УДК 518.4+517.421.1

О. Н. Белоусова

Институт автоматики и электрометрии СО РАН
пр. Акад. Коптюга, 1, Новосибирск, 630090, Россия
E-mail: onb@ngs.ru

ОНТОЛОГИИ В СПЕЦИАЛИЗИРОВАННЫХ ТЕМАТИЧЕСКИХ ИНФОРМАЦИОННЫХ И ОБРАЗОВАТЕЛЬНЫХ ИНТЕРНЕТ-ПРОЕКТАХ*

Введение

В современном обществе, характеризующемся интенсивной компьютеризацией, развитием информационных технологий, а особенно возможностью миллионов пользователей обращаться к глобальной сети Интернет, приводит к насущной необходимости размещения информации по любой, даже узкоспециализированной области знаний, на тематических информационно-вычислительных ресурсах. При этом специфические особенности таких носителей информации как накладывают свои ограничения и требования, так и существенно расширяют возможности представления информации, т. е. это может быть текст, графика (не просто в цвете, но и анимированная), аудио, видео и даже исполняемые программы (например, демо-версии более крупных коммерческих программных продуктов). Приоритетным остается сохранение системности представления знаний, многоуровневости и иерархии размещаемой на серверах информации.

При этом прошедшая в 90-х и начале 2000-х гг. реформация базовых решений по работе с данными, построению клиент-серверных приложений, средств и платформ объектно-ориентированного программирования и компонентного проектирования и моральное устаревание систем, основанных на реляционных СУБД и технологиях построения систем, основанных на хранилищах данных, привели к необходимости построения новой методологии и соответствующих современных решений в архитектуре информационных систем.

Создание технологии двухуровневых построений по данным (данные и метаданные) с использованием онтологий документов, размещаемых на создаваемых тематических информационных интернет-ресурсах, представляется одним из таких решений.

Электронные документы и информационные системы

Информация, зафиксированная в компьютерной памяти, поскольку она уже представлена в форме, обработанной компьютерным вводом (машинные коды), тем самым является **данными**.

В отличие от «знаний», данные могут не обладать свойствами комплексности и включенности в систему взаимных связей. Данным обычно не свойственны динамические процессы, они стационарно связаны со своим носителем. Так что документ вне процесса коммуникации содержит именно **данные**, а не *знания* и не *информацию.*

Само понятие документа также нуждается в определении. По законодательству РФ, документ – это материальный объект с зафиксированной на нем информацией в виде текста, звукозаписи или изображения, предназначенный для передачи во времени и пространстве в целях хранения и общественного использования. Документ обязательно содержит *реквизиты*, позволяющие однозначно идентифицировать, содержащуюся в нем информацию.

* Работа выполняется при финансовой поддержке междисциплинарной интеграционной программы СО РАН (грант № 10) и РФФИ (грант № 07-07-00251-а).

Таким образом, документ – это не только письменный текст на бумаге, но любой носитель сведений, в частности – музейный экспонат с одной стороны, и файл на сервере, доступный через Интернет – с другой. В большинстве случаев той формой документа, которая предназначена для передачи информации, является письменный текст на естественном языке. Но это не единственный способ фиксировать информацию в документе; в последнее время все большее значение приобретает зрительная и слуховая (аудиовизуальная) информация.

На рубеже XXI в. распространилась принципиально новая форма оперативного введения информации в оборот – размещение данных в сети **Интернет**. Сюда же примыкает практика тиражирования данных на переносимых электронных носителях – сначала на магнитных лентах, затем на сменных дисках и не имеющих вращающихся частей флеш-носителях. Эти электронные документы невозможно отнести ни к заведомо опубликованным, ни к неопубликованным. Даже выпускаемые от имени солидных издательств электронные документы не имеют пока стандартной формы редакционной подготовки, поскольку такие стандарты еще не вошли в общую практику. С другой стороны, доступность электронных документов неограниченному кругу пользователей не дает права относиться к ним как к личным непубликуемым документам. По степени «физической опубликованности» следует различать следующие виды электронных документов:

– **в локальных базах данных** – имеют ограниченное распространение в пределах сети доступа к базе данных;

– **на переносимых носителях** (дисках, флеш-носителях) – подлежат копированию и распространению через торговую и библиотечную сеть наравне с книгами;

– **в общепользовательных сетях** (таких, как Интернет) – документы доступны неограниченному кругу пользователей;

– **в радио- и телевещании** – документы существуют только в форме сообщения, широко распространяемого по неконтролируемому кругу получателей.

