OSTIS-2012
(Open Semantic Technologies for Intelligent Systems)

**УДК 004.822:514**

# СЕМАНТИЧЕСКИЙ ПОИСК КАК СОСТАВЛЯЮЩАЯ УПРАВЛЕНИЯ ЗНАНИЯМИ В SEMANTIC WEB

Рогушина Ю.В.

*Институт программных систем НАНУ, г.Киев, Украина*

**ladamandraka2010@gmail.com**

Проанализированы проблемы, возникающие в процессе управления онтологическими знаниями в Web, в частности, связанные с интеграцией знаний из различных источников, извлечением новых знаний из доступной информации и поиском тех знаний, которые нужны пользователю для решения конкретной задачи. Предложены методы автоматизации создания метаописаний информационных ресурсов и персонификации поиска на основе тезаурусов и онтологий, характеризующих предметную область, интересующую пользователя. Предложенные методы реализованы при разработке информационно-поисковой системы МАИПС, в которой поиск персонифицируется на основе агентного подхода и онтологического анализа.
**Ключевые слова:** онтология, Semantic Web.

## ВВЕДЕНИЕ

Сегодня значительная часть современных Web-приложений являются в определенной степени интеллектуальными, то есть каким-то способом используют знание относительно соответствующей предметной области (ПрО) и способны сами создавать новые знания.

На современном этапе развития ИТ большинство интеллектуальных Web-приложений используют технологии и стандарты, разработанные в рамках проекта Semantic Web [Хорошевский, 2008]. Управление знаниями в среде Semantic Web требует разработки соответствующих средств получения, сохранения, поиска и использования знаний с учетом таких свойств среды Web, как динамичность и гетерогенность.

Центральным компонентом концепции Semantic Web является использование онтологий, которые позволяют формализовать знания о ПрО и, в отличие от XML Schema, являются представлением знаний, а не форматом сообщений. Над онтологиями можно выполнять операции логического вывода. Инструментальные средства обеспечивают создание онтологий и их связывание с различными ИР; проверку онтологий на непротиворечивость, усовершенствование онтологий; выполнение операций логического вывода над онтологиями. В существовавшем ранее программном обеспечении для создания, обработки и использования знаний эти знания не были интероперабельными, и невозможно было перенести базу знаний из одного приложения в другое без значительных переработок. Но в современных Web-приложениях для представления знаний широко используют онтологии, которые имеют существенный теоретический базис (в частности, дескриптивные логики) и обеспечивают повторное использование знаний в различных ИС.

Сегодня в рамках проекта Semantic Web уже разработан ряд стандартов и языков для управления знаниями (рис.1.).

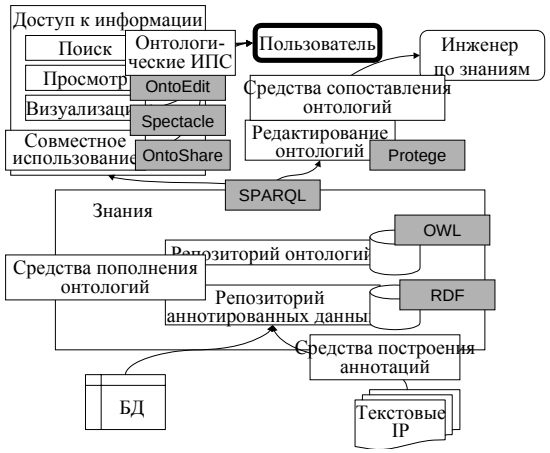


Рисунок 1 – Архитектура управления знаниями на основе Semantic Web

## ПРОБЛЕМЫ УПРАВЛЕНИЯ ЗНАНИЯМИ В WEB

Основные проблемы управления знаниями в Web связаны с: *интеграцией знаний,* полученных от разных ИР (например, интеграция онтологий

нескольких разных Про или ИР в одной Про); *поиском противоречий* между знаниями, которые отображены в контенте разных ИР, оценкой их достоверности и надежности*; получением новых знаний* из уже имеющихся и их представлением в форме, понятной пользователю; *поиском знаний, нужных конкретному пользователю* для решения определенных задач*; автоматизацией создания метаданных,* корректно описывающих контент ИР (как текстовых, так и мультимедийных) на уровне содержания, и поиском в таких метаописаниях.

Проблемы, возникающие в процессе управления знаниями в Web, сводятся к следующим четырем.

