УДК 004.89:004.4

СИСТЕМА ИНТЕГРАЦИИ ИНФОРМАЦИИ И ЗНАНИЙ С ИСПОЛЬЗОВАНИЕМ СЕМАНТИЧЕСКИХ ТЕХНОЛОГИЙ

А.Ф. Тузовский*, А.В. Черний

Институт «Кибернетический Центр» ТПУ
*Томский научный центр СО РАН
E-mail: cherny@tpu.ru

Статья посвящена применению методологии Semantic Web в организациях для решения задачи интеграции разнородных ресурсов информации и данных. Описан подход к разработке такой системы, предложенная архитектура, используемые технологии и программные продукты. Система основана на наборе онтологий и нацелена на извлечение, интеграцию, категоризацию, поиск разнородных объектов знаний организации, на основе имеющихся в организации информационных систем и документов. Система позволяет интегрировать знания нескольких организаций-партнеров. Методология и система представляют собой универсальную платформу для создания Систем Управления Знаниями и могут быть применены в любой предметной области.

Ключевые слова:
Онтология, Semantic Web, OWL, RDF, триплеты, Системы Управления Знаниями, семантическое аннотирование, семантический поиск, интеграция информации.

Key words:
Ontology, Semantic Web, OWL, RDF, triples, Knowledge management systems, semantic annotating, semantic search, information integration.

В настоящее время ведутся активные исследования и разработки в области реализации концепции Semantic Web. Данная концепция предложена консорциумом W3C в качестве новой модели развития Web сети. Основным содержанием данного подхода является использование семантических моделей (онтологий), для описания метаданных ресурсов всемирной сети. Появились стандарты языков описания онтологий и метаданных (например, RDF [1], OWL [2], SPARQL [3]), а также множество инструментов для работы с онтологиями, таких как редакторы онтологий (например, Protege) и системы логического вывода (например, Pellet, Racer, Joseki) [4]. Однако следует отметить трудность реализации концепции Semantic Web в рамках глобальной сети (Интернет) в ближайшее время, ввиду проблем с созданием и поддержкой семантических описаний документов, а также в связи со сложностью интеграции ресурсов. Но идеи и существующие технологии Semantic Web могут быть использованы для развития информационных систем (ИС) в рамках локальных сетей (интранет) организаций. В данной статье описан подход к развитию существующих информационных систем организаций с помощью семантических технологий, использующихся в глобальной сети, таких как смысловой поиск, категоризация, навигация по ресурсам, интеграция и т. д.

Описание основных понятий

Все информация и данные организации содержатся во множестве разнообразных источников R (документы, файлы, базы данных, внешние ресурсы, на которые есть ссылки) $R=\{R_1, \ldots, R_n\}$. Кроме этого, в организации имеется множество потребителей информации, (специалисты, каталоги, бизнес-процессы и входящие в них задачи), которые заинтересованы в своевременном получении некоторого вида информации.

Для организации работы с разнотипной и распределенной информацией и данными предлагается использовать онтологическое моделирование [5]. Под онтологической моделью (онтологией) $\mathcal{O}$ в данной статье понимается знаковая система $\langle C, T, P, F, L, A\rangle$, где, C – множество элементов, которые называются понятиями; T – частичный порядок на множестве C, задающий отношения «подкласс» и «суперкласс»; P – множество элементов, которые называются свойствами (двуместными предикатами); F – функция, которая назначает каждому элементу множества P множество элементов из множества C (с учетом их иерархии в T) , к которым оно применимо (область действия, область, *domain*) и множество элементов из множества C или *литералов* (экземпляров примитивных типов, таких как строки и числа), которые могут быть их значениями (область возможных значений, интервал, *range*); $L=\{L^C, L^P, \alpha^C, \alpha^P\}$ – множество текстовых меток, которые определяют профессиональные термины организации и их соответствие, а именно соответствие α^C – элементам множества C, α^P – элементам множества P; A – набор аксиом онтологии – утверждения об элементах предметной области, которые считаются верными, выраженных с использованием соответствующего логического языка.

