БАЗИ ДАНИХ, БАЗИ ЗНАНЬ ТА ІНЖЕНЕРІЯ ПРОГРАМНОГО ЗАБЕЗПЕЧЕННЯ

УДК 004:504(045)

Ф.И. Андон, В.А. Резниченко
Інститут програмних систем Національної академії наук України

УПРАВЛЯЕМЫЕ СЛОВАРИ, ТАКСОНОМИИ, ТЕЗАУРУСЫ И ОНТОЛОГИИ В СЕМАНТИЧЕСКОМ ВЕБЕ

В статье рассматривается взаимосвязь понятий управляемого словаря, таксономии, тезауруса и онтологии с точки зрения семантического веба.

У статті розглядається взаємозв'язок понять керованого словника, таксономії, тезауруса і онтології з точки зору семантичного вебу.

The relationship of controlled vocabulary notionts, taxonomies, thesauris and ontologies in terms of the Semantic Web is considered in the article.

Ключевые слова: *управляемый словарь, таксономия, тезарус, онтология, семантический веб*

1. Веб и семантический веб

Традиционный веб – это совокупность взаимосвязанных ресурсов. В вебе абсолютно нет никаких различий между имеющимися ресурсами и ссылками, которые эти ресурсы связывают. Одна из основных задач семантического веба – предоставить как можно большую семантику ресурсам и связям между ними. Причем эта семантика может либо определяться простым введением осмысливаемых и однозначно понимаемых человеком терминов, либо дополнительной формализацией этих терминов (понятий), что позволяет компьютеру «понимать» эти понятия и манипулировать ими строго в соответствии с предоставленной формализацией. Приведенные рассуждения иллюстрируются приводимым ниже рисунком.

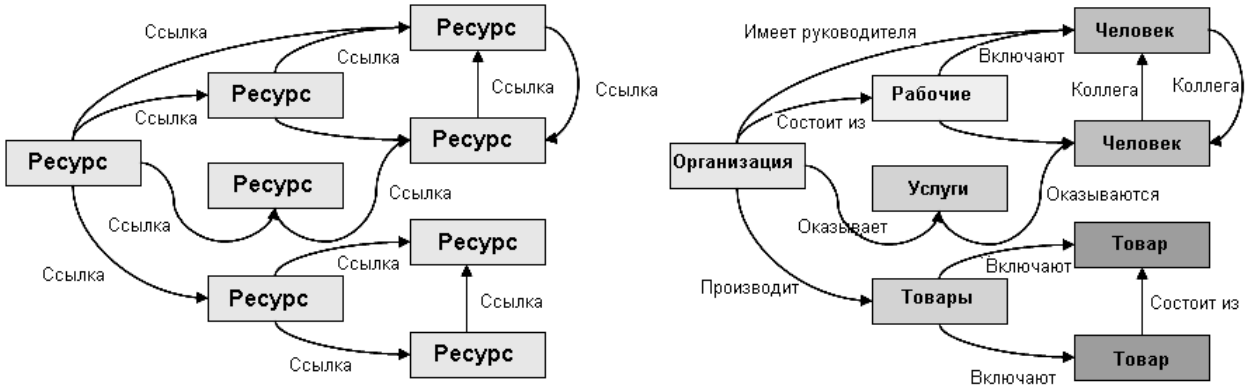

Рис. 1. Веб и семантический веб

Общепринятым в семантическом вебе подходом по приданию смысла понятиям и связям между ними является использование онтологий.

В этом разделе делается попытка определить это понятие. Существует два подхода по определению понятий. Первый заключается в том, что определяется контекст,

в котором это понятие используется, указываются различные взаимосвязи с этим контекстом и указываются отличия определяемого понятия от такого контекста. Второй подход заключается в формулировке определения понятия и, если необходимо, последующего раскрытия смысла используемых в этом определении понятий. В данной статье рассматривается первый подход, то есть смысл онтологии раскрывается через такие понятия как управляемый словарь, таксономия, и тезаурус.