По характеру содержимого различаются документы:

– текстовые;

– звуковые (аудиозаписи, звукозаписи);

– зрительные (визуальные) изображения, видеозаписи;

– мультимедийные (аудиовизуальные).

В состав мультимедийных документов разработчики обещают включить компоненты восприятия и другими органами чувств – обонянием, осязанием, вкусом. Подобные формы представления информации в экспериментальном порядке создаются уже сейчас; что дает право говорит об особом классе документов, относящихся к **«виртуальной реальности»**.

Аналогично печатным изданиям, электронные документы классифицированы по их тематическому содержанию, включая выделение основных классов, таких как монографии, сборники, учебники, нормативные источники, рекламные документы и более конкретные подкатегории.

Таким образом, к **информационным средствам** относятся как традиционные средства накопления и передачи информации – от устных сообщений до библиотек и видеофильмов, так и современные электронные средства представления информации, где любые данные записываются на магнитных или оптических дисках в двоичном коде с соблюдением определенных правил записи. Понятие информации в рассматриваемом смысле оказывается тесно связанными и с общим понятием **языка**, который является основным средством хранения, накопления и передачи информации.

Методы и средства информатики материализуются и доходят до потребителя в виде **новых информационных технологий**, под которыми подразумеваются современные виды информационного обслуживания, организованные на базе средств вычислительной техники и средств связи.

Формально информационная технология (ИТ) может быть определена как совокупность методов, производственных и программно-технологических средств, объединенных в технологическую цепочку, обеспечивающую сбор, хранение, обработку, вывод и распространение информации. Основное назначение ИТ – снижение трудоемкости процессов использования информационных ресурсов.

Одновременно с широким использованием новых информационных технологий появилось понятие **«информационная система»** (ИС), т. е. организационно упорядоченная совокупность документов (массивов документов) и информационных технологий, в том числе с использованием средств вычислительной техники и связи, реализующих информационные процессы. ИС осуществляет сбор, передачу и переработку информации об объекте, снабжающую специалистов различного уровня информацией для реализации функции управления.

Внедрение ИС повышает эффективность производственно-хозяйственной деятельности предприятия за счет не только обработки и хранения информации, автоматизации рутинных работ, но и за счет принципиально новых методов управления, основанных на моделировании действий специалистов при принятии решений (методы искусственного интеллекта, экспертные системы и т. п.), использовании современных средств телекоммуникации (электронная почта, телеконференции), глобальных и локальных вычислительных сетей и т. д.

По сфере применения информационные системы классифицируются следующим образом:

– ИС для научных исследований;
– ИС автоматизированного проектирования;
– ИС организационного управления;
– ИС управления технологическими процессами.

Научные ИС используются для автоматизации научной деятельности, анализа статистической информации, управления экспериментом.

Совокупность существующих документов образует информационные ресурсы (ИР).

В широком смысле ИР – это совокупность данных, организованных для эффективного получения достоверной информации.

По законодательству РФ, ИР – это отдельные документы и отдельные массивы документов, документы и массивы документов в информационных системах: библиотеках, архивах, фондах, банках данных, других видах информационных систем.

Информация в глобальных компьютерных сетях

В настоящее время эффективность использования глобальных информационных ресурсов в образовательных и исследовательских целях напрямую зависит от знаний, умений и навыков пользователей в области организации и проведения поиска информации в сети.

Не обсуждая ограничений, порожденных техническими аспектами функционирования Интернета, остановимся на проблематике, связанной с использованием Интернета как источника специализированной информации.

Концепция «информационного взрыва» и связанного с ним «информационного коллапса», активно обсуждаемая аналитиками с середины XX в., в последние годы приобрела качественно иное звучание.

Так, согласно оценкам компании IDC, в 2006 г. в мире был произведен 161 млрд гигабайт (161 экзабайт) данных. Для сравнения приведем тот факт, что в 2003 г. специалисты Университета Беркли (Калифорния, США) оценили объем произведенных за год данных в 5 экзабайт, причем в расчет принималась вся информация в целом, включая бумажные носители. Общие объемы архивов данных, хранящихся на сегодня в мире, IDC оценивает в 185 экзабайт (185 млрд гигабайт).