Создание единой онтологии для детального описания модели знаний организации является весьма трудоемкой задачей. Для решения задачи построения онтологии организации эффективно использовать онтологическую модель следующей структуры [6]: $\mathcal{O}=\{\mathcal{O}_o, \mathcal{O}_p, \mathcal{O}_з\}$, где $\mathcal{O}_o$ – онтология организации, $\mathcal{O}_p$ – онтология информационных ресурсов, а $\mathcal{O}_з=\{\mathcal{O}_1, \ldots, \mathcal{O}_m\}$ – иерархически организованная, последовательно расширяемая система он-

тологий основных областей знаний $\mathcal{O}_i$, значимых для работы организации. Выделение иерархии областей знаний организации дает возможность создавать отдельно онтологии разных подобластей знаний, которые могут иметь разную детальность, в зависимости от потребностей их моделирования. Онтология организации $\mathcal{O}_o$ включает основные понятия, которые описывают структуру, состав элементов и работу организации (подразделения, специалисты, заказчики, проекты, бизнес-процессы и пр.). Онтология информационных ресурсов $\mathcal{O}_p$ включает описание всех видов ресурсов данных и информации организации (документы, файлы, базы данных, программы и пр.).

На основе такой онтологической модели знаний, описание ресурса (объекта) R_i может быть представлено в виде набора семантических метаданных [6] следующей структуры $M_i=(M_{ki}(\mathcal{O}),M_{ci}(\mathcal{O}))$, где M_{ki} $(\mathcal{O})$ – это *контекстные метаданные* объекта знаний, описывающие взаимосвязи объекта с другими объектами и понятиями организации или литералами, а $M_{ci}(\mathcal{O})$ – *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 организации, т. е. $M_{ki}(\mathcal{O})=(p_1(R_i,v_1), \vee p_2(R_i,v_2) \vee ... \vee p_r(R_i,v_r))$, где утверждение $p_i(R_i, v_i)$ состоит из предиката (отношения) $p_i \in P$, описанного в онтологии организации, *URI* (универсальный идентификатор ресурса) описываемого объекта $\mathcal{O}_i$, для которого определяются метаданные и значения v_i, которое может быть либо *URI* некоторого экземпляра понятий онтологии организации, либо некоторый литерал (текст, число, дата) в соответствии с областью действии и областями возможных значений свойства p_i. *Контентные метаданные* $M_{ci}(\mathcal{O})$ описываются наборами утверждений из $\mathcal{O}_3$, т. е. семантические метаданные объектов знаний описываются отношениями и понятиями из онтологий основных предметных областей знаний организации в виде набора кортежей: $M_{ci}(\mathcal{O})=(\{p_1(s_1,v_1),k_1\} \vee \{p_2(s_2,v_2),k_2\} \vee ... \vee \{p_k(s_k,v_k),k_k\})$, 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*. Утверждение $p_i(s_i,v_i)$ состоит из предиката (отношения) $p_i \in P$, описанного в онтологии областей знаний, объекта s_i, которым может быть понятие онтологии $c_j \in C$ или экземпляр онтологии *i* (ссылка на контекстные метаданные некоторого экземпляра онтологии) и значения v_i, которое может быть либо *URI* некоторого экземпляра, либо некоторым литералом (текстом, числом, датой).

Следует отметить, что не все экземпляры понятий онтологии будут соответствовать ресурсам информации или данных организации, например, в онтологии могут быть также такие понятия, как Специалист, Заказчик, Проект и т. п.

Совокупность описаний онтологической модели и метаданных всех ресурсов организации составляют ***онтологическую базу знаний*** системы.

