УДК 004.82

УПРОЩЕННАЯ МЕТОДИКА ПОСТРОЕНИЯ МНОГОСЛОЙНОЙ ОНТОЛОГИЧЕСКОЙ МОДЕЛИ ПРЕДМЕТНОЙ ОБЛАСТИ

Савченко Андрей Павлович
к. физ.-мат. н., доцент
Кубанский государственный университет, Краснодар, Россия

В работе обсуждаются недостатки существующих языков описания онтологий и проблемы их практического использования. Формулируются требования к процессу создания онтологий, предложена упрощенная методика построения онтологической модели предметной области на базе оригинального языка SXML

Ключевые слова: СЕМАНТИЧЕСКАЯ РАЗМЕТКА, ИНФОРМАЦИОННАЯ СИСТЕМА, МНОГОСЛОЙНАЯ ОНТОЛОГИЯ, НАУЧНЫЕ ИССЛЕДОВАНИЯ, СЕМАНТИЧЕСКИЙ ПОИСК

UDC 004.82

SIMPLE TECHNIQUE OF MULTI-LAYERED DOMAIN ONTOLOGY MODELS CONSTRUCTION

Savchenko Andrey Pavlovich
Cand.Phys.-Math.Sci., associate professor
Kuban State University, Krasnodar, Russia

The drawbacks of the existing ontology languages and problems of their practical application are discussed in this paper. The requirements for the process of ontologies creating are laid down. A simplified method of constructing the ontological domain model on the basis of the original language SXML is proposed

Keywords: SEMANTIC MARKUP, INFORMATION SYSTEM, MULTI-LAYERED ONTOLOGY, RESEARCH STUDY, SEMANTIC SEARCH

Работа выполнена при финансовой поддержке РФФИ (проект №14-07-31145 мол_а).

Введение

Данная статья является частью проекта по разработке информационной модели и программного инструментария для создания системы поддержки научных исследований. Проектируемая система позволит реализовать концепцию единой информационно-коммуникационной среды, выполняющей функции интеграции субъектов научных исследований на разных уровнях: региональном (межвузовском), внутривузовском, индивидуальном.

Базовым компонентом такой информационной системы служит онтология предметной области (ПрО).Она выступает как часть базы знаний информационной системы (ИС) и определяет структуру хранения информации.

Информационная модель проектируемой системы включает следующие элементы: 1) многослойная онтология ПрО;2) информационные ресурсы;3) технологии поиска и работы с информационными ресурсами[1].

В настоящей работе излагаются основные положения методики построения базовой онтологии предметной области с помощью оригинального упрощенного языка SXML.

Технологии онтологического моделирования и семантической разметки

На сегодняшний день большинство информационных систем в той или иной степени используют словари, тезаурусы и онтологии. Под онтологией понимается структурная спецификация предметной области, т.е. формализованное представление основных понятий и связей между ними [2].Результаты эволюции программного обеспечения подтвердили необходимость использования онтологий при работе с масштабными предметными областями[3, 4].Основное предназначение онтологии– обеспечение связности информационного контента в системе. Кроме того, онтология имеет собственную ценность, поскольку содержит тезаурус, систему отношений между концептами ПрО и таким образом несет в себе сведения о ней.

Вопросы использования онтологий для создания информационных систем и структурирования информационных ресурсов исследуются в России и за рубежом уже более 20 лет. Анализ литературы позволяет выделить несколько основных направлений этих исследований:

– использование методов искусственного интеллекта для создания пространств знаний (например, инициатива KA2) [5];

– трансформация сетевых информационных технологий в направлении создания социальных сетей и баз знаний в Интернете (Freebase, микроформаты);

– применение методов онтологического инжиниринга для автоматического структурирования и семантической разметки сетевых информационных ресурсов (концепция Семантического веба, проект DBpedia)[6].