Обращает на себя внимание как достаточно высокая степень дублирования информации – объемы уникальной информации, произведенной в 2006 г., оцениваются в 40 экзабайт, так и тот факт, что около половины всей хранимой информации на компьютерах всего мира либо вообще не используется, либо используется однократно.

Тем не менее, даже если исключить дублируемую информацию, 40 экзабайт данных составляют примерно 300 000 объемов крупнейшей в мире Библиотеки Конгресса США.

Кроме того, впервые в истории наблюдений темпы роста объемов хранимой информации в процентном отношении превысили темпы роста объемов производимых носителей информации, таких как DVD-диски и жесткие диски.

Для сравнения приведем оценки количества веб-сайтов в конце 2006 г. (100 млн сайтов, половина из них обновляется) и 10 лет назад (18 тыс. сайтов).

Столь взрывной рост объема данных обусловлен множеством факторов, главными из которых могут быть названы следующие:

– популяризация цифрового контента, в особенности видеоконтента, произошедшее повсеместное открытие в Интернете пользовательских видеоархивов. Объем неструктурированных данных растет особенно быстрыми темпами. Уже сейчас цифровые изображения, голосовые пакеты и музыкальные записи составляют 95 % всей информации. Но такую информацию очень трудно искать. IDC полагает, что эту проблему можно решить тремя способами: добавлением метаданных, применением средств автоматической классификации (например, распознавания лиц) и разработкой систем доступа, переводящих неструктурированные данные в структурированную форму;

– законодательные требования некоторых стран, обязавшие телекоммуникационных и интернет-провайдеров хранить данные и журналы активности интернет-пользователей;

– стремительный рост покупок с помощью электронных карт и их аналогов, бурное развитие «цифровой» экономики. «Полагаем, что особенно напряженной ситуация с хранением данных будет в банках, так как большая часть розничных покупок в мире к 2010 году будет производиться с помощью пластиковых карт, и объемы банковских архивов будут колоссальными», – прогнозирует отчет IDC;

– опережающие темпы развития Интернета в большинстве стран мира, лишь в немногих развитых странах они близки к насыщению.

В будущем аналитики видят такую картину: к 2010 г. объемы хранящихся архивов достигнут 601 экзабайта, а объем информации, произведенной в 2010 г. достигнет 988 экзабайт (почти 1 зеттабайт).

Можно прогнозировать появление в ближайшее время принципиально нового класса программ, которые будут изучать данные и на основе алгоритмов, заданных пользователями или администраторами, уничтожать лишние данные и сжимать необходимые, экономя место на жестких дисках.

Фактически экстраполяция нынешнего развития Сети приводит ее к «информационному коллапсу», определяемому как «гипотетическое состояние сетевого информационного пространства, угрожающее его стабильности и нормальному функционированию при резком снижении пропускной способности каналов связи, возникает ситуация, когда существующие технологии не в состоянии передать нарастающие объемы трафика».

С точки зрения внешних нетехнических факторов можно охарактеризовать информационный коллапс следующими основными характеристиками:

– стоимость поиска нужных сведений в Интернете возрастает непропорционально быстро по отношению к стоимости повторных исследований с целью их получения;

– объем информации, доступной посредством Интернета, возрастает существенно быстрее развития возможностей ИПС;

– достоверность информации существенно снижается, в связи как с наличием большого количества «версий» одного и того же документа, лишь незначительно отличающихся друг от друга, так и все более характерных тенденциях «вброса» в Сеть заведомо ошибочной или ложной информации.

С точки зрения «внутренних» технически обусловленных угроз можно выделить следующие основные характеристики:

– резкое снижение пропускной способности из-за перегруженности систем;

– атаки хакеров с целью уничтожения или искажения информации, блокирования узлов и «обходных маршрутов» трафика;

– случайные или преднамеренные аварии коммуникационных каналов;

– несовершенство информационно-поисковых систем;

– «моральное» старение протоколов.