Измерение близости метаописаний ресурсов

Учитывая, что модели описаний объектов знаний связаны за счет использования единой модели знаний организации, имеется возможность оценки их подобия (сходства) между собой на основе некоторой метрики подобия $\Phi(M_i,M_j)$. В статье предложены методы оценки подобия контекстных и контентных метаданных объектов знаний. Данный функционал описывает близость между отдельными элементами знаний, содержащимися в разных объектах знаний.

Формализованное представление онтологической модели знаний, а также метаописаний ресурсов создает возможность для измерения близости (подобия) ресурсов в интеллектуальном пространстве организации. Подобие между метаданными $\Phi(M_i,M_j)$ может быть определено через подобие входящих в них утверждений:

$$\Phi(M_i,M_j) = \sum_{I_i \in MD_i} \sum_{I_j \in MD_j} sim(T_i,T_j),$$

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

$$sim(T_i,T_j) = sim((c_i,r_j,i_k k_l),(c_x,r_y,i_z k_w)) =$$

$$\left(\frac{simc(c_i,c_x) + simr(r_j,r_y) + simi(i_k,i_z)}{3} \right) \cdot f(k_l,k_w),$$

где $simc(c_i,c_x)$ – семантическая близость понятий, используемых в утверждениях; $simr(r_i,r_y)$ – семантическая близость отношений онтологий; $simi(i_k,i_z)$ – семантическая близость контекстных метаданных экземпляров понятий онтологий; $f(k_l,k_w)$ – функция учета коэффициентов важности утверждений (используются разные варианты).

Оценка семантической близости двух понятий

Поясним вычисление семантической близости более подробно на примере ее оценки для двух понятий.

В онтологии на множестве понятий *C* задано отношение нестрогого – частичного порядка T_C. Утверждение T_C (c_k, c_l) означает, что c_k предшествует c_l или что c_l следует за c_k. Причем T_C задано так, что среди элементов множества *C* существует *единственный минимальный* (базовый) элемент $c_{top} \in C$. Иерархия понятий с единственной вершиной (таксономия понятий), заданная отношением T_C, используется для определения семантической близости понятий.

Для каждого понятия $c_i \in C$ существует множество $C_{ANC}(c_i)$, являющееся подмножеством *C* и со-

держащее понятия, предшествующие понятию c_i, а также само понятие c_i:

$$C_{ANC}(c_i)=\{ c_j \in C \mid T_C(c_j, c_i) \vee c_j = c_i \}.$$

Для оценки семантической близости двух понятий $sim_C(c_k, c_l)$ вводятся два показателя, основанные на сравнении множеств $C_{ANC}(c_i)$:

$$sim_C(c_k, c_l) = k_{st} * \frac{|C_{ANC}(c_k) \cap C_{ANC}(c_l)|}{|C_{ANC}(c_k) \cup C_{ANC}(c_l)|},$$

$c_{top} \in C_{ANC}(c_k) \cap C_{ANC}(c_l)$ и $c_{top} \in C_{ANC}(c_k) \cup C_{ANC}(c_l)$ при любых $c_k \in C$ и $c_l \in C$.

$$k_{st} = \begin{cases} 0, & \text{если } C_{ANC}(c_k) \cap C_{ANC}(c_l) = c_{top} \\ 1, & \text{иначе} \end{cases},$$

$$sim_C(c_k, c_l) \in [0;1].$$

На основе использования данного функционала могут быть решены разные задачи интеграции разнородной информации и данных: смысловой поиск документов (любого информационного ресурса, описанного с помощью метаданных), классификация объектов знаний по системе рубрик, рекомендация (предоставление) новых объектов знаний специалистам.

Семантический поиск

Семантический (смысловой) поиск состоит в поиске объектов базы знаний, у которых семантические метаданные близки (значение семантической близости превышает некоторое пороговое значение) семантическому описанию поискового запроса. Объектом-эталоном при семантическом поиске является поисковый запрос, представленный в виде контентных метаданных. Процедура формирования множества объектов-кандидатов для выполнения среди них семантического поиска заключается в выборе пользователем тех понятий из онтологии информационных ресурсов, которым соответствуют требуемые типы объектов. В результате семантического поиска пользователю выводится список найденных по запросу объектов, упорядоченных по величине семантической близости и дополнительно сгруппированных по типам объектов.