На сегодняшний день существует большое количество языков описания онтологий, среди которых можно выделить две крупные группы:

– языки с традиционным синтаксисом, основанные на логике первого порядка и принципах декларативного программирования (Cycl, Common Logic, F-Logicи др.);

– языки разметки, в основе которых лежит идея аннотирования информационных единиц, их семантической разметки (OIL, OWL, RDFS).

Один из самых прогрессивных и активно развивающихся методов моделирования ПрО–RDF, изначально разработанный для машиночитаемого описания веб-ресурсов, но позволяющий строить и универсальные информационные модели. В рамках модели RDFможно выделить несколько уровней стандартов формализации знаний:

– синтаксические стандарты: RDF/XML, SPARQL, XML, HTTP;

– языки описания онтологий и схем: OWL, Schema.org, RDFS;

– стандарты семантической разметки: микроформаты, FOAF, RDFa, микроданные.

Структура приведенных стандартов схожа со структурой сетевых протоколов в модели OSI: каждый уровень стандартов регламентирует один из уровней формализации знаний. Набор совместимых стандартов, обеспечивающих построение целостной модели можно обозначить как стек семантических стандартов. В настоящее время активно развиваются стеки на базе стандарта RDF, например, RDF/OWL/FOAF, нацеленный на создание модели машиночитаемых домашних страниц и социальных сетей.

Серьезной альтернативой стеку RDF/OWL выступает проект Schema.org, который продвигается такими лидерами в области веб-поиска, как Google, Yahoo! и Яндекс. Несмотря на то, что основным синтаксическим

стандартом в Schema.orgслужит протокол HTTPи язык HTML5, данный стек поддерживает создание онтологий и в формате RDFS. Основной целью проекта заявлена попытка обеспечить более простой способ аннотирования веб-ресурсов с помощью машинно-читаемых тегов, чем аналогичные подходы с использованием RDFa и микроформатов. Использования единого глобального словаря терминов позволяет минимизировать усилия по семантической разметке ресурсов и обеспечивает совместимость моделей, построенных разными разработчиками. К недостаткам данного подхода можно отнести тот факт, что глобальный глоссарий создается только на английском языке и его локализация не предусмотрена. Однако, на наш взгляд, именно этот стандарт наиболее перспективен с точки зрения практического использования благодаря активной поддержке крупнейших поисковых систем, которые одновременно выступают и создателями стандарта и основными «потребителями» семантической разметки.

Проведенный анализ показал, что при всем многообразии существующих языков и инструментов описания предметной области пока рано говорить об их эффективном использовании в прикладных проектах. Скорее такое многообразие свидетельствует об активных теоретических исследованиях, проводимых в этом направлении. Отсутствие же масштабных прикладных проектов связано с комплексом проблем при использовании онтологического подхода к моделированию знаний. Рассмотрим их подробнее.

Проблемы онтологического подхода к формализации знаний

Несмотря на большой выбор методов и инструментов онтологического моделирования, практическое использование онтологий ограничено рядом проблем теоретического и прикладного характера.

1.Проблема *онтологии верхнего уровня*. Возможность разделить всё сущее на концепты ставил под сомнение еще Аристотель в работе «Топика», с тех пор мало что изменилось и на вопрос возможно ли построить онтологию, охватывающую все материальные и нематериальные сущности, единого ответа до сих пор нет.

2.Проблема *унификации онтологий*. Онтологии одной ПрО, созданные в разных экспертных коллективах, могут довольно существенно различаться: начиная с использования разных терминов для обозначения одного явления и заканчивая разной структурой связей между сущностями;

3.Проблема *объединения онтологий*. Эта проблема тесно связана с предыдущей. При попытке объединить несколько неунифицированных онтологий (а таких сегодня большинство), разработанных для смежных предметных областей, возникают большие сложности, связанные с различием терминологий, принятых в разных онтологиях и др.;

4.Проблема *полноты онтологий*. Поскольку онтология должна, по возможности, полно (с учетом заданного уровня абстракции) описывать систему понятий предметной области, а количество базовых понятий даже в отдельной узкоспециализированной ПрО может измеряться сотнями и даже тысячами, практическое создание такой онтологии может представлять серьезную проблему [7].

