

ПРИМЕНЕНИЕ СЕМАНТИЧЕСКОГО ПОДХОДА ПРИ ОРГАНИЗАЦИИ МЕТАДАННЫХ

Козлов А.С.¹, Зажогин С.Д.², Кабарухин А.П.³, Ангапов В.Д.⁴

¹Козлов Александр Сергеевич - старший системный администратор,
филиал

Корпорация "Алайн Текнолоджи Ресерч энд Девелопмент, Инк", г. Москва;

²Зажогин Станислав Дмитриевич - старший разработчик,

Международный IT интегратор «Hospitality & Retail Systems»;

³Кабарухин Алексей Павлович - старший инженер сопровождения разработки,
Акционерное общество "Нэксайн";

⁴Ангапов Василий Данилович - старший системный архитектор,
Digital IQ,
г. Улан-Удэ

Аннотация: в статье проводится детальный анализ метаданных, структурированных с использованием семантического подхода. Рассматриваются популярные научные работы в рамках тематики данной статьи.

Глубоко изучается вопрос определения значимости метаданных для хранилища данных, описываются функции метаданных в хранилищах данных, в виде схемы демонстрируются компоненты метаданных аналитической системы. Детально изучается работа современных Web-порталов. Приводится их описание и классификация по различным критериям. Особенно отмечается проблематичность вопроса расширения объема обрабатываемой информации, поскольку снижается качество информационных процессов.

Предоставляются убедительные аргументы, что успешное решение такого рода проблемы возможно, если использовать семантический подход.

Осуществляется анализ современного состояния и перспективы развития семантических технологий и возможности по их использованию в Web-порталах. Доказывается, что фундаментальную роль при структуризации семантики информации занимают онтологические модели, которые не до конца исследованы при организации метаданных.

Рассматривается структура онтологий, которая необходима для обеспечения работы семантического ядра. Исследуется метод формирования семантических метаданных, который дает возможность описывать семантику необходимых объектов, что является весьма важным, поскольку позволяет решать вопросы организации метаданных разной природы происхождения.

Ключевые слова: информационные процессы, атрибуты, алгоритм, хранилища данных, метаданные, аналитическая система, Web-портал, онтологические модели, семантический подход.

APPLICATION OF SEMANTIC APPROACH IN ORGANIZATION OF METADATA

Kozlov A.S.¹, Zazhugin S.D.², Kabarukhin A.P.³, Angapov V.D.⁴

¹Kozlov Aleksandr Sergeevich - Sr. System Administrator,
"ALIGN TECHNOLOGY RESEARCH AND DEVELOPMENT INCORPORATED", EMEA RUSSIAN REGION,
MOSCOW;

²Zazhugin Stanislav Dmitrievich – Senior Software Engineer;

³Kabarukhin Aleksei Pavlovich - Senior DevOps Engineer,
NEXIGN, JSC;

⁴Angapov Vasilii Danilovich - Senior Systems Architect,
DIGITAL IQ,
ULAN-UDE

Abstract: the article provides a detailed analysis of metadata that is structured using a semantic approach. Popular scientific works within the scope of this article are considered.

The issue of determining the significance of metadata for the data warehouse is studied in depth, the functions of metadata in the data warehouse are described, and the components of the metadata of the analytical system are shown as a diagram. It describes the work of the basic modern technologies of the Web-portal and their classification by different criteria.

The work of modern Web-portals is studied in detail, and the problem of expanding the volume of processed information is especially emphasized, as the quality of information processes is reduced.

There are strong arguments that a successful solution to this kind of problem is possible if you use a semantic approach.

An analysis of the current state and prospects of the development of semantic technologies and opportunities for their use in Web portals is carried out. It is proved that ontological models, which have not been fully explored in the organization of metadata, play a fundamental role in structuring information semantics.

The structure of ontologies, which is necessary for ensuring the work of the semantic kernel, is considered. The method of formation of semantic metadata is investigated, which makes it possible to describe the semantics of the necessary objects, which is very important because it allows to solve the issues of organization of metadata of different nature of origin.

Keywords: *information processes, attributes, algorithm, data warehouses, metadata, analytical system, web portal, ontological models, semantic approach.*

УДК 004.043

Введение

Улучшение известных и изобретение новых подходов к накоплению, хранению, обработке, обмену и популяризации разного рода информации является одним из главных приоритетов развития информационных технологий (ИТ). Важность такого рода процессов во многом обусловлена постоянным непрерывным ростом баз электронных документов и их доступности, а при недостаточной структурированности хранилищ электронных метаданных возникает ряд сложностей при работе пользователя с информацией. Состав метаданных может быть разнообразным: каталоги, справочники, реестры [1-3].