По мнению аналитиков, на сегодняшний день примерно 75 % цифровой информации создают и копируют индивидуальные пользователи, а 25 % – организации, но к 2010-му доля последних увеличится до 30 %, поскольку компьютеры все шире используются на предприятиях малого и среднего бизнеса, требования регулирующих органов ужесточаются (т. е. информацию нужно долго хранить), расширяется применение отраслевых приложе-

ний (например, средств электронной коммерции, обслуживания клиентов, камер наружного наблюдения и т. д.). При этом растет не только объем информации, но и число контейнеров для ее хранения, т. е. файлов, пакетов и цифровых изображений.

И хотя основной вклад в информационный бум вносят индивидуальные пользователи, за хранение и защиту 85 % данных отвечают организации (предприятия, агентства, госучреждения, ассоциации). Это налагает на них серьезные требования с точки зрения управления огромными объемами данных и их защиты.

Уже сейчас организации напрасно тратят массу времени при работе с данными. Так, по оценке IDC, предприятие с тысячей сотрудников в среднем ежегодно теряет 5,7 млн дол. из-за необходимости переформатировать информацию и 5,3 млн дол. – из-за невозможности ее найти.

Даваемые аналитиками рекомендации носят по преимуществу общий характер. Так, по мнению IDC, организациям следует применить комплексный и упорядоченный подход к хранению информации. В частности, стоит обратить внимание на технологию управления жизненным циклом информации (Information Lifecycle Management, ILM). Важное значение имеют и новые технологии виртуализации и сервисно-ориентированного программирования, которые повышают гибкость связей между компьютерами, запоминающими устройствами и приложениями. Такой подход позволяет объединить изолированные информационные островки в единый пул и отделить данные от инфраструктуры их хранения. Но для реализации такой архитектуры предприятия должны по-новому взглянуть на свою ИТ-инфраструктуру, повысить ее динамичность и больше внимания уделить вопросам управления информацией.

Тематические интернет-серверы и их роль в систематизации знаний

В настоящее время особую роль приобретает создание специализированных интернет-порталов, ориентированных на конкретные предметные области, особенно в применении к образовательным и исследовательским задачам, и позволяющих гибко решать как задачи поиска и каталогизации актуальной тематической информации, так и выполнять организационно-административные функции.

Понятие информационного тематического портала появилось не так давно, но стало уже довольно распространенным. Первые информационные порталы были «единой точкой входа» в Интернет для пользователей, интересующихся какой-либо темой. Создатели порталов стремились собрать и представить в систематизированном и удобном виде как можно больше полезных информационных ресурсов, распределяя их для удобства по категориям (рубрикам). Некоторые ресурсы были представлены на порталах непосредственно своими материалами, а большинство – гиперссылками.

Интернет-порталы позволяют широкому кругу пользователей находить необходимую информацию наиболее удобным способом, обращаться с запросами и получать ответы организации-владельца портала ежедневно и круглосуточно. Особенность современных интернет-порталов состоит в том, что они способны «подстраиваться» под каждого пользователя, предоставляя ему в первую очередь ту информацию, которая его наиболее интересует, помогая избежать ошибок и недоразумений.

Естественно, что, создавая портал, нужно четко представлять, кто составит его целевую аудиторию, какую информацию посетители найдут на его страницах, кому будет поручена ее актуализация, какие услуги смогут получить пользователи с помощью этого портала.

Накопленный опыт разработки и поддержки сложных тематических интернет-порталов позволяет сформулировать перечисленные ниже требования к структуре и наполнению специализированного портала.

В современных условиях чрезвычайно трудно в одном интернет-ресурсе отслеживать весь быстрорастущий объем информации об исследованиях даже в конкретной предметной области. Выходом является скоординированное создание и сопровождение самостоятельных интернет-ресурсов всеми головными субъектами, отвечающими за развитие соответствующих научных и образовательных направлений.

В этом случае специализированный хостинг интернет-портала обеспечивает возможность консолидации в рамках распределенного портала отдельных тематических сайтов путем фор-

мирования единого поискового каталога интернет-ресурсов предметной области и единой новостной ленты, автоматически собираемой из новостей с отдельных подсайтов.