На наш взгляд, основной барьер на пути широкого использования онтологий в информационных системах заключается в сложности процесса создания и наполнения онтологий. Ввиду огромного потенциального объема онтологических моделей единственным реальным способом их наполнения является привлечение большого количества пользователей к этому процессу. Причем, в идеале, наполнение должно происходить не *до начала работы* с системой, а *в процессе работы* с ней. Для начала работы достаточно создания некоторой единой базовой метаонтологии, имеющей ограниченный размер и приемлемую сложность, создание которой должны вы-

полнять специалисты по представлению знаний и эксперты в данной предметной области. Это позволит избежать фундаментальных противоречий в структуре онтологий при их дальнейшем наполнении пользователями.

Для обеспечения такого режима работы необходимо, чтобы процесс наполнения онтологий удовлетворял ряду условий:

– процесс должен быть естественным образом встроен в основные информационные процессы пользователей и не должен отвлекать их от решения основных задач;

– процесс должен требовать минимальных затрат ресурсов пользователей, т.е. быть достаточно простым в исполнении и удобным в использовании;

– процесс не должен требовать от пользователя привлечения дополнительной информации, кроме той, что уже известна пользователю по роду его профессиональной деятельности;

– процесс должен учитывать ограниченность знаний конечного пользователя и допускать неопределенности и/или пропуски в формализованной структуре знаний.

Анализ использования существующих методы и инструментов создания онтологий показывает, что они не удовлетворяют сформулированным условиям, соблюдение которых позволит вывести онтологические проекты за рамки экспериментов и перевести их в прикладную плоскость. К недостаткам существующих технологий семантической разметки можно отнести:

– ориентация практически всех существующих технологий разметки только на профессиональных разработчиков информационных систем, что существенно снижает возможности использования неявных знаний пользователей и их перенос в систему;

– необходимость создания веб-страниц для описания объектов и процессов в предметной области; если для информационной сети наличие

таких страниц является естественным, то для других областей это требует отдельных ресурсов и затрат;

– неэффективное использование созданной онтологии: существующие технологии предусматривают только семантическую идентификацию информационных единиц и их элементов, тогда как внутренняя структура онтологии не используется для уточнения результатов поиска, реализации ассоциативного поиска и др.

Работы по созданию онтологий и разработке программного инструментария онтологического инжиниринга ведутся в России и за рубежом давно и достаточно активно. Российские исследования и разработки находятся в русле уже освоенных в США и Европе методов и средств работы с онтологиями, усилия сконцентрированы на методических аспектах онтологического инжиниринга и развитии инфраструктуры для дальнейших исследований, формированию которой в России серьезно мешает разобщенность коллективов и отсутствие общих стандартов для онтологического инструментария. Прикладное использование онтологий чаще всего ограничено структурированием содержимого баз данных или знаний, тогда как перспективы использования онтологий значительно шире. Онтологии можно рассматривать как универсальный «каркас» для информационных систем, позволяющий проводить семантическую разметку и автоматическую обработку практически любой информации.

Подводя итог, можно сказать, что, несмотря на активное развитие направления, связанного с созданием и использованием онтологий, большинство работ носят пока научно-исследовательский или экспериментальный характер. Между тем существующие возможности использования онтологий уже сегодня можно применять для увеличения эффективности работы в информационных сетях, особенно для повышения качества информационного поиска.

В связи с этим автором разрабатывается упрощенная методика создания онтологической модели предметной области на основе оригинального языка описания Semantic XML. Цель создания методики– устранение присущих существующим методикам онтологического моделирования недостатков, которые препятствуют широкому распространению информационных систем на базе онтологий.