*Проблема 1*. Выбор средств представления знаний, достаточно мощных для удовлетворения потребностей пользователей, но пригодных как для быстрой автоматизированной обработки, так и для понимания человеком. Сегодня для представления знаний ПрО широко применяют онтологии, но до сих пор не предложено общепринятого определения понятия "онтология". Под онтологией ПрО обычно понимают ту часть знаний ПрО, которая предполагает ее относительную неизменность и ограничивает значение терминов ПрО .

*Проблема 2.* Создание новых знаний по имеющимся ИР (например, создание метаописаний и онтологии ИР, логический вывод, выполнение запросов к БЗ). В современном Web ИР содержат неявные нечеткие и противоречивые знания, а количество ИР вызывают необходимость в автоматизации их извлечения. Наличие языка создания метаописаний (RDF) – необходимое, но не достаточное условие формирования таких метаописаний. Чтобы автоматизировать, например, создание метаописания полнотекстового документа, нужно, во-первых, использовать методы лингвистического анализа, а во-вторых, применять формализованные знания соответствующей ПрО в интероперабельном представлении (онтологии ПрО). Кроме того, необходимо разработать специализированные методы индуктивного, традуктивного, дедуктивного вывода, ориентированные на обработку именно таких структур знаний (например, индуктивное обобщение знаний, представленных тройками RDF).

*Проблема 3.* Сопоставление различных информационных объектов на семантическом уровне (например, интеграция или поиск отличий двух онтологий, сопоставление информационного запроса и ИР, соответствующих этому запросу, определение ПрО ИР по его контенту). Эта проблема достаточно нетривиальна и не сводится к обычному поиску. Она требует анализа закономерностей ПрО, наличия средств их формального представления и разработки алгоритма сравнения таких описаний.

*Проблема 4.* Оценка качества новых знаний (достоверности, непротиворечивости, актуальности, полноты). Это требует анализа разных моделей представления знаний, использования соответствующего математического аппарата (например, теории высказываний первого порядка) и оценки качества онтологий, рассматривая для этого не только пару «онтология - реальный мир», для которой устанавливается соответствие, но и тройку - «реальный мир - неформализованное знание о мире (существующее соглашение в некотором сообществе) - формальное представление знания о мире».

## ОНТОЛОГИИ КАК СРЕДСТВО ПРЕДСТАВЛЕНИЯ ЗНАНИЙ

Анализ публикаций показывает, что именно онтологии являются адекватным и эффективным средством для моделирования представлений о разнообразных ПрО, объектах и информационных ресурсах.

В различных источниках предлагаются разные формальные модели представления онтологий. Во всех них присутствует: 1) множество *терминов* (понятий, концептов), которое может подразделяться на множество классов и множество экземпляров; 2) множество *отношений* между понятиями, в котором могут явным образом выделяться отношения «класс-подкласс», иерархические (таксономические) отношения и отношения синонимии (подобия), а также функции - специальный случай отношений, для которых n-й элемент отношения однозначно определяется n-1 предшествующими элементами; 3*) аксиомы и функции интерпретации* понятий и отношений.

Формально онтология представляется тройкой (X,R,F), где X - множество концептов, R – множество отношений между концептами, F - функции интерпретации концептов из множества X и отношений с R. Данная модель носит общий характер, в то время, как на практике пользуются более точными моделями. Например, в [Cimiano, 2006] онтология определяется как структура, которая включает идентификаторы концептов, идентификаторы отношений, идентификаторы атрибутов и типы данных, а также иерархию концептов и иерархию отношений.

В [Euzenat, 2007] онтология определяется как кортеж, который, кроме множеств классов, экземпляров, отношений и типов данных, содержит множество значений и ряд отношений (специализацию, исключение, создание экземпляра и присваивание). Для онтологии ее моделью является интерпретация, которая удовлетворяет всем утверждением онтологии.

Проанализировав выразительные возможности разных средств представления онтологий и формальные модели онтологий, можно утверждать, что существующие технологии Semantic Web предлагают разные средства описания онтологий, которые отличаются по своим выразительным возможностям и по своей сложности: RDF Schemas предоставляет простейший уровень для

представления онтологий, а OWL Full – наиболее сложный.

Выбор средства представления онтологии зависит от специфики проблемы, для которой она разрабатывается.

Таким образом, можно говорить о целесообразности управления знаниями в Web на основе онтологического анализа.

