УДК 004; 621.398; 681.5

Д.П. Кузнецов, С.Ю. Ржеуцкая

МЕТОД ИНТЕГРАЦИИ ОНТОЛОГИЙ РАЗНОРОДНЫХ ИСТОЧНИКОВ ДАННЫХ В АСУП

Предлагается способ интеграции разнородных источников данных в АСУП на основе базовой онтологии предметной области. Представлен алгоритм определения семантической близости концептов и пример практического использования предложенного способа интеграции.

Интеграция разнородной информации, онтология, дескрипционная логика, OWL, онтологический инжиниринг, семантическая близость.

The paper proposes the way to integrate the disparate sources of data in enterprise management systems based on the ontology of knowledge domain. The algorithm for determining the semantic proximity of concepts and an example of practical use of the offered method of integration are presented.

Integration of heterogeneous information, ontology, descriptive logics, OWL, ontological engineering, semantic proximity.

Введение

Современные промышленные предприятия имеют сложную информационную инфраструктуру, как правило, включающую множество разнородных источников данных (базы данных различных форматов, электронные таблицы, текстовые файлы и т.д.). Такое положение, сложившееся на протяжении многих лет «островковой» автоматизации, породило проблему интеграции разнородных источников данных, широко обсуждаемую в настоящее время в научном сообществе [1] – [3]. Надежным и проверенным способом решения задачи интеграции разнородных данных в АСУП является создание единого корпоративного хранилища данных (Data Warehouse).

В силу возможного (и весьма вероятного) наличия дублирования информации, а также структурных, лексических и семантических конфликтов в различных источниках данных, формирование хранилища представляет собой сложный многоэтапный процесс, который нужно рассматривать в контексте более общей задачи управления корпоративными знаниями [3]. В статье представлен метод формирования корпоративного хранилища данных промышленного предприятия, основанный на сравнении онтологий, извлеченных из разнородных источников данных, с базовой (эталонной) онтологией предметной области с целью обнаружения и разрешения конфликтов в процессе интеграции данных. Данный метод опирается на некоторые идеи, изложенные в [1] – [3], тем не менее, он является оригинальной авторской разработкой, не представленной ни в одном из известных авторам источников.

Этапы формирования корпоративного хранилища данных

Суть предложенного метода состоит в том, что в процессе интеграции онтологии, извлеченные из различных источников данных, поочередно сравниваются с базовой онтологией, которая играет роль эталона при обнаружении лексических и семантических конфликтов и сама может пополняться в процессе интеграции. Метод предполагает выполнение следующих этапов в процессе формирования корпоративного хранилища:

1. Преобразование данных разнородных источников в единый онтологический формат.
2. Разработка базовой онтологии предметной области на основе экспертных знаний и нормативных документов предметной области.
3. Нахождение орфографических и синонимических сходств концептов (понятий) и их атрибутов в базовой и интегрируемых онтологиях.
4. Нахождение семантической близости концептов интегрируемой и базовой (расширенной базовой) онтологий.
5. Подтверждение онтологом-интегратором семантической близости концептов интегрируемой и базовой (расширенной базовой) онтологий.
6. Интеграция семантически близких концептов интегрируемой и базовой (расширенной базовой) онтологий, уточнение структуры хранилища данных.
7. Передача данных в хранилище.

Безусловно, этот сложный многоэтапный процесс требует участия специалистов. Так, разработка базовой онтологии (этап 2) может быть выполнена только с помощью опытных экспертов предметной области, принятие важных проектных решений выполняется онтологом-интегратором (этап 5). Предложенный подход позволяет соединить экспертные знания информацией, извлеченной из разнородных источников данных предприятия, при этом появляется возможность полной или частичной автоматизации наиболее трудоемких этапов формирования хранилища.

Поясним кратко представленные этапы, остановившись более подробно на тех из них, для которых не имеется типовых решений.

1 этап. Первоочередной задачей является получение из интегрируемых разнородных источников данных их описаний в едином онтологическом формате, на котором доступна семантика для эффективной интеграции. Для детального описания связей и свойств онтологии был выбран диалект языка описа-

ния онтологий OWL-DL, который имеет неплохие выразительные и вычислительные возможности. При апробации метода для автоматизации данного этапа было использовано бесплатное программное средство D2RQ.

