметили это явление в двух проявлениях.

Есть знание, которое известно и в наивной картине мира, но из-за того, что в повседневной жизни некоторая сущность чаще всего встречается в той или иной форме, то эта форма и считается основной для сущности.

Это явление хорошо видно на примере веществ и их агрегатных состояний и проявляется уже в различиях в толкованиях, которые даются в толковых словарях и энциклопедических словарях.

Так, в толковом словаре [4] первое значение слова *вода* таково: *1. Бесцветная прозрачная жидкость, представляющая собою химическое соединение водорода и кислорода и содержащаяся в атмосфере, почве, живых организмах и т.п.*

В Большой Советской энциклопедии термин *вода* имеет такое определение: *окись водорода, H₂O, простейшее устойчивое в обычных условиях*

химическое соединение водорода с кислородом (11,19% водорода и 88,81% кислорода по массе), молекулярная масса 18,0160; бесцветная жидкость без запаха и вкуса (в толстых слоях имеет голубоватый цвет),

Как следствие, в тезаурусе РуТез установлено отношение *ВОДА – ЖИДКОСТЬ*, в Онтологии по естественным наукам и технологиям *ВОДА – это СОЕДИНЕНИЕ КИСЛОРОДА С ВОДОРОДОМ, ОКСИД НЕМЕТАЛЛА*. Вводится дополнительное понятие *ЖИДКАЯ ВОДА* (вода в жидкой фазе, вода в жидком состоянии), которая и является видом понятия *ЖИДКОСТЬ*.

При этом образованным современникам отлично известно, что как химическое соединение, вода бывает в разных агрегатных состояниях, но установить отношение именно между понятиями *ВОДА* и *ЖИДКОСТЬ* в общезначимом ресурсе удобно, так как жидкое агрегатное состояние воды является наиболее обсуждаемым, другие агрегатные состояния *ВОДЯНОЙ ПАР* и *ЛЕД* воспринимаются как производные от основного.

Еще один элемент антропоцентрической картины мира в тезаурусе РуТез – это наличие таких оценочных понятий как *СТИХИЙНОЕ БЕДСТВО*, которое оценивает воздействие тех или иных явлений на человеческое существование и включает такие понятия как *ЗЕМЛЕТРЯСЕНИЕ*, *СМЕРЧ*, *НАВОДНЕНИЕ* и др. Как представляется естественнонаучная онтология должна избегать таких оценочных понятий как *СТИХИЙНОЕ БЕДСТВО* и должна использовать нейтральные классификации: *СЕЙСМИЧЕСКОЕ ЯВЛЕНИЕ*, *МЕТЕОРОЛОГИЧЕСКОЕ ЯВЛЕНИЕ* и т.п.

5 Пример

В качестве примера сравним описание понятия *АЗУРИТ* в составе Тезауруса РуТез и Онтологии по естественным наукам и технологиям.

Азурит – достаточно известный минерал, используется для получения меди и медного купороса, а также для изготовления синей краски, поэтому соответствующее понятие давно описано в составе Тезауруса РуТез.

Описание понятия *АЗУРИТ* в тезаурусе РуТез таково:

АЗУРИТ

син АЗУРИТ

син МЕДНАЯ ЛАЗУРЬ

ВЫШЕ *МИНЕРАЛ*

син МИНЕРАЛ

син МИНЕРАЛЬНОЕ ВЕЩЕСТВО

син МИНЕРАЛЬНЫЙ

АСЦ1 *МЕДЬ*

син МЕДНЫЙ

син МЕДНЫЙ КОНЦЕНТРАТ

син МЕДЬ

син МЕДЬСОДЕРЖАЩИЙ

а в Онтологии по естественным наукам и технологиям:

АЗУРИТ (МИНЕРАЛ)

син АЗУРИТ

син МЕДНАЯ ЛАЗУРЬ

син МЕДНАЯ СИНЬ

ВЫШЕ *ГИДРОКСОКАРБОНАТ МЕДИ*

син (CUOH)2CO3

син CU2(CO3)(OH)2

син CU2CO3(OH)2

син CUCO3-CU(OH)2

син ВОДНЫЙ КАРБОНАТ МЕДИ

син ГИДРОКСОКАРБОНАТ МЕДИ

ВЫШЕ *МИНЕРАЛ МЕДИ*

син МЕДНЫЙ МИНЕРАЛ

син МИНЕРАЛ МЕДИ

син ПРИРОДНАЯ МЕДЬ

ВЫШЕ *ПРИРОДНЫЕ КАРБОНАТЫ*

син КАРБОНАТНЫЙ МИНЕРАЛ

син МИНЕРАЛ КЛАССА

син КАРБОНАТОВ

син ПРИРОДНЫЕ КАРБОНАТЫ

Рисунок 1 показывает рабочий экран ведения РуТез*Онтологий. В левой верхней части экрана помещен список понятий, курсором выделено рассматриваемое понятие - *АЗУРИТ*. В левой нижней части экрана показаны текстовые входы для понятия. Правая верхняя часть экрана представляет список понятий, связанных отношениями с рассматриваемым. Курсор установлен на отношении с понятием *ГИДРОКСОКАРБОНАТ МЕДИ*. Правая нижняя часть экрана показывает текстовые входы понятия, выделенного курсором в правой части экрана.