2. От управляемого словаря к онтологии

Один из способов изучения того или иного понятия является нахождение его места в некоторой классификационной системе. Именно так мы сейчас подойдем к понятию онтологии, указав ее взаимосвязь с такими понятиями, как управляемый словарь, таксономия и тезаурус. Эта связь графически представлена на рисунке ниже (см., например,.[1])

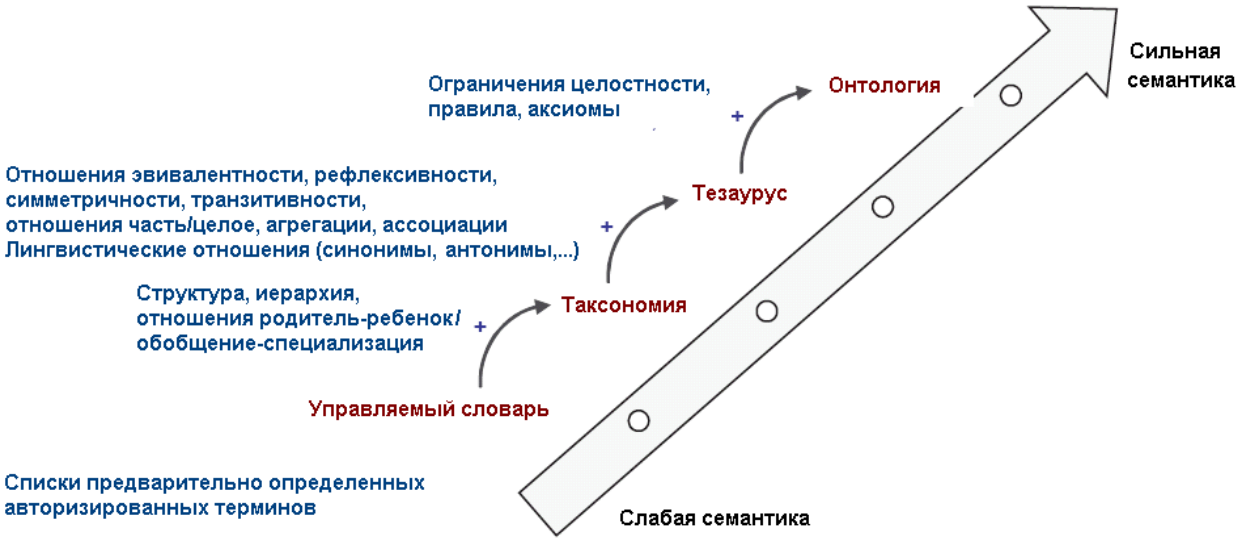

Рис. 2. От управляемых словарей к онтологиям

Кратко прокомментируем этот рисунок.

Одним из начальных этапов познавательной деятельности является формулировка понятий. В нашем случае под понятием мы подразумеваем следующее: понятие – это целостная совокупность суждений, то есть мыслей, в которых что-либо утверждается об отличительных признаках исследуемого объекта, ядром которого являются суждения о наиболее общих и в то же время существенных признаках этого объекта [2, с 456]. Понятие включает **объем (экстенсионал)** – класс обобщенных в понятии индивидов (объектов) и **содержание (интенсионал)** – совокупность (обычно существенных) признаков, по которым произведено обобщение и выделение объектов в данном понятии.

Выделение понятий – это основа классификации. Управляемые словари – это, по сути, множества индивидов объемов понятий, которые обладают признаком устойчивости в той или иной предметной области.

Следующий шаг – систематизация понятий. Это, как правило, установление между понятиями отношений одного определенного типа – обобщения/специализации. Это дает нам таксономию.

Следующий шаг – установление между понятиями произвольных связей (отношений), некоторые из которых (но не обязательно все) обладают строгим смыслом. Это тезаурусы

Наконец, строгая формальная спецификация смысла понятий и связей между ними преобразует тезаурус в онтологию.

Далее детально рассматриваются все эти понятия.

2.1. Управляемые словари

Управляемые словари (controlled vocabulary) – это способ организации знаний с целью облегчения их представления и последующего поиска.

Управляемый словарь – это список предварительно определенных, явно заданных тщательно отобранных терминов (слов, фраз или нотаций), а не произвольных слов естественного языка, для описания понятий, связанных с информационными ресурсами. Все термины в управляемом словаре должны иметь однозначное и неизбыточное толкование (определение). Управляемые словари – это основа классификации.

Управляемые словари возникают тогда, когда относительно того или иного понятия достаточно знать только его экстенсионал.

Например, в подавляющем числе случаев использования информационных ресурсов относительно такого понятия, как «день недели», вполне достаточно иметь перечень (экстенсионал) этих дней без какого-либо раскрытия смысла этого понятия (интенсионала). То же можно сказать, например, относительно понятия «язык». Если наша предметная область относится к электронному каталогу, в котором для любой публикации следует указывать его язык, то для нас вполне достаточно иметь список существующих языков.