Категоризация

Под задачей категоризации понимается распределение ресурсов информации между иерархически организованным набором рубрик (категорий). Каждая рубрика описана семантическими метаданными, описывающими информацию, которая должна в них содержаться. Родительские и дочерние рубрики в иерархии классификации связаны между собой отношением включения (*subsumption*) между их семантическими метаданными. При решении задачи классификации в качестве объекта-эталона выступают метаданные рубрики классификатора. Из базы знаний извлекаются семантические метаданные всех объектов (за исключением рубрик), для которых нужно проверить соответствие рубрике. В процессе обработки результатов из множества кандидатов удаляются те объекты, контентные метаданные которых не содержат хотя бы одного понятия или экземпляра, присутствующего в семантических метаданных рубрики. Во множестве объектов-кандидатов остаются лишь те объекты, которые имеют показатель релевантности больше нуля, то есть относятся к рубрике.

Архитектура системы

Информационная система организации состоит из множества компьютеров объединенных в локальную сеть, которые содержат разнообразные информационные ресурсы разных форматов. Рассматриваемая в статье система позволяет интегрировать и структурировать все общие ресурсы информации и данных. Она представляет набор серверов и клиентских приложений, которые должны устанавливаться в локальной сети организации. Структура системы показана на рис. 1.

Сервера отвечают за совместное использование онтологической базы знаний и решения базовых задач по работе с онтологической моделью (редактирование и пополнение онтологии), метаданными (формирование семантических метаданных – аннотирование, хранение базы знаний) и информационными ресурсами (поиск, категоризацию, навигация по метаданным). Архитектура системы приведена на рис 2.