2 этап. Разработка базовой онтологии – самый трудоемкий и ответственный «ручной» этап, поскольку качество его выполнения, напрямую зависящее от квалификации и опыта экспертов предметной области, определяет успешность выполнения всех остальных работ.

Первые два этапа являются подготовительными в процессе интеграции и могут быть выполнены одновременно. После их завершения можно приступать к интеграции онтологий.

3 этап. С целью устранения некоторых лексических конфликтов, которые могут присутствовать в наименованиях концептов и их атрибутов, требуется применить один из алгоритмов нахождения орфографических сходств. На сегодняшний день их существует множество, например, метод Левенштейна, метод Q-грамм, алгоритм нечеткого поиска, алгоритм Soundex и т.д. В данной работе использован алгоритм нечеткого поиска, так как он позволяет эффективно находить сходства сущностей и атрибутов онтологий. Для сравнения концептов определялось сходство по их названию и по названию всех атрибутов. Для нахождения синонимических сходств использовался словарь синонимов.

4 этап. Нахождение семантической близости концептов онтологий

Многие концепты онтологий могут не иметь орфографических сходств, но при этом быть очень близкими по семантике. Для нахождения семантической близости концептов онтологий будем применять единую онтологическую модель, используемую в [2]:

$$O = < C,\ T,\ P,\ F,\ A >,$$

где C – концепты (основные понятия предметной области); T – частичный порядок на множестве C; P – свойства (двуместные предикататы); F – функция, которая назначает каждому элементу множества P множество элементов из множества C (с учетом их иерархии в T); A – набор аксиом онтологии (утверждения об элементах предметной области).

На основе такой онтологической модели знаний O описание ресурса (объекта) R_i может быть представлено в виде набора семантических метаданных следующей структуры:

$$M_i = (M_{ki}(O), M_{ci}(O)),$$

где $M_{ki}(O)$ – это контекстные метаданные объекта знаний, описывающие взаимосвязи объекта с другими объектами и понятиями рассматриваемой предметной области или литералами, а $M_{ci}(O)$ – контентные метаданные ресурса, описывающие информацию, которая содержатся в объекте. Контекстные метаданные соответствуют набору значений свойств понятия ($c_j \in C$), экземпляром которого является описываемый объект в онтологии, т.е.

$$M_{ki}(O) = (p_1(O_i,\ v_1),\ \cup\ p_2(O_i,\ v_2)$$
$$\cup \ldots\ \cup\ p_r(O_i,\ v_r)),$$

где утверждение $p_i(O_i,\ v_i)$ состоит из предиката (отношения) ($p_i \in P$), описанного в онтологии. Контентные метаданные $M_{ci}(O)$ описываются наборами утверждений из O_3.

$$M_{ci}(O) = \{p_1(s_1,\ v_1),\ k_1\} \cup \{p_2(s_2,\ v_2),\ k_2\} \cup \ldots$$
$$\ldots \cup\ \{p_k(s_k,\ v_k),\ k_k\},$$

где $\{p_i(s_i,\ v_i),\ k_i\}$ – кортеж, включающий утверждение $p_i(s_i,\ v_i)$ (соответствующее RDF триплету $(s_i,\ p_i,\ v_i)$) и k_i – важность данного утверждения для описания контента объекта знаний i. Подобие между метаданными Ф(MDi, MDj) может быть определено через подобие входящих в них утверждений:

$$\sum\nolimits_{I_i \in MD_i} \sum\nolimits_{I_j \in MD_j} \mathrm{sim}(T_i, T_j),$$

где $\Phi(MD_i,\ MD_j)$ – величина близости метаописаний объекта i и объекта j; $\mathrm{sim}\ (T_i,\ T_j)$ – величина близости утверждений (триплетов) T_i и T_j, входящих в сравниваемые метаописания. Величины $\mathrm{sim}\ (T_i,\ T_j)$ могут быть определены с использованием следующего выражения:

$$\mathrm{sim}(T_i, T_j) = \mathrm{sim}((c_i, r_j, i_k k_i), (c_x, r_y, i_z k_w)) =$$
$$= \frac{\mathrm{sim}(c_i, c_x) + \mathrm{sim}(r_j, r_y) + \mathrm{sim}(i_k, i_z)}{3} \cdot f(k_l, k_w),$$