Списки допустимых терминов/значений, которые могут использоваться для тех или иных целей (определения, ввода, индексации, поиска) являются управляемыми словарями. Ниже на рисунках приводятся примеры трех управляемых словарей, используемых при вводе данных, а также при формулировке поисковых запросов.

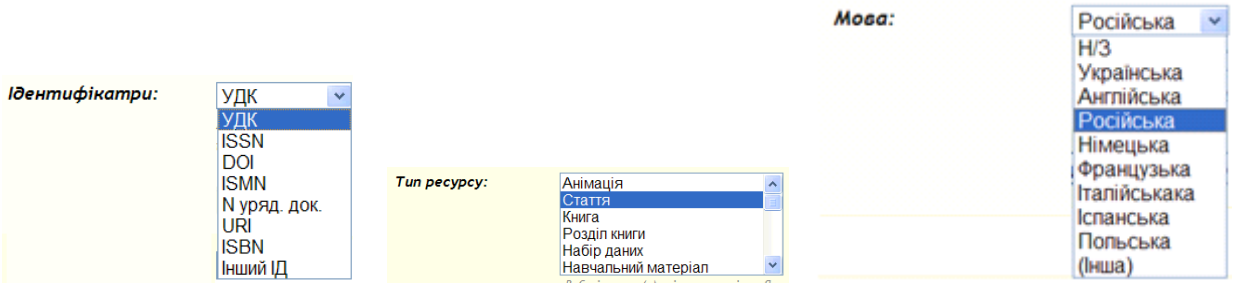

Рис. 3. Примеры управляемых словарей

Управляемые словари используются в схемах предметной индексации, в предметных рубриках, в таксономиях и тезаурусах и в других системах организации знаний.

В библиотечных и информационных системах управляемый словарь представляет собой список тщательно отобранных слов и фраз, которые используются для разметки информационных единиц (документов или слов) с тем, чтобы их можно было проще найти. Управляемые словари используются в языках индексирования, например, в языках управляемого индексирования, в которых могут использоваться только специально отобранные термины.

Управляемые словари фиксируют возможные варианты выбора значений. Например, при каталогизации пользователям предоставляется список допустимых значений тех или иных понятий, которые можно использовать. В этом случае предотвращается возможность использования тех или иных терминов по желаю пользователя, которые могут оказаться неоднозначными, бессмысленными или указанными с грамматическими ошибками.

Управляемые словари решают проблемы синонимии, антонимии, полисемии путем установления соответствия между понятиями и авторизованными терминами (см. далее о таксономии). Другими словами, управляемые словари уменьшают неопределенность и неоднозначность, присущую естественному языку, когда одному и тому же понятию могут придаваться различные имена и наоборот.

2.2. Таксономии

Таксоно́мия (от др.-греч. τάξις — строй, порядок и νόμος — закон) — учение о принципах и практике ***классификации и систематизации*** [3]. Термин «таксономия» впервые был предложен (в 1813 г.) для классификации растений и животных согласно некоторым естественным взаимоотношениям между ними, и изначально применялся только в биологии. Позже этот термин стал использоваться для обозначении общей теории классификации и систематизации сложных систем как в биологии, так и в других областях знаний, в лингвистике, географии, геологии.

Таксономия – это предметная классификация, которая группирует термины в виде управляемых словарей и упорядочивает их (словари) в виде иерархических структур. Ниже приводятся примеры двух таксономий – биологическая таксономия, ведущая к человеку, и таксономия «управляющего объекта»

Фауна: мир животных
Тип: хордовые
Подтип: позвоночные
Класс: млекопитающие
Подкласс: звери
Отряд: приматы
Подотряд: антропоиды
Суперсемейство: гоминоиды
Семейство: гоминиды
Род: Homo
Вид: Sapiens
Класс: диапсиды (рептилии, динозавры, птицы)