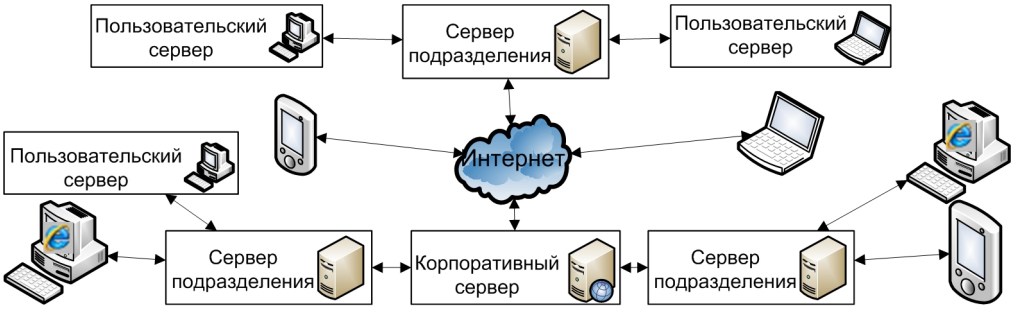

Рис. 1. *Структура системы*

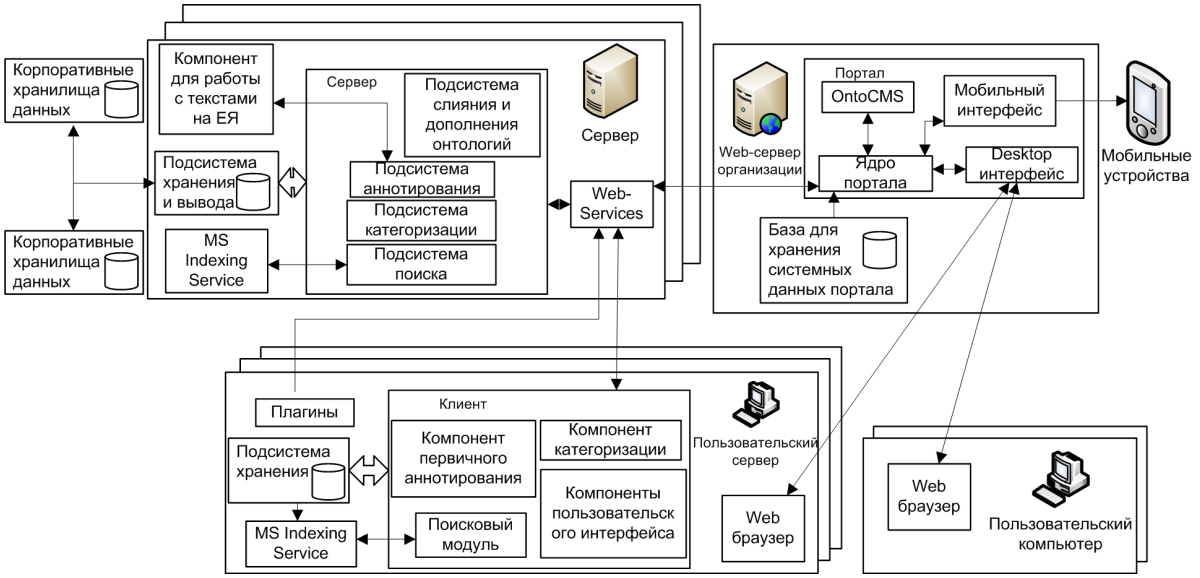

Рис. 2. *Архитектура системы*

Описание компонентов системы

В состав сервера входят подсистемы хранения и логического вывода, аннотирования, поиска, пополнения онтологии и категоризации ресурсов.

Подсистема хранения и логического вывода

Данная подсистема состоит из трех частей: модуля логического вывода, модуля хранения и модуля интеграции.

Модуль логического вывода отвечает за построение иерархий и создание новых триплетов на основе базовой онтологии за счет механизмов дескриптивной логики. Таким образом, в системе появляются связи, явно не заданные, но актуальные.

Модуль хранения обеспечивает хранение описания онтологической модели в реляционных базах данных.

Модуль интеграции

Данный модуль предоставляет функционал по интеграции онтологии системы и баз данных. Это дает возможность, не изменяя существующие в организации информационные системы, использовать данные этих ИС и иметь в онтологии всегда актуальные данные без приложения дополнительных усилий. Каждая таблица в БД ставится в соответствие с понятием онтологии. Каждый столбец этой таблицы является свойством понятия, а каждая строка – экземпляром (рис. 3).

При интеграции в свойствах каждого понятия указывается, с каким полем таблицы связано каждое свойство этого понятия. В БД добавляются триггеры, которые позволяют отслеживать обновления экземпляров в БД и актуализировать эти значения в онтологии.

Выработанная организацией базовая онтология является доступной для партнеров организации информацией (только онтология, без метаданных). Таким образом, становится возможным слияние онтологий из разных источников, что дает партнерам использовать все преимущества единого информационного пространства нескольких взаимодействующих организаций. И, как следствие, более полно описанные документы, семантически описанные совместные бизнес-процессы и т. п.

Слияние онтологий выполняется совместно менеджерами по работе со знаниями организаций и основано на свойствах языка OWL, который позволяет упростить этот процесс, находя, например, идентичные понятия, ставя их в однозначное соответствие, и так далее.