## СЕМАНТИЧЕСКИЙ ПОИСК КАК ОСНОВНАЯ СОСТАВНАЯ УПРАВЛЕНИЕ ЗНАНИЯМИ

При современном уровне развития ИТ и Web-технологий задача поиска информации трансформировалась из выявления документов, которые содержат определенные ключевые слова, в поиск знаний, необходимых для решения определенной задачи. Семантический поиск – это процесс поиска документов по их смыслу. В дальнейшем будем рассматривать семантический поиск как вид автоматизированного полнотекстового информационного поиска с учетом смыслового содержания слов и словосочетаний запроса пользователя и предложений текстов проиндексированных информационных ресурсов [Gladun, 2009].

При семантическом поиске, в отличие от обычного, предметом поиска может быть не просто ИР (документ или его фрагмент), а информационный объект определенного класса, то есть пользователь может (явно или неявно) указать класс искомого объекта. Это может быть довольно простой и распространенный класс, например, «человек» или «мультимедийный объект», либо класс, специфический для данной ПрО, например, «научная публикация».

Нередко задача семантического поиска состоит не просто в нахождении объекта определенного класса, который удовлетворяет ряду условий (например, «научный сотрудник», который работает в «организации XXX» возрастом до 30 лет), а в выявлении набора объектов разных классов, которые находятся в заданных отношениях и удовлетворяющих условия поискового запроса. При этом нередко надо использовать правила и закономерности ПрО поиска. Например, относительно простую задачу «найти группы людей разного возраста, которые проживают по одному адресу» довольно легко трансформировать в запрос «найти семьи», но чтобы запрос «разработать за 2 месяца программу автоматизации работы отдела №13, которая поддерживает одновременное обслуживание до 100 клиентов» был трансформирован в «создать группу из 3 человек в составе Иванова, Петрова и Сидорова для написания программы на C++ со следующими функциями ...», нужно применить большое количество правил, которые описывают и компетенции исполнителей, и функции отдела, и свойства программирования. Тем не менее в перспективе такие запросы тоже должны выполняться автоматизированно.

Такие сложные запросы пользователя (следует отметить, что под пользователем следует понимать как человека, так и некую программную сущность, которая имеет цели и намерения) намного труднее формализовать: 1) пользователь должен описать ту проблему, для решения которой ему нужны искомые сведения; 2) надо знать, какие именно сведения уже есть у пользователя в наличии; 3) необходимо понимать, какую информацию (в какой форме, какого уровня сложности и т.п.) пользователь способен воспринять и обработать; 4) информация может содержаться в доступных при поиске документах неявно (то есть необходимо сначала выполнить над ней операции логического вывода, обобщение, сравнение и т.п.).

Таблица 1 – Сравнение традиционных и семантических ИПС

| | Традиционные ИПС | ИПС, использующие семантический поиск |
|---|---|---|
| Запрос | набор ключевых слов | Информационная потребность в сведениях определенной ПрО |
| Персонификации поиска | История запросов пользователя | Модель пользователя и его информационных потребностей |
| Результат поиска | Документ, содержащий ключевые слова запроса | Знания, извлеченные из документов, релевантных запросу и описывающих интересующий пользователя объект |
| Источник сведений об ИР | Индекс ИПС | Индекс ИПС и метаданные о доступных ресурсах |
| Описание ПрО | - | Онтология ПрО |

Рассмотрев средства представления метаданных о гетерогенных (в том числе и мультимедийных) ИР, которые публикуются в распределенной динамической среде Web, можно сделать вывод о том, что, несмотря на многообразие подходов к отображению семантики информационных ресурсов, на современному уровне развития информационных технологий в большинстве случаев наиболее релевантным остается информационный поиск по ключевым словам. Эффективность такого поиска довольно низка, однако ее можно значительно повысить за счет использования контекста запроса и сведений о конкретном пользователе, который предоставляет запрос, и об его специфических информационных интересы, и учета предыстории его обращений с запросами к этой ИПС [Гладун, 2009]. Для

определения контекста запроса представляется целесообразным использовать онтологический подход к описанию предметной области поиска.

Как показывает анализ публикаций, один из перспективных подходов к задаче контекста поиска базируется на онтологиях, которые содержат перечень основных терминов ПрО, связи между ними и правила вывода.