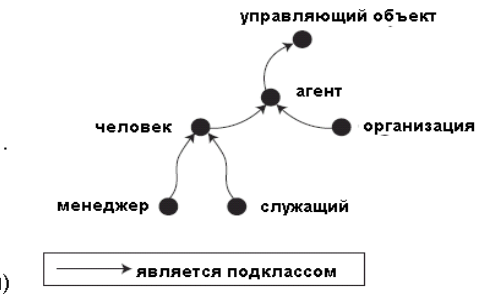

Рис. 4. Примеры таксономий

Таксономическая структура. Математически таксономией является древообразная структура классификаций определенного набора объектов. Вверху этой структуры — объединяющая единая классификация — корневой таксон — которая относится ко всем объектам данной таксономии. Таксоны, находящиеся ниже корневого, являются более специфическими классификациями, которые относятся к поднаборам общего набора классифицируемых объектов. Современная биологическая классификация, к примеру, представляет собой иерархическую систему, основание которой составляют отдельные организмы (индивидуумы), а вершину — один всеобъемлющий таксон; на различных уровнях иерархии между основанием и вершиной находятся таксоны, каждый из которых подчинён одному и только одному таксону более высокого ранга.

Примером таксономии является УДК, фрагмент которого приведен ниже.

0 Загальний відділ
00 Загальні питання науки та культури
001 Наука та знання в цілому. Організація розумової праці
002 Документація. Книги. Письменництво. Авторство
003 Системи письма та писемності
004 Комп'ютерна наука та технологія. Застосування комп'ютера
004.2 Комп'ютерна архітектура
004.3 Апаратне забезпечення комп'ютерів
004.4 Програмне забезпечення
004.5 Взаємодія людини і комп'ютера. Інтерфейс користувача
004.6 Дані
004.7 Комп'ютерні мережі
004.8 Штучний інтелект
004.9 Прикладна техніка, що базується на комп'ютерних системах. Прикладні інформаційні системи

Рис. 5. Фрагмент таксономии УДК

Следует отметить, что НЕ ВСЯКАЯ иерархия определяет таксономию. Для иерархического упорядочения понятий могут использоваться различные принципы. Например, один из них – это наличие между ними взаимосвязи «один ко многим» и при этом каждое из понятий имеет самостоятельный смысл. На рис. 6 приведен пример такой иерархической структуры «организация-сотрудник-проект».

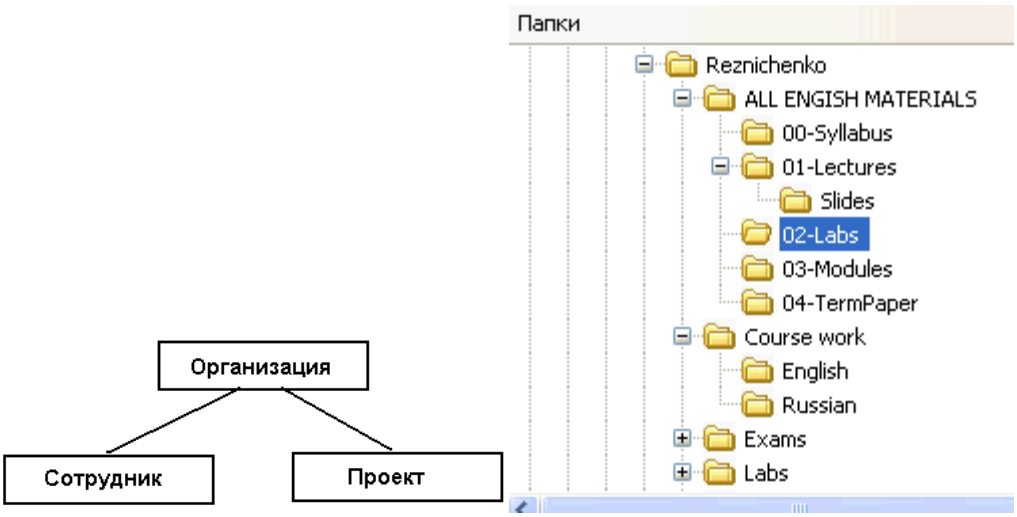

Рис. 6. Примеры иерархических структур, не являющихся таксономиям

Каждое из этих понятий является независимым в том смысле, что, например, «сотрудник» НЕ является ни «организацией» и ни «проектом». И так для любой пары понятий. Далее, в каждой организации работает множество сотрудников и имеется множество проектов, но каждый сотрудник работает в одной организации и каждый проект выполняется в одной организации. Это классическое понимание иерархии (дерева) с точки зрения **иерархической модели данных в базах данных**. Другим примером иерархии, не являющейся таксономией, является структура папок в операционной системе компьютера. Хотя иерархическая структура папок более приближена к понятию таксономии в связи с тем, что эта иерархия определена на одном множестве объектов, а именно – на множестве папок, тем не менее это все же не таксономия.