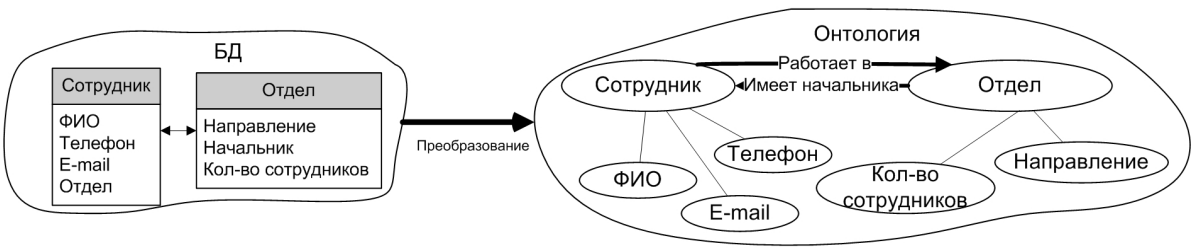

Рис. 3. *Формирование онтологической модели на основе БД*

Подсистема аннотирования

Подсистема аннотирования выделяет из документов понятия, экземпляры понятий и связи между ними. Она основывается на использовании программного обеспечения, выполняющего грамматический разбор текстов на естественном языке. На основе этого строится описание (граф) документа, который сохраняется в виде триплетов на сервере и (если позволяет формат документа) в самом документе. С каждым триплетом, полученным из документа, связывается значение, определяющее соответствие триплета смыслу документа – значимость этого утверждения. По полученным аннотациям впоследствии производится семантический поиск.

Аннотации, полученные таким путем, считаются первичными, и пользователю предоставляется возможность после ознакомления с документом в дружелюбном пользовательском интерфейсе откорректировать и дополнить автоматическую аннотацию.

Подсистема поиска

Как известно, семантический поиск дает намного более релевантные результаты, чем полнотекстовый. Считается [7], что лучшим вариантом является их совмещение. В связи с этим в системе интеграции должны быть реализованы оба вида поиска, чтобы пользователь мог выбрать какой вид поиска его интересует: только семантический, только полнотекстовый или совмещенный, когда пользователь задает набор триплетов и уточняющий полнотекстовый запрос, либо указывает при наборе в строку полнотекстового запроса, что его необходимо интерпретировать и как семантический. Во втором случае система разбирает его по тем же правилам, что и тексты при аннотировании и на этой основе строит триплеты для запроса. Таким образом, пользователь получает упорядоченные результаты, которые удовлетворяют семантическому запросу и содержат искомый текст в порядке убывания релевантности.

Пользовательский клиент

Данный модуль включает возможности категоризации, поиска (как было описано ранее), а также подсистемы взаимодействия с пользователем. Главной частью этой подсистемы являются модули интеграции приложений.

Модули интеграции приложений с системой обеспечивают взаимодействие различных приложений по работе с документами с серверами системы для аннотирования и поддержки связи информации документов с онтологической базой знаний. Например, в приложении Microsoft Word модуль интеграции позволяет выделять понятия предметной области и выполнять с ними требуемые манипуляции. Например, при работе с некоторым документом, в котором есть ссылка на другой документ, имеется возможность (выполнив щелчок кнопкой «мыши» на упоминание о нем), перейти к описанию данного документа и узнать, кем и когда он был создан, кем изменялся и пр., либо сразу открыть этот документ. Такого рода модули могут быть разработаны для браузеров и офисных приложений.

Подсистема Web-сервера

Web-серверы могут устанавливаться как в подразделениях, так и могут быть вынесены на корпоративный уровень. Основной составляющей Web серверов системы интеграции является семантический портал. Он отвечает за создание и доступ к функционалу системы через web-интерфейс и состоит из базовой части и системы управления контентом на основе семантических метаданных.

Базовая часть семантического портала предоставляет набор классов и интерфейсов для реализации других модулей и компонент портала. Система является платформой для создания систем интеграции информации и данных и обеспечивает только базовый функционал семантической системы управления знаниями, востребованный в любой организации. Но для каждой организации необходимо разрабатывать дополнительные специализированные модули, предоставляющие дополнительную (специфическую для конкретной организации) функциональность, что и реализуется с помощью API.