где $\mathrm{sim}(c_i,\ c_k)$ – семантическая близость понятий, используемых в утверждениях; $\mathrm{sim}(r_i,\ r_y)$ – семантическая близость отношений онтологий; $\mathrm{sim}(i_k,\ i_z)$ – семантическая близость контекстных метаданных экземпляров понятий онтологий; $f(k_I,\ k_w)$ – функция учета коэффициентов важности утверждений (используются разные варианты). Для каждого понятия ($c_i \in C$) существует множество $C_{anc}(\mathrm{c_i})$, являющееся подмножеством C, и содержащее понятия, предшествующие понятию c_i, а также само понятие $C_{anc}(c_i) = \{c_j \in C \mid T_c(c_j, c_i) \vee c_j = c_i\}$.

Для оценки семантической близости двух концептов $\mathrm{sim}C\ (c_k,\ c_l))$ применим формулу $sim_c(c_k, c_l) = k \cdot \frac{|C_{anc}(c_k) \cap C_{anc}(c_l)|}{|C_{anc}(c_k) \cup C_{anc}(c_l)|}$, используемую в [1] и добавляем к ней меру близости, использующую глубину иерархии вершин концептов (ранг множества) и меру близости атрибутов иерархии. Оценивать семантическую близость концептов числителя и знаменателя на каждом ранге будем отдельно исходя из следующей формулы:

$$100\% = (\frac{Ctop}{2^{k-1}} + \frac{C(k)}{2^{k}} + \frac{C(k+1)}{2^{k+1}} + ... + \frac{C(k+n)}{2^{k+n}}) \cdot 100\% , \quad (1)$$

где n = 1,2,3…, k, где: k – ранг множества.

Поместим иерархии сравниваемых онтологий в массивы значений (Матрицы A и B) – A:(k x n), B:(1 x m) где k, l – количество рангов рассматриваемых при нахождении семантической интеграции; n, m – количество концептов рангов k, l. Алгоритм сравнения концептов использует тезаурус предметной области, для нахождения синонимов.

$$\begin{bmatrix} A(1,1), A(1,2), ..., A(1,n) \\ A(2,1), A(2,2), ..., A(2,n) \\ ..... \\ A(k,1), A(k,2), ..., A(k,n) \end{bmatrix} \overset{\text{Тезаурус}}{\cong} \begin{bmatrix} B(1,1), B(1,2), ..., B(1,m) \\ B(2,1), B(2,2), ..., B(2,m) \\ ..... \\ B(l,1), B(l,2), ..., B(l,m) \end{bmatrix}.$$

где $\cong$ – семантическая близость.

Чтобы получить семантическую близость двух рассматриваемых концептов, необходимо на основании формулы (1) вычислить взвешенную сумму семантических близостей каждого ранга Ключей и Атрибутов интегрируемой и базовой онтологий.

Получаем:

$$\mathrm{sim}_c(c_k, c_l) = \frac{1}{2+|k-l|} \cdot \sum_{n,m} \sum_{k,l} \frac{|C_{anc}(c_{k,n},) \cap C_{anc}(c_{l,m})|}{|C_{anc}(c_{k,n}) \cup C_{anc}(c_{l,m})|} +$$

$$+ \frac{1}{2+|k-l|} \cdot \sum_{k,l} (\frac{F_{k,n}(C_{anc}(A_{k,n})) \cap F_{l,n}(C_{anc}(A_{l,m})}{F_{k,n}(C_{anc}(A_k)) \cup F_l(C_{anc}(A_{l,m})}$$

$$C_{top} \in C_{anc}(c_k) \cap C_{anc}(c_l) \text{ и } C_{top} \in C_{anc}(c_k) \cup C_{anc}(c_l)$$

при любых $(c_k \in C)$ и $(c_l \in C)$, F_p – близость атрибутов иерархии, k, l – ранги множеств. $C_{top} \in C_{anc}(c_k) \cap C_{anc}(c_l)$ и $C_{top} \in C_{anc}(c_k) \cup C_{anc}(c_l)$ при любых $(c_k \in C)$ и $(c_l \in C)$.

$$\mathrm{sim}_c(c_k, c_l) \in [0;1].$$

Алгоритм нахождения семантической близости концептов на основании формулы (2) представлен на рис. 1.