Метаданные, прежде всего, содержат сведения о составе данных, содержании, статусе, происхождении, местонахождении, качестве, форматах и формах представления, условиях доступа, приобретения и использования, авторских, имущественных и смежных с ними правах на данные и др.

Существующие на сегодня подходы к работе с информацией являются недостаточно целесообразными и эффективными. С целью решения подобного рода проблем и улучшения доступа пользователя к регулярно растущему объему информации, ИТ специалистами была предложена концепция Web-порталов. Web-портал следует рассматривать, как, некую программную систему, которая имеет цель обеспечить унифицированный доступ к данным и находится во множестве разнообразных информационных источников [4].

Главным заданием Web-портала, является структурирование информации и предоставление средства для ее поиска. На данный момент различные виды Web-порталов разрабатываются и внедряются как в нашей стране, так и за рубежом. Успешная практическая реализация данного подхода и положительные перспективы усовершенствования и структуризации информационных процессов отмечается ведущими аналитиками и подтверждается пользователями.

Если говорить о работе Web-порталов, то в них информация обрабатывается на синтаксическом уровне, то есть без учета таких свойств естественного языка как синонимия, полисемия и омонимия. Такой подход к обработке информации приводит к уменьшению необходимого качества обработки информации и, в том числе, к неудовлетворительным и далеким от точных результатам поиска [2].

Для перехода на новый качественный уровень организации метаданных необходимо вести обработку информации на семантическом уровне, то есть учитывать ее смысл. Следует отметить, что современные тенденции в развитии информационных процессов, направлены на решения проблемы учета семантики в рамках существующих информационных систем.

Кроме того, данное направление исследует семантические технологии, которые предоставляют возможность создавать новый класс информационных систем.

Анализ последних исследований и публикаций. Современные исследования в области информационных технологий и систем ведутся как в нашей стране, так и за рубежом. Значительное внимание в них отведено области развития и практической реализации семантических и порталных технологических процессов с целью успешной организации работы пользователя с необходимой ему информацией.

Значительные результаты исследования по тематике данной статьи можно найти в работах Дуброва А.М., Беспаловой Г.В., Ивановой Е.С., Михалева О.В., Ковалева А., Котлера Ф., Хопкинса К. и других [1-5].

Но при этом отметим, что существует значительный разрыв по количеству и результативностью исследований в этой области между отечественным и зарубежным научным обществом.

Анализ литературных источников показывает, что в теории существуют многие правила для организации метаданных, но они не всегда подтверждаются на практике.

Статья «Metadata» Википедии на английском языке – P. Bagley. Extension of Programming Language Concepts. Philadelphia: University City Science Center, November, стоит у истоков использования термина метаданные. Проведенный статистический анализ ACM SIGMOD Anthology показал: в базах данных термин метаданные начал активно использоваться на пороге 1980-х гг., хотя в одиноких случаях и немного в другом контексте использовался и ранее. Последующий рост его использования связан с изобретением XML-технологий и концепции Semantic Web; в их понимании, он чаще всего выступает как описание контента информационных ресурсов [1].

Отметим также, что развитие и организация метаданных с помощью семантического подхода является важным этапом на пути успешного развития и внедрения семантических технологий. В этом контексте исследователи выделяют два основных направления: исследования, которые происходят в области

структуры семантических метаданных и исследования, которые возникают в области языков описания семантических метаданных.

Из литературных источников мы можем увидеть, что развитие в области структуры семантических метаданных связано с различными сферами применения семантических метаданных. То есть структура разрабатывается таким образом, что берется к вниманию определенные типы описываемых объектов [3].

Самой универсальной структурой семантических метаданных является стандарт Dublin Core. Используя данную структуру, становится возможным описание разнотипных объектов. Атрибуты, которые входят в структуру метаданных Dublin Core (DC) можно классифицировать на три группы: семантические атрибуты, права собственности и системные атрибуты. Для каждого атрибута задано его конкретное назначение, которое должно строго учитываться при использовании этой структуры.