## ИСТОЧНИКИ ОНТОЛОГИЧЕСКИХ ЗНАНИЙ ДЛЯ СЕМАНТИЧЕСКОГО ПОИСКА

На сегодня задача поиска извлечения из ресурсов необходимых пользователю знаний остается одной из наиболее актуальных в ИТ. При этом уже разработан ряд средств и методов, которые могут использоваться для решения этой задачи. Средствами пополнения онтологических знаний являются: непосредственное (не автоматизированное) построение онтологии или тезауруса специалистом предметной области, автоматизированная обработка метаданных об ИР, получение онтологических знаний из естественноязыковых текстов, применение методов индуктивного вывода и логические операции над существующими онтологиями.

Пользователю нужно создать онтологию той области, к которой относятся его информационные интересы, чтобы потом использовать ее при поиске наиболее пригодных ИР. Эта довольно сложная задача. Конечно, пользователь может использовать какую-то общую онтологию, которая была создана ранее другими исследователями и покрывает область его интересов. Но из-за сложности структуры и большого объема таких онтологий пользователю тяжело вносить в них изменения и дополнения. Кроме того, общие онтологии могут не соответствовать убеждениям и знаниям пользователя и не отображать его персональные предпочтения. С другой стороны, для самостоятельного создания онтологий пользователь должен не только четко представлять себе структуру интересующей его ПрО, основные ее понятие и связи между ними, но и обладать знаниями и привычками инженера по знаниям.

Поэтому, кроме редактора онтологий, который позволяет пользователю непосредственно строить онтологию, целесообразно предоставить ему определенные программные средства, способные помочь пользователю в формализации его знаний в виде онтологии и тезаурусов. Методы индуктивного вывода и лингвистический анализ позволяют извлекать термины и связи между ними из документов ПрО. Оба подхода дополняют друг друга.

Тезаурус является частным случаем онтологии, его проще формировать и обрабатывать. Тезаурус – это полный систематизированный набор данных о любой области знаний, который позволяет человеку или компьютеру в ней ориентироваться.

В связи с необходимостью анализа большого количества ИР предлагается использовать упрощенный алгоритм построения их тезауруса: по полному перечню слов, используемых в ИР, строится словарь терминов, из которого удаляются слова, помещенные в специально разработанному пользователем список («стоп»-слова). Этот алгоритм применяется только для тех ИР, которые не сопровождаются метаописаниями. В противном случае из метаописаний (в формате RDF ли OWL) приобретаются термины тезауруса и связи между ними, которые дополняют построенный по контенту ИР словарь. Потом тезаурус ИР сравнивают с тезаурусом пользователя.

Пополнение онтологии ПрО может осуществляться также и в результате лингвистического анализа текстов, выбранных пользователем в соответствия с представлениями о своих информационных потребностях. В результате семантико-синтаксического анализа в документах пользователя выделяются классы онтологии, экземпляры классов и отношения между классами и экземплярами. Для этого необходимо установить соответствие между фрагментами ЕЯ-текста (словами и словосочетаниями) и элементами онтологии (классами и отношениями).

Чтобы формализовать эту связь, создается и пополняется специальная лексическая онтология. В лексической онтологии хранятся словоформы для классов, экземпляров классов и отношений онтологии ПрО. Например, отношению «состоит из» соответствуют такие фрагменты ЕЯ-текста, как «сделан из», «изготовлен из», «содержит в себе».

Если в одном предложении встретились два фрагмента, которые связаны с экземплярами классов лексической онтологии, но в онтологии ПрО не зафиксированы отношения между этими классами, то необходимо спросить пользователя о необходимости пополнения онтологии ПрО новым отношением. Онтология пополняется новым классом, если в одном абзаце текста есть фрагменты, связанные с уже существующими классами, а также фрагменты, связанные с отношением. Тогда в предложении выделяется словоформа для нового класса онтологии.

Стандарты Semantic Web обеспечивают интероперабельное представление знаний в Web: RDF позволяет создавать метаописания ИР, в которых явным образом описывается их семантика; OWL позволяет представлять знания о ПрО в виде онтологий, которые можно использовать и обрабатывать в различных приложениях; URI позволяет однозначно идентифицировать различные ресурсы (причем не только ресурсы Web, но и абстрактные понятия и объекты реального мира) для их автоматизированной обработки; язык запросов SPARQL позволяет извлекать из метаописаний RDF и онтологий OWL необходимые пользователю сведения.