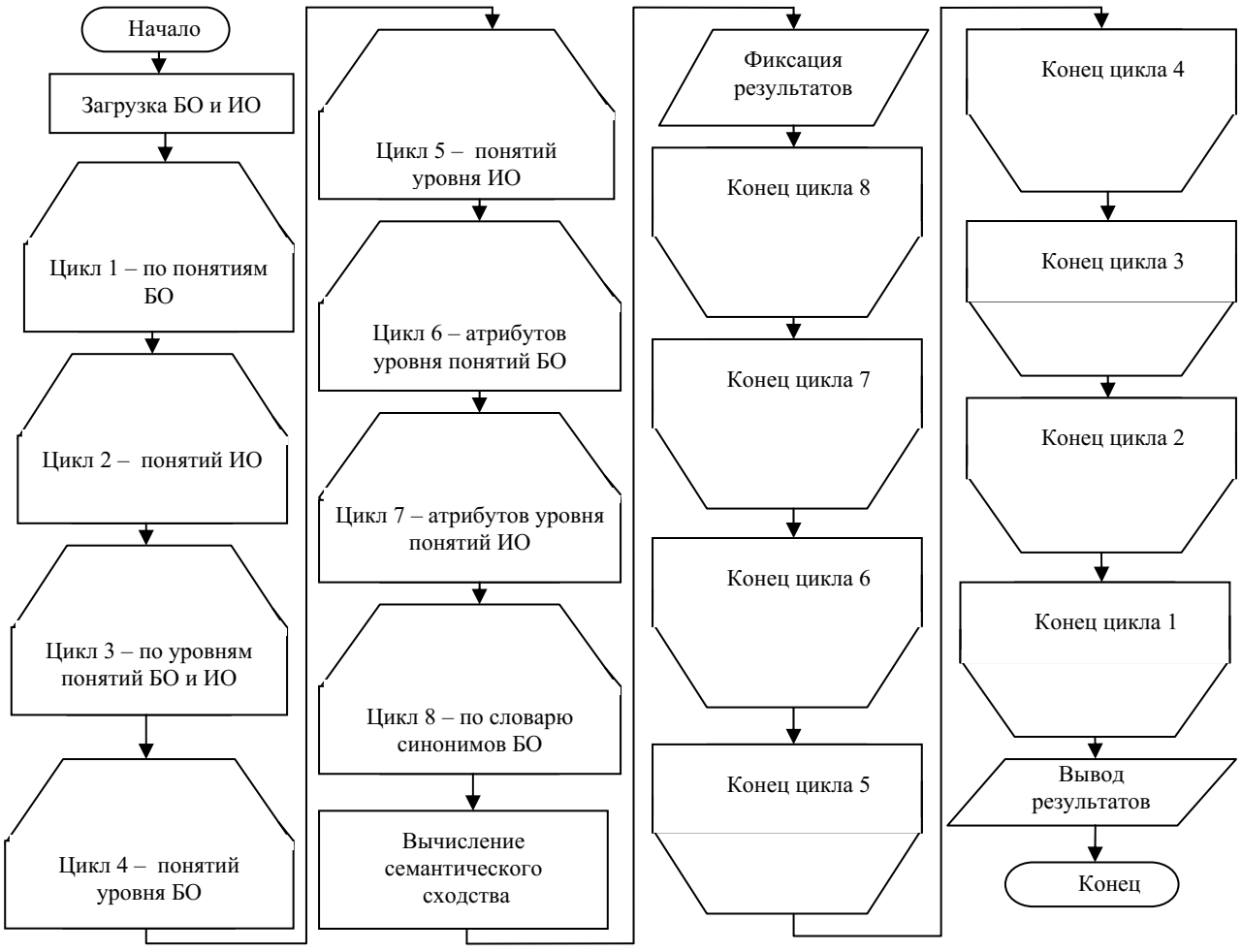


Рис. 1. Нахождение семантической близости концептов онтологий. Алгоритм

5 этап. После нахождения семантических сходств интегрируемой и базовой онтологий онтологу-интегратору предоставляются результаты поиска с вычисленным процентным соотношением близости концептов, которую он должен либо подтвердить и произвести интеграцию, тем самым изменяя структуру хранилища, либо выбрать другой вариант.

6 этап. Интеграция семантически близких концептов онтологий. Для формализации процесса интеграции использовалось расширение математической модели интеграции онтологий, представленной в [2], с учетом специфики решаемой задачи.

В рамках модели определен набор функций, необходимых в процессе интеграции онтологий: функции объединения, разделения и преобразования.