Предложенные выше два варианта понятия иерархического упорядочения несут очень слабую и абсолютно другую семантику по сравнению с иерархическим упорядочением в таксономии.

Если каждый таксон определить посредством «интенсионала» - множества индивидов, которые принадлежат таксону, и «интенсионала» - множества свойств (атрибутов, характеристик,), которые характеризуют экстенсионал таксона, то таксономическое отношение между таксоном A (верхнего уровня) и таксоном B (нижнего уровня) означает, что интенсионал A содержится в интенсионале B и экстенсионал A содержит экстенсионал B (см. рис. 7)

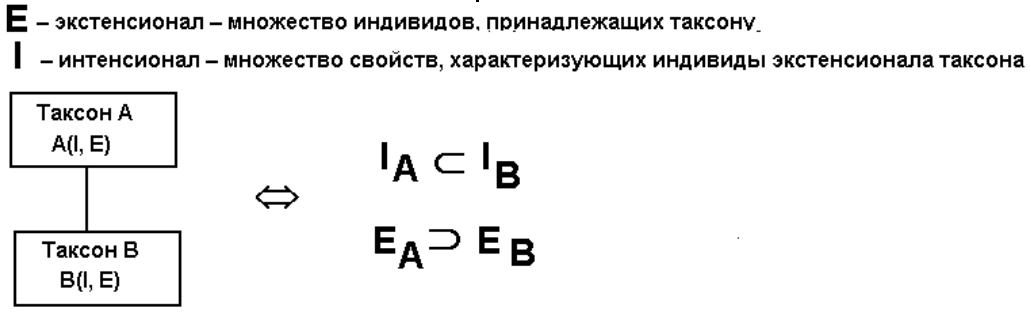

Рис. 7. Смысл таксономического иерархического упорядочения

Таксономическая иерархическая структура отличается от иерархической модели данных в базах данных следующим образом:

В иерархической модели данных каждое понятие может иметь свое собственное множество индивидов. Таксономия определяется на одном множестве индивидов.

В иерархической модели данных иерархическая связь между понятиями – это связи «один-ко многим» между индивидами каждого из понятий. В таксономии это специальная связь, смысл которой описан выше.

Таксономические отношения также называются:

отношением «обобщения/специализации»,

«родовидовым» отношением;

отношением «супертип/подтип»;

отношением «суперкласс/подкласс»;

в англоязычной литературе также говорится об отношении «Is a».

В некоторых случаях для упорядочения понятий вводят отношение «***является частью***» (***часть/целое***) и рассматривают такую структуру как ***таксономию агрегации***.

Следует отметить, что древовидные таксономии (иерархии) иногда приводят к дублированию в том случае, когда некоторый класс является подклассом двух суперклассов, или когда индивид принадлежит двум классам. Для снятия такого дублирования вводят понятие ***сетевой таксономической структуры***.

Итак, введение таксономии расширяет семантику ПО. В этом случае смысл любого понятия раскрывается через указания его взаимосвязи с другими понятиями «вверх» и «вниз» согласно заданной таксономической структуры.

2.3. Тезаурусы

Тезаурус (от греч. thesaurós — сокровище, сокровищница) в широком смысле определяется как описание системы знаний о действительности.

Тезаурус является расширением таксономии в том смысле, что в тезаурусе помимо родо-видовых отношений могут существовать любые другие отношения, которые на множестве понятий формируют сложную сетевую структуру. Тезаурусы, особенно в электронном формате, являются одним из действенных инструментов для описания отдельных предметных областей.

С точки зрения лингвистики тезаурус - это множество смысловых единиц некоторого языка с заданной на нём системой семантических отношений. Тезаурус фактически определяет семантику языка (национального языка, языка конкретной науки или формализованного языка для автоматизированной системы управления).

Например, лингвистический, тезаурус содержит (см., например, [5]):

морфологические и синтаксические свойства (часть речи, род, склонение, корень, словоформы в различных падежах, родах и числах);

семантику (значение, синонимы, антонимы, гиперонимы, гипонимы);

родственные слова;

происхождение;

фразеологизмы и устойчивые сочетания.

В 70-х гг. 20 в. получили распространение информационно-поисковые тезаурусы В этих тезаурусах выделены специальные лексические единицы — дескрипторы, по которым можно осуществлять автоматический поиск документальной информации. С каждым словом такого тезауруса сопоставляется дескриптор (дескрипторная статья),

Дескрипторная статья может иметь следующую структуру:

- заглавный дескриптор;
- ключевые слова из класса эквивалентности;
- дескрипторы, подчиняющие заглавный;
- дескрипторы, подчиненные заглавному;
- дескрипторы, ассоциированные с заглавным.