Многочисленные результаты исследований и внедрений семантических технологий в структуру Web-порталов, можно встретить в трудах зарубежных ученых, таких как: Aaker D., Myers J., Dahlen M., Flores L., Pavlou P., Stewart D и др. Среди них следует обратить внимание на такие масштабные проекты как «OntoWeb: Ontology-based information exchange for knowledge management and electronic commerce» или «ODESeW: Automatic generation of knowledge portals for intranets and extra- nets» [1-3].

В результате изучения известных научных работ в рамках данной статьи установлено, что выполненных исследований недостаточно для организации успешной работы с метаданными с использованием семантических технологий для описания семантики контента необходимых объектов. Это можно объяснить тем, что каждый объект может быть рассмотрен в трех разных аспектах — структура, контекст и контент. В большинстве известных исследований семантические подходы используются только для описания контекста объекта, в то время как в Web-порталах в приоритете является описание семантики объектов в плане контента [4].

Изучение литературных источников показало, что тема организации метаданных на основе семантического подхода — это сравнительно новое направление в мире эпохи больших данных и требует дополнительных исследований с целью расширения известных методов при работе с информацией.

Формулирование целей статьи. Провести исследование организации структуры метаданных с использованием семантических технологий.

Изложение основного материала исследования. Многогранность определения значения метаданных можно пояснить тем, что они имеют главной целью удовлетворить технические и семантические потребности всех существующих групп пользователей информационных систем. Каждая определенная группа пользователей ИТ имеет свои требования к метаданным.

Пользователи информационных систем хотят иметь ответ на интересующий их вопрос в терминах, которые будут тривиальными для этой группы пользователей.

Например, специалисты бухгалтерских систем ожидают, чтобы такая система была адаптирована на их профессиональный язык. Аналитиков, интересуют более сложные вопросы, в частности, о природе возникновения и достоверности данных компании [2]. Разработчиков приложений интересует информация о построении модели данных для информационных систем с целью расширения или же практического внедрения дополнительных бизнес приложений.

Таким образом, видим, что присутствие многообразных интересов в метаданных порождает большое число различных элементов, которые, в целом, и составляют структуру метаданных.

Проектирование и разработка организации метаданных является одной из самых главных и трудоемких задач при разработке информационных систем.

Крупнейшее в мире хранилище информации World Wide Web (WWW), имеет свои преимущества: мгновенный доступ, качественный поиск по ключевым словам; и свои недостатки: информация отлично понятна людям, но непонятна машинам, это можно пояснить тем, что компьютеры осуществляют обработку документов как некий набор символов, но при этом не понимают их смысл [1].

Успешным решением недостатков WWW, можно назвать Semantic Web. Основным предназначением Semantic Web является поиск и интеграция данных. К основным преимуществам Semantic Web можно отнести следующие возможности: семантический поиск, объединение данных; логический вывод, агенты.

Для описания онтологий в Semantic Web используется Web Ontology Language (OWL). В системах управления мультимедийными ресурсами онтологии могут применяться для определения основных классов и понятий, а также для определения отношений между ними. Не менее важно, что при смене онтологий структура Resource Description Framework (RDF) утверждений остается неизменной. Для описания метаданных в Digital Asset Management (DAM) системе, где требуется четкость, важным является поддержка OWL-аксиом (таких, как subClassOf, disjointWith, и т. д.) и структурирование классов (unionOf, IntersectionOf и т.д.) [3].

Если изучать особенности организации метаданных и их предназначения в широком контексте, то они восходят к онтологиям. Они способны определять концепты из некой предметной области, с помощью чего машина может понимать семантику.

Если рассматривать организацию метаданных с технической стороны, то это прежде всего множество с определенными характеристиками и данными, которое предоставляет информацию о степени распространения предметной области в информационной системе, данные относительно архитектуры самой

системы и прочие существенные сведения. Кроме всего этого метаданные содержат семантическую интерпретацию относительно данных, которые находятся в информационной системе [5].

Рассмотрим наиболее важные компоненты информационной системы (ИС), представленные на рис. 1, при организации метаданных, которые характеризуют следующее:

- 1) происхождение и структуру информационных пополнений, первоисточники данных в информационном хранилище;
- 2) обработку данных при их передаче из неких оперативных источников в информационную базу;
- 3) детализацию о состоянии данных в информационной базе;
- 4) историю редактирования существующих данных в хранилище.