Основные понятия языка Semantic XML

В основе проектируемого языка семантической разметки Semantic XML лежит стандартная модель описания предметной области (ПрО), аналогичная модели RDF. Онтология представляется как совокупность сущностей и отношений между ними. Отношения задаются в виде формулы«Субъект + Предикат + Объект». Во избежание усложнения и противоречивости онтологии, проектируемая методика описания ПрО предусматривает использование только однонаправленных связей между классами.

Информационная модель предметной области строится в рамках объектно-ориентированного подхода и складывается из двух компонентов.

1. *Структура классов* ***K***.

Под классом *K* понимается абстрактная сущность, описывающая множество однотипных реальных или виртуальных сущностейe_i, для которых можно выделить общую структуру (набор свойств s_j) и характерное поведение (набор методовm_k).

Классы могут описывать множества материальных и нематериальных, статичных и динамических сущностей, протяженных во времени и пространстве. Между классами устанавливаются разнородные именованные отношения (связи). Множество типов отношений является открытым и может быть дополнено в процессе использования онтологии. В технологии также предусмотрен ряд предопределенных типов связей, использование которых позволит стандартизировать отношения между классами, что

важно для автоматической обработки онтологии. К стандартным типам относятся следующие:

– часть (partof);

– разновидность (kindof);

– синоним (knownas).

Лингвистически связь описывается простым предложением вида «Класс А имеет связь с классом В». Синтаксическая формула такой связи: «Субъект + Предикат + Объект». При этом в качестве объекта (дополнения) в таком предложении должны выступают независимые сущности, а в качестве субъекта (подлежащего) – зависимые. Предикат используется в качестве имени отношения. Например, необходимо установить связь между сущностями «Университет» и «Ректор». В данном случае «Ректор» – зависимая категория, поскольку имеет смысл только при наличии объекта управления, а вот «Университет» – независимая, поскольку может (хотя бы теоретически) существовать вне зависимости от сущности «Ректор». Таким образом, связь определяется как «Ректор управляет Университетом», а не «Университет управляется Ректором».

Набор отношений, в которых заданный класс выступает субъектом, будем называть свойствами этого класса. Среди множества свойств класса можно выделить обязательные, или *классовые признаки*, и дополнительные.

Иногда сложно определить какой из классов в отношении выступает субъектом. В этом случае рекомендуется построить дочерний класс для одного из участников отношения, чтобы снять эту неопределенность. Например, при попытке установить связь между классами «Человек» и «Учебное заведение» оба класса представляются независимыми. Однако если построить класс «Студент», который является наследником класса «Человек», то станет очевидно, он выступает в роли зависимого, т.к. обучение в каком-либо заведении служит необходимым (классовым) призна-

ком для понятия «студент», следовательно, связь будет выглядеть как «Студент обучается в Учебном заведении».

К другой категории неопределенности относится случай, когда оба участника отношения взаимозависимы и возникают сложности с поиском независимой сущности. Такую неопределенность можно снять путем введения в структуру отношения сущности-посредника. В этом качестве могут выступать физические и абстрактные объекты, коммуникационные каналы, среда, интерфейсы и др.

Например, неопределенность в отношении «Покупатель – Продавец» снимается добавлением сущности «Товар», тогда:

– «Продавец *продает* Товар»;