Для дескрипторов явным образом указываются семантические отношения, например, род - вид, часть - целое, цель - средство. Обычно принято разделять родовидовые, агрегативные и ассоциативные отношения.

ANSI/NISO Monolingual Thesaurus Standard (NISO - National Information Standards Organization) определяет тезаурус как:

«упорядоченный управляемый словарь, структурированный таким образом, что в нем между терминами четко определены и идентифицированы отношения эквивалентности, гомографии, иерархии и ассоциации с использованием стандартизированных индикаторов этих отношений… Первичная задача тезаурусов заключается в том, чтобы облегчить поиск документов и достигнуть согласованности в выполнении индексации письменных или другим способом полученных документов» [6]

Таким образом, согласно этому определению в тезаурусах существует четыре различных типа связей: ***эквивалентность, омонимия, иерархия, ассоциация***

Ниже в таблице 1 приводится: синонимы названий этих отношений, определение их смысла и примеры

Таблица 1

Отношения и их эквиваленты	Смысл	Примеры
эквивалентность, синоним, аналогично используется вместо	Термин X имеет тот же или почти тот же смысл, что и термин Y	«электронная библиотека» является синонимом для «цифровая библиотека»
омоним, имеет такое же написание гомография	Термин X имеет такое же написание, что термин Y, но они имеют различный смысл	Закрыть замок на замок, чтобы замок не замок
шире, чем иерархия: является родительским для	Термин X шире по смыслу, чем Y	“Организация” шире по смыслу, чем «финансовое учреждение»
уже, чем иерархия: является дочерним для	Термин X уже по смыслу, чем Y	«финансовое учреждение» ”уже по смыслу, чем «организация»
ассоциация, ассоциируется с связан с используется для	Термин X ассоциируется с термином Y. Существует какая-то неспецифицируемая связь между ими.	«гвоздь» связан с «молотком»

Далее на рис. 8 приводится пример тезауруса, который состоит из двух таксономий и двух ассоциативных связей между ними. Две таксономии встраивают иерархию двух систем понятий, а ассоциации устанавливают взаимосвязь между некоторыми парами понятий этих систем.

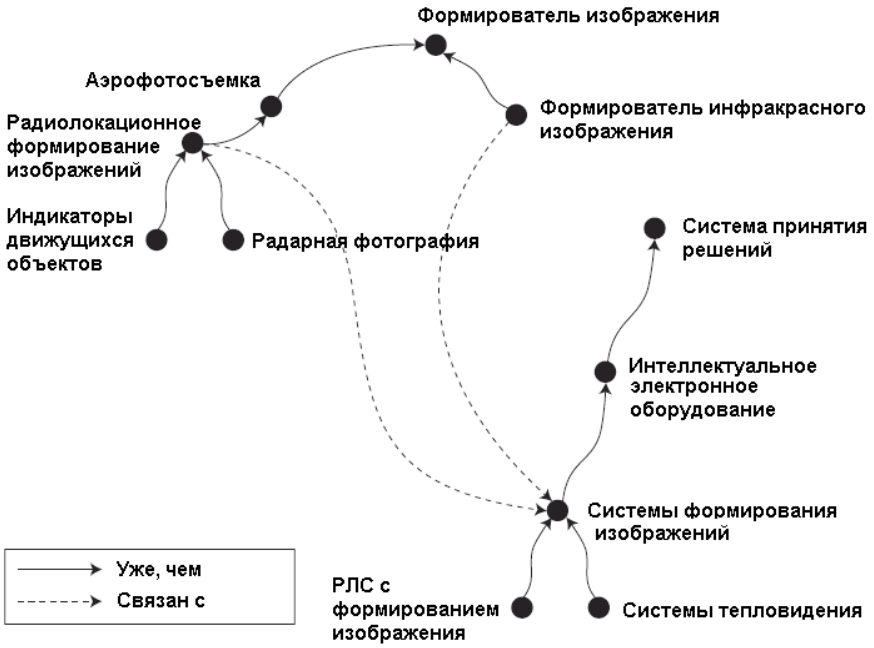

Рис. 8. Пример тезауруса с родовидовыми и ассоциативными связями

Наконец, приведем пример фрагмента тезауруса для понятия «успеваемость». Он включает два родовидовых отношения (Более узкое понятие, Более широкое понятие) и два ассоциативных отношения (Используется для, Связан с).