Рис. 1. Компоненты метаданных информационной системы

Большинство из представленных на рис. 1 компонентов принято называть обязательными атрибутами метаданных, иные – необязательными, но при этом также достаточно важными. Метрику необходимо понимать, как некую функцию, которая определяет расстояние между данными, или упорядочение информационных данных. Она позволяет определить допустимые операции над занесенной в информационное хранилище информацией (суммирование, ранжирование, иерархический порядок и т.п.). Синонимы – это ссылочные данные, которые позволяют установить конкретные ссылки на данные, например, при использовании их под разными именами несколькими пользователями.

Модель данных является граничной в концепции метаданных и определяет возможные пути использования информации из информационной базы. Регламент внесения изменений данных приводит к пополнению системы как из внутренних, так и из внешних оперативных источников.

Приведенная схема на рис. 1 позволяет разработчику и аналитику опереться на детализированную, конкретно обоснованную конструкцию, ориентированную на их потребности. Но для этого должна быть правильно выстроена вся структура дерева информационной поддержки в организации, включая оперативные информационные системы, на внутренней базе модели данных. Описанная модель данных главной своей задачей должна предусматривать интегрированный взгляд на информацию.

Автономные подсистемы должны непрерывно предоставлять возможность использования информации для последующего сохранения и презентации в информационной базе. Но при этом необходимо учитывать предметную среду информационной базы, например, для бизнес-среды это – субъекты представления, как заказчики, продукты, сделки, поставщики и продавцы.

В структуру метаданных включают описание алгоритмов вычислительной среды, существующих предметных областей, информационной безопасности, то есть все то, что имеет влияние на эффективность использования информационной системы, прежде всего, специалистами по разработке и конечными пользователями.

Если говорить о предназначении семантических метаданных, то они описывают контекст и контент объекта. В связи с этим необходимо констатировать, что все структуры метаданных предназначены для описания контекста объекта и не предназначены для его описания с точки зрения контента. Описание контента наиболее важно для тех информационных систем, которые реализуют функции полнотекстовой обработки информации.

С целью визуализации выше сказанного приведем рис. 2, где проиллюстрирована архитектура информационной системы, в которой связующим звеном являются метаданные, поскольку они существенны для приложения, а также для оперативного и неиссякаемого во времени потока данных информационной базы.



Рис. 2. Предназначение метаданных в информационной базе

При структуризации информационного хранилища главным компонентом являются метаданные, поскольку их можно охарактеризовать как данные о данных, или же своего рода каталогом информационной системы.

При разработке языков для описания характеристик метаданных нет необходимости для проведения дополнительных исследований в области их структуры, поскольку требования к структуре метаданных регламентированные. На сегодняшний день, стандартизованным и весьма популярным языком описания семантических данных является язык RDF, который базируется на триплетях и предоставляет возможность описывать метаданные с произвольной структурой, и, кроме того, данный язык удовлетворяет необходимые требования к строению метаданных, в плане, наличия конечного набора атрибутов [2].

Возможности языка RDF позволяют создавать свой список атрибутов или импортировать его из других совместимых RDF-описаний. В созданном перечне каждый атрибут должен иметь имя, поскольку RDF изобретен на стандарте URL, то названные атрибуты сохраняются с уникальностью имен. На практике каждый атрибут предназначается для определенного аспекта объекта, поскольку с помощью языка RDF невозможно выразить смысл атрибута.

Принято, что смысл атрибута обосновывается вербально разработчиками структуры семантических метаданных, но его обязательно необходимо брать во внимание при заполнении метаданных. При перемещении (импортировании) атрибутов их смысл должен быть неизменным, а создание нового атрибута вербально регламентируется.

Для успешной практической реализации с использованием онтологии сервер семантических метаданных существенно взаимодействует с сервером онтологий и предоставляет пользователю такие функции, как: 1) формирование метаданных; 2) хранение метаданных; 3) предоставление метаданных; 4) сравнение метаданных. Описанные функции представим в схематичном виде на рис. 3.