Рисунок 2 представляет экран ввода характеристик понятия *АЗУРИТ*. На экране показан список возможных предметных областей, подмножество которых выбирается для характеристики каждого понятия.

Рисунок 3 показывает верхние уровни иерархии понятия *АЗУРИТ* в Онтологии по естественным наукам и технологиям (за недостатком места приведены в основном только таксономические отношения, причем не все описанные отношения отражены). Ромбиками помечены понятия, которые были экспортированы из тезауруса РуТез.

На рисунке 4 для сравнения показаны верхние уровни иерархии понятия *АЗУРИТ* в тезаурусе РуТез.

Мы можем видеть, что прямые отношения понятия *АЗУРИТ* в тезаурусе РуТез заменились на многоступенчатые структуры, описывающие химический состав минерала.

Заключение

В статье мы рассмотрели типы несоответствий между описаниями одинаковых и близких по смыслу понятий в общезначимой онтологии и предметно-ориентированной онтологии на примере Тезауруса РуТез как общезначимой онтологии и Онтологии по естественным наукам как предметно-ориентированной онтологии.

Одним из наиболее важных выявленных фактов является новый взгляд на структуру «стыка» между такими онтологиями. Стык не представляет собой сплошную полосу понятий, принадлежащих обеим онтологиям. Стык онтологий выглядит как совокупность полос, в которых между уровнями, принадлежащими обеим онтологиям, находятся понятия, принадлежащие только одной из онтологий.

Различия в антропоцентрической «наивной» картине мира и естественнонаучной картине мира, проявляются в несоответствиях между описаниями понятий в соответствующих онтологиях.

Литература

- [1] Апресян Ю.Д., Лексическая семантика. Синонимические средства языка. – М.: Восточная литература. – 1995.
- [2] Добров Б.В., Лукашевич Н.В., Синицын М.Н., Шапкин В.Н., Разработка лингвистической онтологии для автоматического индексирования текстов по естественным наукам // Электронные библиотеки: перспективные методы и технологии, электронные коллекции. Труды Седьмой Всероссийской научной конференции (RCDL'2005) г. Ярославль 4-6 октября 2005г. – Ярославль: ЯрГУ им. П.Г. Демидова, 2005. – С.70-79.
- [3] Добров Б.В., Лукашевич Н.В., Тезаурус и автоматическое концептуальное индексирование в университетской информационной системе РОССИЯ // Третья Всероссийская конференция по Электронным Библиотекам "Электронные библиотеки: перспективные методы и технологии, электронные коллекции" - Петрозаводск, 2001 - С.78-82.
- [4] Ефремова Т.Ф. Современный толковый словарь русского языка. В 3 т. – Изд-во АСТ, 2006.
- [5] Buitellar, P., Sacalenu, B., Extending Synsets with Medical Terms. // Proceedings of the NAACL workshop on WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA, 2001.
- [6] Gomez-Perez A., Fernandez-Lopez M., Corcho O., OntoWeb. Technical Roadmap. D.1.1.2. - IST project IST-2000-29243.
- [7] Guarino N., Understanding, building and using ontologies. Int. Journal of Human-Computer Studies (IJHCS) 46 (1997) 293—310.
- [8] Hovy E.H., Combining and standardizing large-scale, practical ontologies for machine translation and other uses. // Proceedings of the 1st International Conference on Language Resources and Evaluation (LREC). Granada, Spain, 1998.
- [9] Kalinichenko L., Skvortsov N., Ontology reconciliation in terms of type refinement. In Proceedings of the 6th Russian Conference on Digital Libraries RCDL2004, Pushchino, Russia, September 2004
- [10] Magnini B., Speranza M. Merging Global and Specialized Linguistic Ontologies. – In Proceedings of OntoLex 2002.
- [11] McGuinness D.L., Fikes R., Rice J., Wilder S., An environment for merging and testing large ontologies. In: Proc. of the Seventh International Conference (KR'2000). Morgan Kaufmann Publishers, San Francisco (2000)
- [12] Noy N.F., Musen M.A., PROMPT: Algorithm and tool for automated Ontology merging and alignment. In: Proc. of the Seventeenth National Conference on Artificial Intelligence (AAAI-2000), Austin, TX (2000)
- [13] Vossen P., Extending, Trimming and Fusing WordNet for Technical Documents. // Proceedings of WordNet and Other Lexical Resources: Applications, Extensions and Customizations, Pittsburg, USA, 2001.

Reuse of Linguistic Ontologies: Changes in Conceptualization Structure

Dobrov B., Loukachevitch N.

In the paper we describe types of distinctions between descriptions of the same or similar concepts in general linguistic ontology RuThes and Ontology on Natural Sciences and Technologies.

One of the important found phenomena is a new view on the structure of juncture between general and specific ontologies. The juncture between such ontologies is not a continuous zone of concepts belonging to both ontologies. The juncture of ontologies looks as a set of stripes, where levels of common concepts for both ontologies alternate with levels belonging only one ontology.

Distinctions in the anthropocentric naive picture of the world and the scientific picture of the world are reflected in differences between descriptions of concepts in corresponding ontologies.

* Данная работа поддержана грантом РФФИ № 05-07-90391.