Таблица 2

Фрагмент тезауруса понятия «»успеваемость»

Связь	**Термин (понятие)**
Используется для	Средний балл Оценка знаний Оценка преподавания
Более узкое понятие	Успеваемость в школе Успеваемость в колледже Успеваемость в ВУЗе Успеваемость по математике
Более широкое понятие	Успех
Связан с	Мотивация успеваемости Прогнозирование успеваемости Способности к обучению Образование Учебный процесс

2.4. Онтологии

Чтобы понять, чем отличается онтология от тезауруса, следует четко понимать различие между термином и понятием (см. рис. 9, взятый из книги [7])

Термины - это строки символов, которые что-то обозначают, однако не всегда понятно, что именно. Смысл терминов раскрывается через понятия, то есть понятия - это осмысленные термины.

Раскрытие смысла, естественно, происходит через реальный (или воображаемый) мир. Таким образом, если тезаурусы имеют отношение к терминологическим системам, то онтологии - к понятийным системам.

Тезаурусы несут в себе некоторую минимальную семантику, которая раскрывается через наличие классификационной системы (управляемые словари) и задания на ней таксономий/ родовидовых отношений (отношения типа "уже чем" и "шире чем"). В тезаурусах также существует отношение эквивалентности, которые позволяет устанавливать равенство между различными терминами или утверждать, что два термина являются семантически тождественными или нет. Наконец, в тезаурусах имеется отношение

ассоциации, которое фиксирует наличие взаимосвязи между терминами, но семантика таких связей не определяется. Таким образом, тезаурусы несут в себе слабую семантику.

Задача онтологий - максимально расширить семантику терминов (понятий), их свойств и взаимосвязей между ними. Достигается это введением специальных языков для описания действующих в предметной области аксиом (правил, ограничений). Такие правила направлены на описания смысла, как самих сущностей предметной области, так и их свойств и взаимосвязей. На рис. 9 приведены два примера описания семантики. Первый с помощью соответствующей аксиомы раскрывает смысл ассоциативного отношения «сотрудничают»: два исследователя сотрудничают, если они имеют общую цель и предпринимают действия для достижения этой цели. Второй раскрывает смысл свойства индивидов «семейное положение» за счет определения множества допустимых значений этого свойства (холост, женат, разведен вдовец) - то есть экстенсионала этого понятия – и множества допустимых переходов между этими значениями.

В заключение данного раздела отметим, что контролируемые словари, классификаторы, таксономии, тезаурусы и онтологии в своей совокупности предназначены для следующего:

4. Это основа для выработки единой, согласованной, нормативной, однозначно понимаемой, полной и непротиворечивой терминологии, используемой всеми, кто имеет отношение к предметной области (ПО).

5. Это средства, предназначенные для классификации, структурирования, систематизации, моделирований и придания смысла понятиям, их свойствам и связям, относящимся к ПО.

6. Это средства для “исчерпывающего” описания информационной модели ПО, включая и ее семантику, самостоятельно, то есть не зависимо (однако, естественно, с учетом) от задач, которые будут решаться с использованием этой модели (отделение информационных знаний предметной области от оперативных знаний).

4. Это основа для коммуникации между людьми и компьютерными агентами.