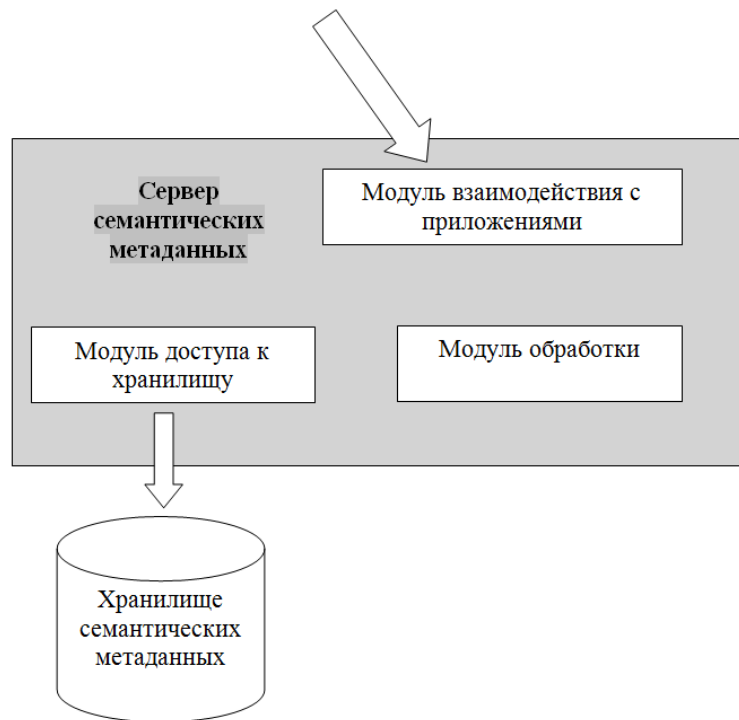


Рис. 3. Структура семантического подхода при организации метаданных

Анализ структуры, представленный на рис. 3, показывает, что модуль взаимодействия с приложениями предоставляет интерфейс доступа к функциям сервера семантических метаданных. Модуль доступа к хранилищу предоставляет непрерывное взаимодействие с подсистемой с целью хранения семантических метаданных. Модуль обработки способен интерпретировать функции для формирования семантических метаданных с использованием онтологии, а также их сравнения на семантическом уровне.

Выводы. В статье проведено глубокое исследование семантического подхода при организации структуры метаданных. На основании осуществленного анализа литературных источников показана доминирующая роль онтологических моделей для целей представления семантики информационных данных и недостаточность исследований и практических разработок в области описания семантики объектов Web-портала с точки зрения контента.

Приведены убедительные аргументы в пользу необходимости использования методов вычисления семантической близости элементов онтологии и метаданных, поскольку они предоставляют возможность количественно оценить общие семантические описания необходимых для пользователя объектов. Исходя из этого, очевидно, что при переходе к семантическим технологиям при организации метаданных будут возникать новые возможности для поиска данных разной природы происхождения.

Для лучшей визуализации, весьма значимой информации в рамках тематики исследования, представлены следующие рисунки:

- 1) Организация компонентной структуры метаданных информационной системы отражена на рис. 1.
- 2) Предназначение метаданных в информационной базе показано на рис. 2.
- 3) Структура семантического подхода при организации метаданных приведена на рис. 3.

В статье установлено, что в результате успешно спроектированной архитектуры информационной системы, максимально использующей уже существующие в системе или в сети знания, позволяет создать качественные результаты со значительно меньшими затратами. Показано, что RDF-модель базы метаданных отлично приспособлена для осуществления обмена между различными системами, позволяя легко импортировать данные или вести совместный поиск во множестве систем без необходимости определения семантического подхода при формировании метаданных.

На основе изложенного материала, можно утверждать, что использование семантики понятий, которая обосновывается за счет наличия знания в предметной области, является весьма актуальным, поскольку ее можно определить благодаря описанию отношений между понятиями. Опытные специалисты по разработке информационных систем, используя семантическую технологию метаданных, будут способны решить большинство задач при настройке информационной системы или коробочного программного продукта, не используя при этом программирование.

1. *Davenport T., Prusak L.* Working Knowledge: how organizations manage what they know. Boston: Harvard Business School Press, 1998. 200 p.
2. *Dahlen M.* Marketing on the web: Empirical studies of advertising and promotion effectiveness. Stockholm Stockholm school of economics. The Econ. research inst. (EFI), 2001.
3. *Parsaye K.* A Characterization of Data Mining Technologies and Processes // The Journal of Data Warehousing. 1998. № 1.
4. *Иванова Е.С.* Коммуникативная эффективность англоязычной рекламы: когнитивно-семантические основания: Автореф. дис. ... канд. филол. наук. М., 2002.
5. *Тузовский А.Ф.* Разработка системы управления знаниями на основе единой онтологической модели // Известия Томского политехнического университета, 2007. Т. 310. № 2. С. 182–185.