Оптимальной по соотношению гибкости и эффективности представляется структура портала, состоящая из двух независимых частей, использующих одинаковые программные компоненты:

– внутренняя часть, обеспечивающая информационное взаимодействие субъектов, участвующих в поддержке портала, подготовку и подписание информации к публикации на портале;

– внешняя часть, реализующая информирование пользователей и предоставление интерактивных сервисов.

В процессе реализации системы разрабатывается схема безопасной репликации (синхронизации) информации между двумя частями в соответствии с модифицируемой политикой публикации информации на портале (электронным административным регламентом публикации).

Главное портальное приложение обеспечивает базовую функциональность системы по управлению информационными объектами (ИО) и формированию комплекса навигационных элементов портала (меню, оглавлений, рубрикаторов и т. п.), управление таксономией, формирование дизайна портала, трансформацию и выдачу опубликованных информационных объектов для доступа пользователей, кэширование страниц портала и др.

Внешняя часть портала имеет блочно-модульную структуру и может разрабатываться постепенно, гибко реагируя на возникающие информационные потребности. Требования к структуре блока и модуля определяются ядром системы. Как правило, эти требования сводятся к подключению модулем инициализирующей части программного кода и дальнейшим соблюдением форматов вывода портала (например, в простейшем случае, к записи формируемого контента в глобальную переменную с фиксированным именем).

Подсистема публикации портала основывается на работе с информационными объектами и их метаописаниями. Подсистема обеспечивает поддержку полного жизненного цикла ИО в соответствии с редакционной политикой портала. Подсистема публикации обеспечивает поддержку маршрутизации (последовательности обработки) ИО.

Необходимыми представляются требования предоставления доступа к новостям и другим информационным материалам портала по протоколу RSS, поддержка интерфейса пользователя на русском и английском языках.

В связи с общей картиной относительной распространенности в настоящее время клиентских браузеров, страницы портала оптимизируются для просмотра с помощью браузеров Internet Explorer 5.5 и выше и Mozilla Firefox 1.0 и выше.

Для упорядочения объектов на портале предусматривается использование многомерной таксономии (системы рубрик), включающей как тематические, так и иные классификаторы, причем каждый объект может описываться сразу несколькими классификаторами одновременно.

Административные интерфейсы портала предусматривают гибкие средства управления таксономией, системой меню, а также различными вспомогательными справочниками и классификаторами, используемыми на портале. Должны быть предусмотрены механизмы загрузки внешних справочников и классификаторов, получаемых из смежных систем и иных источников.

Подсистема дискуссий (форумов) обеспечивает поддержку процессов взаимодействия между пользователями портала, управление и доступ к форумам портала. Подсистема позволяет зарегистрированным пользователям публиковать сообщения в темах и просматривать сообщения других пользователей. Подсистема обеспечивает проведение коллективных обсуждений, в том числе с привязкой к опубликованным информационным объектам.

Подсистема электронной библиотеки обеспечивает автоматизацию процессов создания, хранения и представления содержимого информационных объектов.

Подсистема поиска позволит осуществлять полнотекстовый многокритериальный поиск информации по метаописаниям с учетом морфологии русского языка во всех разделах и подразделах портала, с возможностью использования расширенного языка запросов, включающего стандартные логические операции над ключевыми словами для поиска.

Подсистема уведомлений и почтовых рассылок обеспечивает рассылку по электронной почте тематических информационных материалов портала и автоматических уведомлений

об обновлениях и событиях в соответствии с настройками персонализации пользователя. Подсистема обеспечивает возможность получения периодических дайджестов информационных материалов портала.

Подсистема обмена с внешними информационными источниками используется для:

– импорта информационных объектов (новостей, анонсов, мероприятий и т. п.) из внешних источников;

– экспорта информационных объектов на внешние ресурсы.

Обмен с внешними источниками позволяет:

– экспортировать и импортировать пользовательские права и настройки при работе с доверенными ресурсами;

– осуществлять имплементацию новых стандартов взаимодействия без модернизации базовой части подсистемы.

Подсистема хранения данных портала обеспечивает возможность:

– централизованного управления распределенным хранилищем данных;

– реализации логики хранения данных (в том числе обеспечения их целостности и непротиворечивости в любой момент времени) штатными средствами;

– поддержки процессов полнотекстового поиска информационных объектов по их метаописаниям;

– обеспечения вывода информации в стандартизованных в пределах портала форматах;

– обеспечения разделения прав доступа к хранимой информации;

– доступа к данным согласно стандарту ISO / ANSI SQL 92;

– выполнения регламентных работ (резервное копирование, восстановление после сбоев и т. п.) штатными средствами.