Система управления контентом на основе семантических метаданных (ресурсов информации и данных) представляет собой Content Management System (CMS) с расширенным функционалом для работы с онтологиями и семантическими метаданными. Например, любой элемент управления на web-странице может формировать SPARQL запрос к онтологической базе знаний и представлять полученные результаты пользователю в удобном виде. На основе онтологической CMS реализуется подсистема поддержки задач бизнес-процессов необходимыми информационными ресурсами. Для поддержки бизнес-процессов необходимо создать семантическое описание каждого бизнес-процесса, то есть описать его и все задачи, из которых он состоит в понятиях онтологии. Это дает множество преимуществ. При выполнении некоторой задачи бизнес-процесса сотрудник может узнать всю информацию, полезную для ее выполнения, получить описание сотрудников организации, которые выполняли данную задачу ранее (и, при необходимости, связаться с ними), и многое другое.

Реализация системы

Для реализации описанного подхода в качестве компонент используются уже имеющиеся программные продукты. Работа с онтологической моделью (загрузка, поиск, логический вывод) в базо-

вой подсистеме выполняется с помощью пакета OWL API [8]. Для аннотирования применяются компоненты синтаксического и морфологического анализа текстов компании АОТ [9], либо более дорогие и более функциональные компоненты компании RCO [10]. В качестве системы полнотекстового поиска используется встроенный в ОС Windows сервис Microsoft Indexing Service.

Выводы

Для эффективной автоматизации работы с разнородными ресурсами информации и знаний необходимо выполнять смысловое описания (семантики) их содержания. Описание семантики информационных ресурсов и опыта накопленного сотрудниками организации (семантических метаданных) возможно только на основе достаточно выразительных онтологических моделей знаний организации и использования современных языков моделирования, таких как RDF/RDFS и OWL. Применение семантических метаданных позволяет эффективно решать набор таких стандартных задач работы с информацией и знаниями, как поиск и категоризация.

Поддержка работы сотрудников организации с такими сложными понятиями, как онтологические модели и семантические метаданные, требует создания распределенных программных систем, основанных на наборе серверов, поддерживающих создание и обновление онтологической модели знаний организации, а также создание и использование семантических метаданных.

Применение описанного подхода позволяет выполнять эффективную интеграцию ресурсов информации и знаний, создавать системы управления знаниями организаций.

СПИСОК ЛИТЕРАТУРЫ

1. Resource Description Framework Specification [Электронный ресурс]. – 1999. – режим доступа: http://www.w3.org/RDF/. – 17.09.2009.
2. OWL Web Ontology Language Overview [Электронный ресурс]. – 2004. – режим доступа: http://www.w3.org/TR/owl-features/. – 17.09.2009.
3. SPARQL Query Language for RDF (W3C Recommendation) [Электронный ресурс]. – 2008. – режим доступа: http://www.w3.org/TR/rdf-sparql-query/. – 17.09.2009.
4. Allemang D., Hendler J. Semantic Web for the Working Ontologist Modeling in RDF, RDFS and OWL – Morgan Kaufmann Publishers, 2008. – 350 p.
5. Staab S., Studer R. (eds.) Handbook on Ontologies (International Handbooks on Information Systems). – Springer Verlag, 2004. – 660 p.
6. Тузовский А.Ф. Формирование семантических метаданных для объектов системы управления знаниями // Известия Томского политехнического университета. – 2007. – Т. 310. – № 3. – С. 184–188.
7. Добров Б.В., Иванов В.В., Лукашевич Н.В., Соловьев В.Д.. Онтологии и тезаурусы: модели, инструменты, приложения: Лекция № 3 [Электронный ресурс]. – 2008. – режим доступа http://www.intuit.ru/department/expert/ontoth/3/. – 17.09.2009.
8. Documentation about OWL API [Электронный ресурс]. – 2008. – режим доступа: http://owlapi.sourceforge.net/documentation.html. – 17.09.2009.
9. Пакет документации к программному продукту RML [Электронный ресурс]. – 2006. – режим доступа: http://www.aot.ru /technology.html. – 17.09.2009.
10. Описание компонента RCO Fact Extractor [Электронный ресурс]. – 2007. – режим доступа: http://rco.ru/product.asp?ob_no=1131. – 17.09.2009.

Поступила 17.09.2009 г.