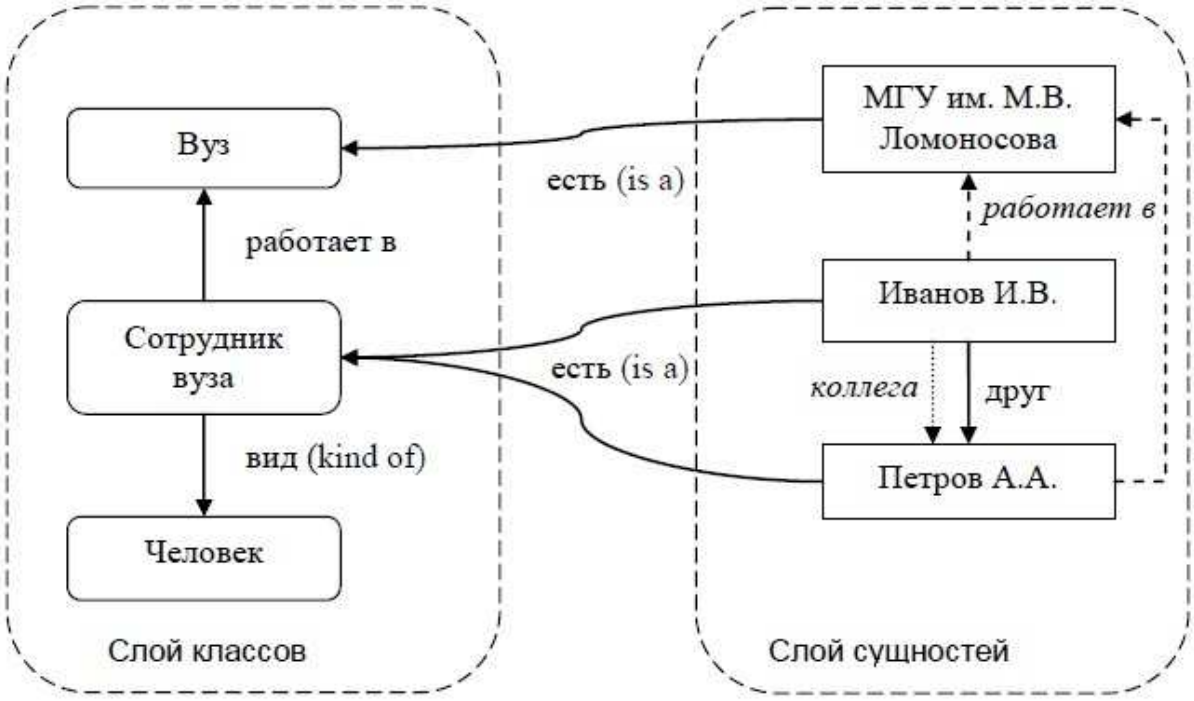
– «Покупатель *покупает* Товар».

2. *Структура сущностей (экземпляров)**E***.

Под сущностью, или экземпляром класса, понимается сущность, имеющая структуру и поведение, характерное для данного класса. Принципиальное отличие сущности от класса заключается в том, что сущность может иметь только одну уникальную реализацию в онтологии, она не может выступать объектом в отношениях типа «разновидность» или «часть», т.е. описывает только один объект или явление реального мира. Имя сущности служит ее уникальным идентификатором. Отметим, что сущность может входить одновременно в несколько классов. Например, сущность «Петров И.В.» может одновременно входит в классы «Преподаватель» и «Исполнитель НИР», при этом сущность наследует свойства обоих классов-родителей.

Между слоем сущностей Еи слоем классов Ксуществует отображение Ф, такое, что $\boldsymbol{K} = \Phi(\boldsymbol{E})$, т.е.для каждой сущности e существует класс (или ограниченное множество классов)$K = \Phi(e)$, причем отображение Ф выражает отношение структуризации вида «isa» («является»).

Между сущностями также возможно задание множества связей. При этом связи между сущностями могут быть наследственными (унаследованными от классов-родителей) и индивидуальными (определенными непосредственно в слое сущностей) (см. рисунок). Возможность создания индивидуальных связей позволит упростить структуру онтологии за счет реализации редких связей без их отражения в слое классов.