Подсистема мониторинга и управления предлагает собственные средства для отслеживания посещаемости и интенсивности использования портала. Данная информация предоставляется в двух формах:

– для нужд управленческого мониторинга с возможность получения уполномоченными лицами данной информации для последующего построения отчетов по степени востребованности информации и сервисов;

– для нужд технического мониторинга с возможностью анализа интенсивности нагрузки на технические средства, выявления угроз безопасности, сбоев и отказов в обслуживании.

Подсистема аутентификации, авторизации и управления пользователями обеспечивает разделение прав доступа пользователей к разделам и сервисам в соответствии с модифицируемой ролевой моделью, как к административному интерфейсу, так и к пользовательской части портала. В состав подсистемы включаются:

– механизмы аутентификации и авторизации пользователей по индивидуальным логину и паролю;

– механизмы создания групп пользователей и управления политикой безопасности для этих групп, включая индивидуальное назначение прав доступа к тем или иным категориям информационных объектов.

На основе данных подсистемы обеспечивается персонализация пользовательского интерфейса (набора информационных объектов, представляемых пользователю).

Портал должен обеспечивать устойчивую работу при нагрузке до 30 000 HTTP-запросов в час, при этом среднее время отклика системы на запрос не превышает 2000 мс, а максимальное – 8000 мс. Данные параметры выдерживаются при выполнении любых комбинаций HTTP-запросов, порождаемых при обращении к произвольным разделам портала.

Таким образом, тематический интернет-портал способен выполнить интегративную, информационную, коммуникативную и обучающую функции в образовательных и аналитических информационных системах.

Онтологии

Практически любой пользователь Интернета хотя бы раз сталкивался с ситуацией, когда при поиске интересующей его информации он помимо прочего получал от поисковой машины множество бесполезных ссылок. Поскольку поиск информации осуществляется вне контек-

ста, никакие уточнения запросов не смогут надежно найти именно то, что нужно. Для качественного осуществления поиска пользователю необходимо понимать все тонкости предметной области, включая ее лексику, термины, определения, иерархии сущностей – одним словом, досконально знать онтологию.

Рассмотрим одно из перспективных направлений решения проблемы влияния стиля документов веб-сервера, связанное с созданием языка онтологий, т. е. общего набора терминов, которые используются для описания и представления объектов в Интернете.

Онтология определяет термины, с помощью которых можно описать предметную область. Использование онтологий особенно необходимо в приложениях-агентах, осуществляющих поиск и объединение информации из различных источников и из разных сред, в которых один и тот же термин может означать разные вещи.

Онтология может описываться различными средствами, и сегодня существует несколько языков описания онтологий, однако ввиду того, что в любой онтологии определяются термины и задаются логические связи между ними, точная семантика описываемых терминов и связей в различных языках будет одна и та же.

Онтология представляет собой совокупность терминов и взаимосвязанных определений, относящихся к некоторой предметной области и выполняющих нормативную функцию. Именно **онтология** формирует самое общее представление об объекте исследования, фиксирует категориальный аппарат концепции (теории).

Развитые **онтологические системы** строятся на основе принципов:

– формализации, т. е. описания объективных элементов действительности в единых, строго определенных образцах (терминах, моделях и др.);

– использования ограниченного количества базовых терминов (сущностей), на основе которых конструируются все остальные понятия;

– внутренней полноты и логической непротиворечивости.

Соблюдение первого принципа дает возможность специалистам в области современных компьютерных технологий (в том числе создания тематических интернет-проектов) сформировать и использовать общий понятийный аппарат. При конструировании определений не допускается применение фраз с нечетким или многозначным смыслом, метафорических выражений и др. Формализованные трактовки терминов фиксируются в тезаурусе (словаре с полной смысловой информацией) или глоссарии. Таким образом, принцип формализации позволяет избежать полисемии (многозначности) понятий, оптимизировать коммуникации между всеми заинтересованными сторонами.