В предлагаемом методе с учетом возможности пополнения базовой онтологии была введена еще функция дополнения:

$$C_{(1,2..x,i)} \cap C_{po,j}$$
$$Fd : A'_{po} = A_{po} + (A_{po} - A_{(1,2..x)})$$
$$A_{(1,2..x)}(N_{(1,2..x)}, T_{(1,2..x)}, SV_{(1,2..x)}),$$
$$A_{po} = (N_{po}, T_{po}, SV_{po})$$
$$Fd(A_{(1,2..x,i)}, A_{po,j}) \rightarrow A'_{po}.$$

Выполнив интеграцию онтологий в соответствии с данной моделью, получим расширенную базовую модель предметной области, на основе которой можно автоматически сформировать схему хранилища данных. Основная часть задачи решена.

7 этап. Для передачи данных в хранилище необходимо сформировать запросы к источникам данных. Полная автоматизация этого процесса представляет собой трудоемкую задачу, которая может быть реализована в развитие данного метода. При наличии хорошего конструктора запросов «ручное» выполнение данного этапа не представляет сложностей.

Экспериментальная проверка предложенного способа интеграции разнородных источников данных

Эксперимент проводился на реальных данных Вологодского комбината хлебопродуктов (мукомольное и комбикормовое производство). Данное предприятие имеет несколько автономных учетных систем на различных технологических платформах с большой степенью дублирования информации, поэтому остро нуждается в интеграции данных с целью оперативного формирования качественной сводной управленческой отчетности. Целью эксперимента явилось подтверждение работоспособности и эффективности предлагаемого метода интеграции, в частности, алгоритмов нахождения семантической близости концептов и интеграции близких концептов онтологий, а конкретным прикладным результатом – сформированное единое корпоративное хранилище данных.

В ходе эксперимента были последовательно выполнены все представленные выше этапы. Совместно с экспертами предметной области была разработана базовая онтология предметной области «Хранение и переработка зерна», которая в очень укрупненном виде представлена на рис. 2.

Проведена интеграция действующих баз данных учетных систем «Корм-оптима» и «Автоматизированной системы количественно-качественного учета» (АСККУ) на основании предлагаемого метода. В результате интеграции онтологий, извлеченных из баз данных, с базовой онтологией получили расширенную онтологию предметной области (рис. 3). Как видно из схемы, базовая онтология не изменила своей структуры, а лишь дополнилась новыми знаниями из интегрируемых источников. На этой основе была сформирована структура хранилища, после чего в него была передана информация обо всей основной деятельности предприятия (после ее очистки).

Результаты эксперимента по процессу интеграции онтологий с участием базовой онтологии представлены в таблице, из которой следует, что предлагаемый метод позволил найти семантически близкие концепты и атрибуты, несмотря на различие в их наименованиях, также удалось выявить концепты и атрибуты, которые имеют одинаковое наименование, но совсем разный смысл. Общее количество найденных семантически близких к базовой онтологии концептов и атрибутов равно 145 и 26 (65 %).

Таблица

Результаты процесса интеграции онтологий с базовой онтологией

Объект	Базовая онтология	Онтология БД АСККУ	Онтология БД Корм-Оптима
Количество концептов и атрибутов	170	220	40
Количество лексически близких к базовой онтологии концептов и атрибутов	-	85	15
Количество семантически близких концептов и атрибутов	-	145	26
Количество семантически близких, но лексически различных концептов	-	64	12
Количество лексически близких, но семантически различных концептов	-	4	2

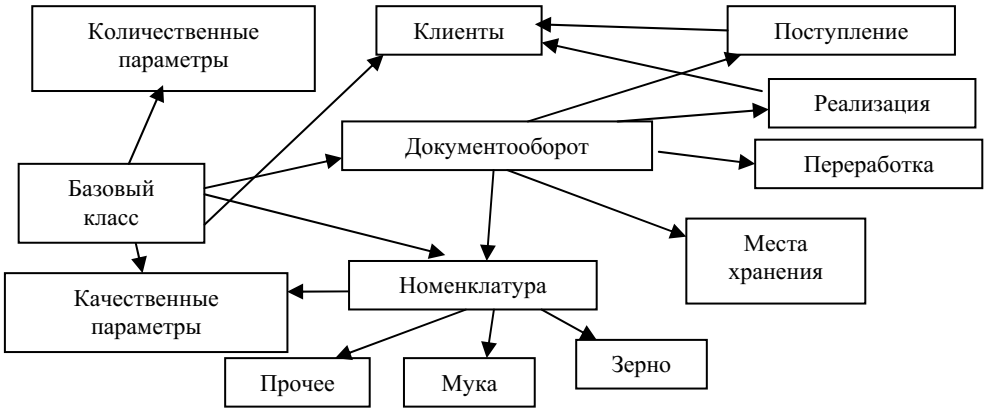


Рис. 2. Базовая онтология предметной области «Хранение и переработка зерна»