Пример структуры отношений между классами и сущностями

(→ –индивидуальные связи; − → – наследственные связи)

Таким образом, предлагаемая концепция описания ПрО обладает следующими специфическими характеристиками:

– для описания онтологии используется язык SXML, имеющий более простую структуру по сравнению с распространенными языками типа OWL или RDFS;

– процессы использования и наполнения онтологии не разделены, т.е. наполнение онтологии осуществляется не *до начала*, а *в процессе работы* с информационной системой на базе этой онтологии;

– методика описания ПрОподдерживает типизацию сущностей, механизмы наследования (включая множественное), а также описание их поведения;

– синтаксис языка допускает неопределенность в описании сущностей, в частности создание неименованных отношений, нетипизированных сущностей;

Указанные свойства языка семантической разметки и методики построения онтологии позволяют снизить сложность процесса описания предметной области и распределить его как во времени, так и в пространстве (имеется в виду разделение задачи между множеством пользователей), процесс наполнения онтологии поэтапным, более гибким и удобным для пользователя.

Однако по-прежнему актуальной остается проблема чрезмерной сложности структуры онтологии с точки зрения пользователя при описании реальных предметных областей. Для устранения этого недостатка автором разработана концепция многослойной онтологии.

Понятие многослойной онтологии

Основная идея многослойной онтологии заключается в разделении всего множества сущностей и связей между ними на так называемые слои. Под слоем понимается множество элементов онтологии, связанных отношениями определенного типас одним или несколькими ключевыми элементами, которые образуют *ядро слоя*.

Слой отражает структуру понятий, связанных с определенным направлением исследования или конкретной задачей. Пользователь получает возможность просматривать отдельные слои онтологии, отображая только информацию об интересующей его задаче и скрывая остальную. Благодаря этому даже сложные онтологии остаются легко читаемыми и

понятными для пользователя. В зависимости от характера отношений между ядром и элементами слоя можно выделить следующие типы слоев:

– функциональный слой (вычислительная модель) – элементы слоя связаны с ядром функциональными отношениями, т.е. одни информационные единицы могут быть вычислены через другие;

– каузальный слой (сценарий) – элементы слоя связаны с ядром причинно-следственными отношениями;

– структурный слой (иерархия) – элементы связаны с ядром отношениями структуризации;

– семантический слой (ситуация) – элементы связаны с ядром разнородными отношениями, но в целом описывают некоторую конечную совокупность явлений / процессов, условно названную ситуацией.

Процесс построения онтологического слоя осуществляется рекурсивно. Элементы, напрямую связанные с ядром, образуют начальный слой, который становится базисом для следующего шага по включению новых элементов и т.д. Очевидно, что мощность онтологического слоя, сформированного на заданном ядре (т.е. количество элементов в нем), зависит от глубины рекурсии в процессе расширения слоя. Глубина определяется пользователем и для больших онтологий ее следует ограничивать, дабы не столкнуться с проблемами бесконечной рекурсии, зацикленных связей и т.д. В проектируемой методике предполагается использовать не явное ограничение числа шагов рекурсии, а установку нижнего предела семантической близости элементов к ядру, при пересечении которого рекурсия останавливается.

Предлагаемая концепция многослойной онтологии представляет собой компромисс между простыми в разработке, но ограниченными в применении «учебными» онтологиями и универсальными, но сверхсложными в разработке онтологиями реального мира. Многослойная онтология позво-

ляет описывать явления и объекты сложных предметных областей в рамках одной онтологии без ее усложнения с точки зрения пользователя.

Выводы

Таким образом, в работе предложена концепция многослойной онтологии, позволяющая эффективно моделировать предметные области со сложной внутренней структурой. Для описания онтологии используется разработанный автором упрощенный язык на базе XML, ускоряющий процесс разработки онтологий и позволяющий создавать онтологии в ручном и автоматизированном режимах.

К преимуществам предложенной модели можно отнести:

– простоту и меньшую строгость синтаксиса предложенного языка описания онтологии SXMLпо сравнению с развитыми языками типа OWL;при этом функциональность языка достаточна для решения большинства практических задач с использованием онтологий;

– концепция онтологических слоев позволяет формировать своеобразные проекции сложных онтологий, такая проекция обеспечивает визуальную простоту онтологии для пользователя даже при сложной внутренней структуре.