Следование второму принципу позволяет реализовать идею «экономии мышления», широко известную как «бритва Оккама». При составлении онтологических систем стараются использовать минимальный набор базовых категорий, исключить близкие по смыслу, синонимичные понятия. Необходимо сохранять открытость онтологии для пополнения новыми понятиями.

В отличие от обычного словаря для онтологической системы характерны внутреннее единство, логическая взаимосвязь и непротиворечивость используемых понятий. Трактовки всех используемых терминов даются в рамках единого методологического подхода, т. е. явно описанной совокупности исходных принципов, аксиом или убеждений создателей онтологии. При этом используемые категории должны охватывать все явления и процессы заданной предметной области. Таким образом, онтология представляет собой концептуальный фундамент теории, ее понятийную основу.

Онтологические системы всегда обращены к идеальным объектам. Этим онтология теснейшим образом связана с моделированием, решающим задачу представления (репрезентации) реальных объектов через идеальные образы.

В работе [Марчук, Осипов, 2000] описывается формирование онтологии неспецифической информации и базирования разрабатываемых информационных систем на этой онтологии. Предлагается методология структуризации данных, содержащая следующие основные элементы:

– выделение оптимального количества ортогональных сущностей, в совокупности характеризующих наиболее существенные моменты описываемых явлений;

– дифференцирование определений на определения сущностей и определения отношений между сущностями;

– отказ от множественности одноименных предикатов (семантических дуг), «выходящих» из одного узла, через обратные ссылки, обладающие свойством единственности;

– усложнение используемых отношений и их «симметризация».

Термин «сущность» является базовым для формирования информационного пространства. Это то, о чем можно сделать высказывание. При этом сущности должны обладать свойством «различимость». Тогда их можно идентифицировать, т. е. присвоить уникальные идентификаторы, по которым они становятся доступны в информационном пространстве. Следующим важным свойством является возможность установления отношений между сущностями, которые в свою очередь могут быть (стать) сущностями.

В документе должны описываться необходимые базовые классы и отношения (properties) данной системы организации информационного пространства:

– суперкласс для всех классов сущностей, имеющий следующие базовые отношения:

• отношение именования;

• отношения, задающие начальную дату сущности и конечную дату;

• отношения, задающие текстовое описание сущности и произвольный текстовый комментарий;

– класс, определяющий базовую сущность «документа»;

– класс, описывающий частный вариант документа – RDF-документ;

– множество произвольных сущностей;

– множество, задаваемое элементами, которые «связываются» с ресурсом;

– директория – специфическое множество, группирующее RDF-документы;

– экземпляр документа – место («файл») для «хранения» содержимого документа;

– отношение, задающее связь между экземпляром документа и документом;

– указание на то, какой экземпляр документа является оригиналом («последним» экземпляром);

– координата экземпляра документа.

Описанная выше OWL-спецификация используемой базовой модели сущностей и отношений охватывает только наиболее общие представления о внешнем мире. Хотя при построении информационных систем требуется большая конкретизация, а при предметной ориентации информационных систем должны быть еще более существенные расширения, все-таки построение основ онтологий информационных систем могут базироваться на предложенных принципах.

Заключение

Наряду с имеющимися достоинствами предложенного подхода, у него есть ограничения и поэтому он малоприменим, например, для задач поддержки делопроизводства, как более «динамичных» областей построения информационных систем. Тем не менее он достаточно актуален при построении различных «статичных» информационных и информационно-аналитических систем, аналогов библиотек, музеев и архивов.

Примерами тематических серверов с онтологической поддержкой систематизации представленных материалов являются:

– сервер «Методы решения условно-корректных задач» (существует в глобальной сети с 2000 г. по настоящее время);

– сервер «Математические проблемы геофизики» (существует в глобальной сети с 2006 г. по настоящее время).

Оба сервера созданы и поддерживаются коллективом, основной состав которого сформирован из специалистов трех научно-исследовательских институтов СО РАН: Института математики, Института автоматики и электрометрии и Института нефтегазовой геологии и геофизики.

Список литературы

Марчук А. Г., Осипов А. Е. К вопросу об идентификации электронных документов и коллекций // Программирование. 2000. № 3. С. 53–62.

Материал поступил в редколлегию 26.09.2007