Для представления и обработки OWL

существует теоретический базис в виде семейства логик DL, обеспечивающий доказательность логического вывода на онтологиях, а различные ризонеры позволяют осуществлять на структурированных данных (OWL и RDF) логический вывод.

Уже существует достаточно большое количество структурированных метаданных, описывающие различные типы объектов, которые базируются как на стандартах Semantic Web, так и на социальных сетях (FOAF и т.д.), а также создано достаточное количество онтологий разного уровня и объема, формализующие знания самых разных ПрО.

Разработан ряд методов и инструментов для автоматизированного построения онтологий и тезаурусов по полнотекстовым ИР, существуют средства сопоставления запросов и ресурсов, ориентированные на семантический поиск Web-сервисов, которые могут использоваться и для поиска других типов ресурсов, а также есть методы и средства сопоставления онтологий (например, онтологии запроса пользователя и онтологии ИР).

Существуют ИПС, ориентированные на поиск среди структурированных данных (в частности, представленных в форматах OWL и RDF).

Но есть и ряд проблем: основная часть ИР, представленных сегодня в Web, не сопровождаются метаданными RDF (а если и сопровождаются, то доверие к этим метаданным остается открытым вопросом); построение онтологий ИР может быть автоматизировано только частично и в любом случает требует участия человека на ряде этапов, оставаясь при этом достаточно длительным и трудоемким процессом; процесс сопоставления двух независимых онтологий является сложной и трудоемкой процедурой.

Таким образом, самое простое и очевидное решение проблемы семантического поиска – построить репозиторий семантических метаописаний всех доступных в Web ИР, а потом сопоставлять их с запросом пользователя, также представленным в виде онтологии, – на сегодня не может быть реализовано. Тем не менее можно предложить следующее альтернативное решение:

Этап 1. Строится формальная модель информационной потребности пользователя (на основе онтологии интересующей его ПрО, онтологии самого пользователя и т.д.);

Этап 2. Эта модель сопоставляется с доступными структурированными данными (например, онтологиями и метаописаниями ИР), причем подобное сопоставление включает: 1) поиск соответствующих онтологий – с помощью специализированных ИПС или в собственном репозитории онтологий; 2) пополнение онтологии запроса; 3) элементы классификации запроса;

Этап 3. Постановка задачи трансформируется (в сторону конкретизации, расширения, связи с конкретными URI и т.д.) на основе анализа этих структурированных данных, и трансформированный запрос передается с помощью МАИПС к внешним ИПС;

Этап 4. Анализируются метаданные и естественно-языковой контент найденных ИР, строятся (или пополняются) онтологии этих ИР, которые затем сопоставляются с онтологией запроса;

Этап 5. Полученные результаты выполнения запроса упорядочиваются с учетом формальной модели (онтологии) запроса и знаний, полученных на предыдущем этапе.

Преимуществами предложенного двухэтапного семантического поиска является:

- используются уже существующие онтологические базы знаний, связанные с предметной областью поиска;
- для поиска соответствующих знаний используются уже существующие поисковые механизмы;
- поиск осуществляется не только среди ИР, сопровождаемых метаданными, а среди всего контента Web;
- построение онтологий ИР и их сопоставление с онтологией запроса надо осуществлять только для относительно небольшого подмножества ИР;
- при построении онтологии запроса можно использовать знания, накопленные при выполнении предыдущих запросов.

## Система семантического поиска МАИПС

Мультиагентная информационно-поисковая система МАИПС с развитыми средствами интеллектуализации ее поведения, которая детально описана в [Рогушина, 2010], функционально направлена на выполнение сложных многоразовых запросов в довольно узких областях, связанных с профессиональными или научными интересами пользователей, и предоставляет пользователю высоко релевантные результаты поиска. Такие результаты достигается благодаря ориентации системы на пользователей, имеющих в сети постоянные информационные интересы и требующих постоянного поступления соответствующей информации – для этого МАИПС позволяет сохранять и повторно выполнять запросы, отслеживать появление аналогичных запросов у других пользователей, сохранять формальное описание области интересов пользователя в виде онтологии и т.д.

Особенностью МАИПС является интегрированное использование ряда семантических технологий. Основой МАИПС являются технологии Semantic Web, в частности, язык представления онтологий OWL и средства его обработки. Для представления знаний об

интересующей пользователя ПрО используются онтологии и тезаурусы ПрО, а теоретико-множественные операции над тезаурусам позволяют более точно описывать нужную ПрО. Реализована генерация тезаурусов по естественноязыковым текстам, которые описывают семантику этих ИР, а объединение таких тезаурусов позволяет формировать тезаурус ПрО.