В дальнейшем разработанную модель можно будет использовать для построения информационно-поисковых систем, реализующих интеллектуальные функции, такие как ситуативный поиск, автоматическая детализация поискового запроса и др.

Библиографический список

1. Региональный сектор экономики знаний: проблемы формирования и управления: монография/В.В. Ермоленко,М.Р. Закарян,Р.М. Закарян,Д.В. Ланская,А.П. Савченко; под ред.В.В. Ермоленко. Краснодар: Кубанскийгос. ун-т, 2013. С. 297–300.

2. Gruber T. Ontology // Encyclopedia of Database Systems / Ed. by Ling Liu, Tamer Özsu M. Springer-Verlag, 2009.

3. Ермоленко В.В. Накопление и воспроизводство интеллектуального капитала в корпорации как функция контроллинга: нейросетевой подход // Научный журнал КубГАУ. - 2010. – №04(58). – URL: http://ej.kubagro.ru/ 2010/04/pdf/07.pdf.

4.Рассел С., Норвиг П., Искусственный интеллект: современный подход. 2-еизд. М., 2006.– 1408 с.

5. Knowledge Annotation Initiative of the Knowledge Acquisition Community[Электронныйресурс]. Режимдоступа: http://hcs.science.uva.nl/usr/richard/ka2/presentation/ppframe.htm.

6. Ефименко И.В., Хорошевский В.Ф. Онтологическое моделирование экономикипредприятий и отраслей современной России. Ч. 1: Онтологическое моделирование:подходы, модели, методы, средства, решения. М.: Изд. дом ВШЭ, 2011. – 76 с.

7. Савченко А.П. Онтологический подход к представлению знаний в интеллектуальных системах: проблемы и перспективы практического применения // Актуальные проблемы управления корпорацией и человеческим капиталом в экономике знаний: сб. науч. тр. / под ред. С.Г. Фалько. Краснодар: Кубанский гос. ун-т, 2011. Вып. 3. С. 374-379.

References

1. Regional'nyj sektor jekonomiki znanij: problemy formirovanija i upravle-nija: monografija/V.V. Ermolenko,M.R. Zakarjan,R.M. Zakarjan,D.V. Lanskaja,A.P. Savchen-ko; pod red.V.V. Ermolenko. Krasnodar: Kubanskijgos. un-t, 2013. S. 297–300.

2. Gruber T. Ontology // Encyclopedia of Database Systems / Ed. by Ling Liu, Tamer Özsu M. Springer-Verlag, 2009.

3. Ermolenko V.V. Nakoplenie i vosproizvodstvo intellektual'nogo kapitala v korporacii kak funkcija kontrollinga: nejrosetevoj podhod // Nauchnyj zhurnal Kub-GAU. - 2010. – №04(58). – URL: http://ej.kubagro.ru/ 2010/04/pdf/07.pdf.

4.Rassel S., Norvig P., Iskusstvennyj intellekt: sovremennyj podhod. 2-eizd. M., 2006.¬– 1408 c.

5. Knowledge Annotation Initiative of the Knowledge Acquisition Community[Jelektronnyjresurs]. Rezhimdostupa: http://hcs.science.uva.nl/usr/richard/ka2/presentation/ppframe.htm.

6. Efimenko I.V., Horoshevskij V.F. Ontologicheskoe modelirovanie jekonomikipredprijatij i otraslej sovremennoj Rossii. Ch. 1: Ontologicheskoe modelirova-nie:podhody, modeli, metody, sredstva, reshenija. M.: Izd. dom VShJe, 2011. – 76 s.

7. Savchenko A.P. Ontologicheskij podhod k predstavleniju znanij v intellektu-al'nyh sistemah: problemy i perspektivy prakticheskogo primenenija // Aktual'nye problemy upravlenija korporaciej i chelovecheskim kapitalom v jekonomike znanij: sb. nauch. tr. / pod red. S.G. Fal'ko. Krasnodar: Kubanskij gos. un-t, 2011. Vyp. 3. S. 374-379.