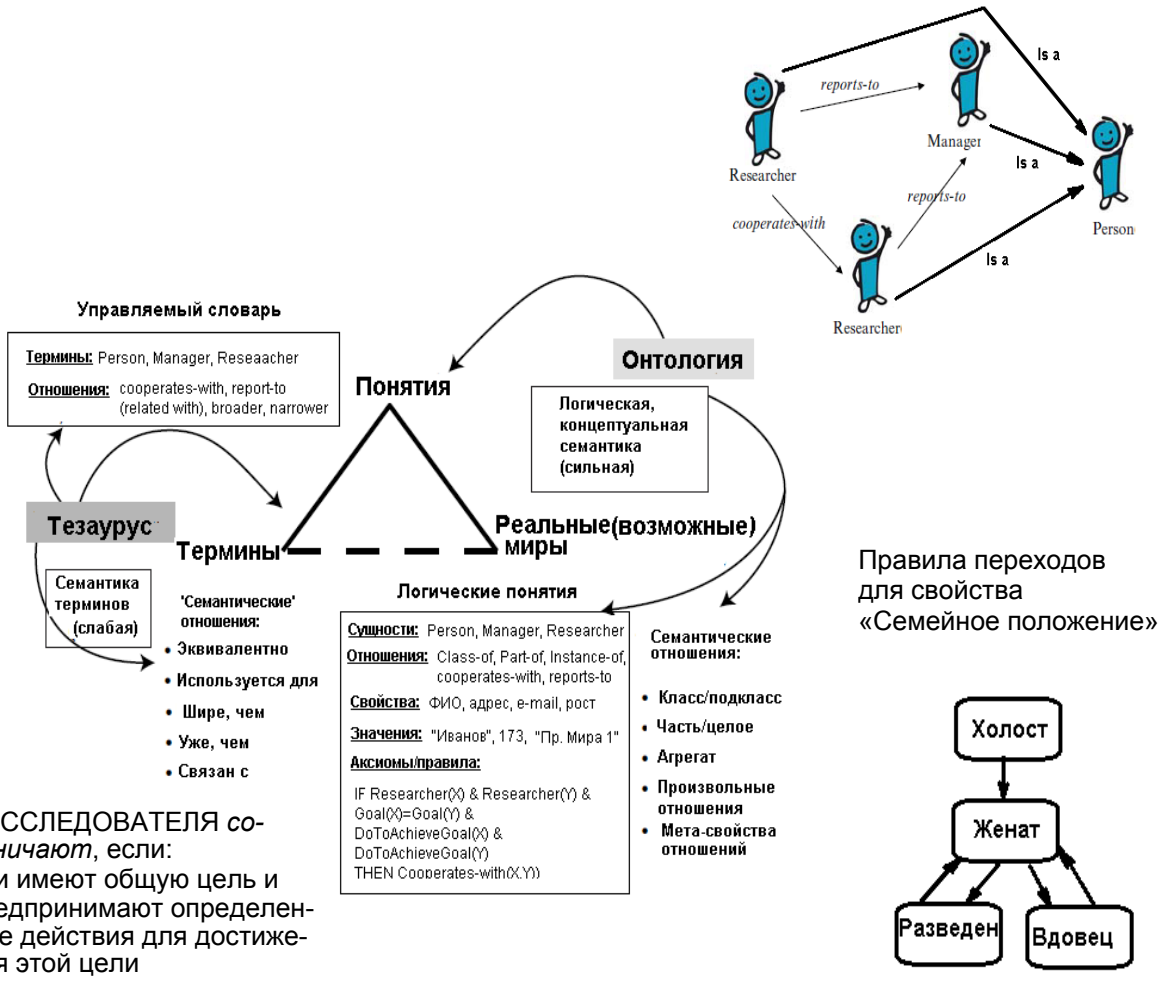

Рис. 9. Тезаурусы и онтологии

Список использованных источников

1) Jorge Cardoso. Semantic Web Services: Theory, Tools and Applications. IGI Global, Hershey, PA, March 2007.

2) Кондаков Н.И. Логический словарь-справочник.- Издательство «Наука», 1975.- 720 с.

3) Таксономия. - http://ru.wikipedia.org/wiki/%D0%A2%D0%B0%D0%BA%D1%81%D0%BE%D0%BD%D0%BE%D0%BC%D0%B8%D1%8F

4) Тезаурус. - http://slovari.yandex.ru/~%D0%BA%D0%BD%D0%B8%D0%B3%D0%B8/%D0%91%D0%A1%D0%AD/%D0%A2%D0%B5%D0%B7%D0%B0%D1%83%D1%80%D1%83%D1%81/

5) Викисловарь - многоязычный открытый словарь. - http://ru.wiktionary.org/wiki/

6) ANSI/NISO Monolingual Thesaurus Standard - ANSI/NISO Z39.19-1993 (R1998)

7) Michael C. Daconta, Leo J. Obrst, and Kevin T. Smith. The Semantic Web: A Guide to the Future of XML, Web Services, and Knowledge Management. - Indianapolis : Wiley Pub., 2003.

Сведения об авторах:

Андон Филипп Илларионович – директор Института программных систем НАН Украины, доктор физ. мат наук, академик НАН Украины.
E-mail: iss@isofts.kiev.ua

Резниченко Валерий Анатольевич – ведущий научный сотрудник Института программных систем НАН Украины, канд. физ-мат. наук, с.н.с.
E-mail: vreznichenko_47@mail.ru