В МАИПС применяются технологии Web 2.0: облака тэгов используются для визуализации поисковых тезаурусов, а социальные сервисы позволяют осуществлять взаимодействие между пользователями системы со схожими интересами.

Для упорядочения информационных ресурсов, найденных системой, разработаны оригинальные алгоритмы, работающие с учетом веса (значимости) онтологических терминов для конкретного запроса. Критерии оценки уровня читабельности текста применяются для поиска информации, которая соответствует персональным потребностям пользователя [Рогушина, 2007]. Методы индуктивного вывода позволяют обобщить опыт работы МАИПС.

Для формализованного описания поведения системы используется мультиагентный подход к созданию модели интеллектуальной информационно-поисковой системы и представление компонентов системы как интеллектуальных BDI-агентов, а парадигма интеллектуальных Web-сервисов – для описания функций агентов системы для их интероперабельного использования.

Пользователь может обращаться к онтологиям, созданных другими пользователями – пересматривать их, задавать за ними контекст поиска, копировать из них нужные фрагменты, но не имеет права изменять их. ИПС может обеспечить поиск онтологий, которые содержат введенные пользователем термины, а также поиск онтологий, похожих на выбранную пользователем онтологию. Это позволяет создавать группы пользователей с общими информационными интересами и предотвратить дублирование в выполнении одинаковых многоразовых запросов разных пользователей.

## Заключение

В работе представлен подход к осуществлению семантического поиска среди ресурсов Web, который, с одной стороны, направлен на максимальное использование имеющихся средств представления и обработки распределенных знаний, а с другой – учитывает недостаточную наполненность существующего контента Web соответствующими семантическими метаописаниями и потому представляет собственные средства и методы для создания и обработки знаний о ресурсах и пользователях. В качестве основного средства представления знаний в работе используются онтологии и тезаурусы.

## Библиографический список


**[Cimiano, 2006]** Cimiano P. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications. – Springer-Verlag New York, Inc. Secaucus, NJ, USA, 2006. – 347 p.

**[Euzenat, 2007]** Euzenat J., Shvaiko P. Ontology matching. – Springer-Verlag Berlin Heidelberg, 2007. – 332 p.

**[Gladun, 2009]** Gladun A., Rogushina J. Use of Semantic Web technologies in design of informational retrieval systems // in Book "Building and Environment", 2009 Nova Scientific Publishing, New-York, USA. – P.89-103.

**[Гладун, 2009]** Гладун А.Я., Рогушина Ю.В. Использование технологии Semantic Web для интеллектуального управления в динамических распределенных системах // International Book Series "Information Sciences and Computing", 2009, Varna, Bulgarien. – P.143-153.

**[Рогушина, 2007]** Рогушина Ю.В. Показатели индивидуальной легкости чтения текста как критерий поиска информационных ресурсов в сети Интернет // УСиМ, № 3, 2007. – С.76-84.

**[Рогушина, 2010]** Рогушина Ю.В., Гришанова І.Ю. Літературний твір наукового характеру "Модель мультиагентної інформаційно-пошукової системи "МАІПС"("Модель МАІПС"). – Свідоцтво про реєстрацію авторського права на твір №32068, 2010.

**[Хорошевский, 2008]** Хорошевский, В.Ф. Пространства знаний в сети Интернет и Semantic Web (Часть 1) / В. Ф. Хорошевский // Искусственный интеллект и принятие решений. - 2008. - № 1. - С.80-97.


## SEMANTIC SEARCH AS A COMPONENT OF KNOWLEDGE MANAGEMENT IN SEMANTIC WEB


Rogushina J.V.

*Institute of software systems of National Acedemy of Sciences of Ukraine, Kiev, Ukraine*

**ladamandraka2010@gmail.com**



The problems of the ontological knowledge management in the Web are analysed, in particular relating to the integration of knowledge from various sources, acqusition of new knowledge from the available information and the retreval of knowledge that the user needs for some purpose. Methods of automatizated creation of meta-description of information resources and of search personalization based on thesauri and ontologies, describing the subject area of interest to the user, are proposed. The proposed methods are implemented in the development of an IRS MAIPS where the search is personified on base of agent approach and the ontological analysis.



**Keywords:** ontology, Semantic Web.