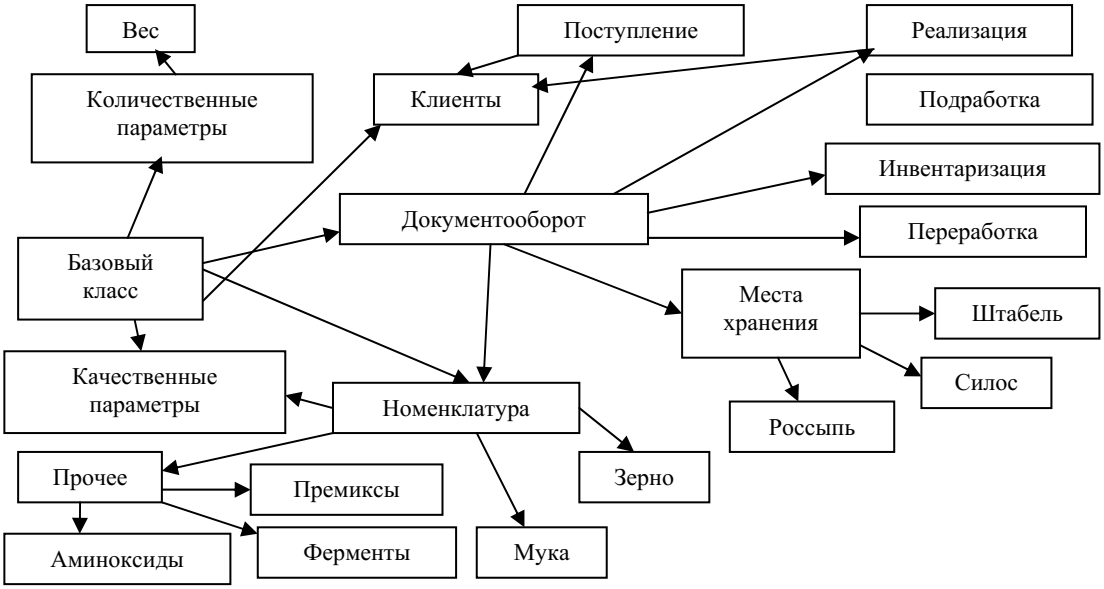


Рис. 3. Онтология предметной области «Хранение и переработка зерна» после интеграции

Важным практическим результатом эксперимента по интеграции разнородных источников данных явилось сокращение времени формирования сводной отчетности предприятия. Так, до интеграции двум бухгалтерам для получения сводного отчета по всем направлениям (комбикормовое, мукомольное) за период 2004 – 2012 гг. требовалось около 5 ч. времени. После интеграции это время сократилось до 15 мин. По одному направлению одному бухгалтеру требовалось около 4 ч. После интеграции – 10 мин. Запросы были выполнены на сервере со следующими параметрами: процессор – intel xeon 2,3 ГГц, оперативная память – 4 Гб.

Данный способ можно рекомендовать к использованию в АСУП для тех предприятий, которые решают задачу консолидации разнородных источников данных, используя экспертный опыт и современные методы управления корпоративными знаниями.

Литература

1. *Васильев, И.А.* Методы и инструментальные средства построения семантических WEB-порталов: дис. ... канд. техн. наук / И.А. Васильев. – Томск, 2005.

2. *Гаврилова, Т.А.* Онтологический инжиниринг / Т.А. Гаврилова. – URL: http://www.big.spb.ru/publications/bigspb/ (11.09.2004).

3. *Михайлов, И.А.* Математическое и программное обеспечение структурной и семантической интероперабельности информационных систем на основе метамоделей: дис. ... канд. техн. наук / И.А. Михайлов. – М., 2008.